LoRA vs. Full Fine-Tuning: A Theoretical Perspective

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning adapts a pre-trained model to downstream tasks using a small amount of labeled data. Low-Rank Adaptation (LoRA) is an efficient fine-tuning method that reduces memory and computation costs while often achieving performance close to full fine-tuning. Despite its widespread use, the theoretical behavior of LoRA in terms of generalization is not yet well understood. In this paper, we study LoRA in a simple linear regression setting and compare its excess risk with that of full fine-tuning. Our analysis identifies regimes in which LoRA achieves lower excess risk than full fine-tuning in both overdetermined and underdetermined settings. Specifically, our theory predicts that LoRA can outperform full fine-tuning when the difference between the pretraining and the downstream tasks is effectively low-rank. We further show how the choice of LoRA rank affects generalization performance, explaining why using a very small rank can improve test accuracy in certain settings, even though it limits model expressivity. Finally, we support our theoretical results with experiments on practical tasks, suggesting that the identified tradeoffs and insights extend beyond linear regression.

1 Introduction

Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023), have achieved strong performance across a wide range of tasks. However, training these models from scratch requires optimizing billions of parameters over massive datasets, resulting in substantial computational, financial, and environmental costs. As model scales continue to grow, full pre-training has become increasingly impractical and often yields diminishing returns relative to its cost.

Fine-tuning (FT) offers an efficient alternative by adapting a pre-trained LLM to a downstream task using a smaller task-specific dataset (Ziegler et al., 2019; Dodge et al., 2020). The straightforward approach, *full fine-tuning (FFT)*, updates all model parameters but incurs substantial memory and computational costs, as gradients must be computed and stored for every trainable weight.

To mitigate these overheads, Hu et al. (2022) proposed using *Low-Rank Adaptation (LoRA)*. Instead of updating the entire weight matrix of each trained layer, LoRA constrains updates to lie in a low-dimensional subspace. This is done by expressing the update as a low-rank matrix factorization. This approach significantly reduces the number of trainable parameters while maintaining performance close to full fine-tuning (Biderman et al., 2024; Shuttleworth et al., 2025) with an appropriate choice of rank. Due to its good performance and efficiency, LoRA has since become one of the most widely adopted fine-tuning methods for LLMs.

Despite LoRA's widespread adoption and strong empirical performance, its statistical generalization behavior remains only partially understood. A limited number of works have begun to study LoRA from a theoretical perspective. For example, Zeng & Lee (2024) analyze the expressive power of LoRA, while focusing on the rank required for fine-tuning a model to match a given target model. In a complementary direction, Zhang et al. (2025) investigate the behavior of LoRA parameters after a single step of gradient descent (GD), and show that even one update can align LoRA with some singular subspaces of the corresponding full fine-tuning solution. In addition, an extensive empirical study by Shuttleworth et al. (2025) demonstrates that LoRA with low ranks can significantly underperform FFT on a coding dataset, while it can achieve almost the same accuracy on a math

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dataset using moderate ranks. Overall, in spite of this recent progress, the performance of LoRA in terms of *generalization* remains largely unclear, particularly compared to other fine-tuning methods.

In this work, we address this gap by studying the generalization of LoRA in the simplified setting of linear regression, focusing on its expected excess risk over the population distribution. We compare its generalization performance to that of FFT in both overdetermined and underdetermined regimes. The goal is to understand how the imposed low-rank constraint influences generalization and to identify regimes under which one method generalizes better than the other. We begin in Sec. 2 by formulating FFT and LoRA within a linear regression framework and presenting their solutions in closed form. Here we present new results on the roles of implicit bias and regularization in solution selection. In Sec. 3, we review existing results on the excess risk of FFT, which correspond to the least-squares and minimum-norm solution in the overdetermined and underdetermined settings, respectively. We then present our main theoretical contributions: novel excess risk bounds for LoRA in both regimes, together with an analysis of their asymptotic behavior in high dimensions. Finally, in Sec. 4, we validate our theoretical findings through controlled linear regression experiments and extend our study to LLM fine-tuning to examine how these insights carry over beyond the linear setting.

2 Problem setting and foundational results

In this paper, we study fine-tuning in a linear regression setting and compare FFT and LoRA in terms of excess risk. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ be a fine-tuning dataset of n i.i.d. random samples. For each sample $(\mathbf{x}_i, \mathbf{y}_i) \sim \mathbb{P}_{\mathbf{x},\mathbf{y}}$, the feature vector $\mathbf{x}_i \in \mathbb{R}^{d_x}$ and its label $\mathbf{y}_i \in \mathbb{R}^{d_y}$ are related through an underlying linear map $\mathbf{A}^\star \in \mathbb{R}^{d_y \times d_x}$ as

$$\mathbf{y}_i = \mathbf{A}^\star \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{d_y}$ is Gaussian noise with distribution $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})$ i.i.d. independent of $\mathbf{x}_i$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d_y \times n}$ be the feature and label matrices, respectively, where the samples are placed in columns. Let $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}$ be the population and empirical second moment matrices of the feature vectors $\{\mathbf{x}_i\}$, defined as

$$\boldsymbol{\Sigma}_{\mathbf{xx}} \coloneqq \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}} \coloneqq \frac{1}{n}\sum\nolimits_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{n}\mathbf{X}\mathbf{X}^\top. \tag{2}$$

Throughout the paper, we assume that $\boldsymbol{\Sigma}_{\mathbf{xx}}$ has full rank. Given an initial (pre-trained) model $\mathbf{A}_0 \in \mathbb{R}^{d_y \times d_x}$, we define the ground-truth adaptation as

$$\boldsymbol{\Delta}^\star = \mathbf{A}^\star - \mathbf{A}_0, \tag{3}$$

where $\boldsymbol{\Delta}^\star \in \mathbb{R}^{d_y \times d_x}$ captures the difference between the given pre-trained model $\mathbf{A}_0$ and the underlying ground-truth map $\mathbf{A}^\star$. This formulation allows us to consider a wide range of fine-tuning scenarios: when $\boldsymbol{\Delta}^\star = \mathbf{0}$, the pre-trained model equals the target linear map, while a nonzero $\boldsymbol{\Delta}^\star$ quantifies the degree of task shift. In general, we do not impose any constraint on $\boldsymbol{\Delta}^\star$, and it can be arbitrary. However, in the following sections, we will analyze how specific structural assumptions on $\boldsymbol{\Delta}^\star$, such as low or full rank, affect the generalization performance of LoRA and FFT.

The goal of fine-tuning is to find an accurate estimator $\widehat{\mathbf{A}} \in \mathcal{A}$ for $\mathbf{A}^\star$, where $\mathcal{A}$ is some feasible set. Here we consider empirical risk minimization with the squared loss, defined as

$$\widehat{\mathcal{R}}(\mathbf{A}) \coloneqq \frac{1}{n}\sum\nolimits_{i=1}^n \left\|\mathbf{A}\mathbf{x}_i - \mathbf{y}_i\right\|^2. \tag{4}$$

Thus, using the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the estimator is given by

$$\widehat{\mathbf{A}} \in \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} \widehat{\mathcal{R}}(\mathbf{A}) = \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2, \tag{5}$$

Below, we formulate FFT and LoRA in linear regression and characterize their solutions in both the overdetermined and underdetermined regimes. Building on that, we derive new results on the roles of implicit bias and regularization in solution selection.

2.1 Full fine-tuning

FFT minimizes the empirical risk without any constraints or regularization, while initializing the weights at the pre-trained model $\mathbf{A}_0$. In certain settings, such as the overparameterized regime

($d_x > n$), the empirical risk may admit multiple global minima. In this work, we opt to consider a solution that minimizes the Euclidean distance to the pre-trained model (initialization). Formally,

$$\min_{\mathbf{A}} \|\mathbf{A} - \mathbf{A}_0\|_{\mathrm{F}} \qquad \text{subject to} \qquad \mathbf{A} \in \operatorname*{arg\,min}_{\mathbf{A}\in\mathbb{R}^{d_y\times d_x}} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 . \tag{FFT}$$

We focus on minimum-norm solutions[1] for four reasons: (*i*) they preserve proximity to the pre-trained model, consistent with the motivation for fine-tuning; (*ii*) they arise from standard optimization and regularization procedures for both FFT and LoRA under mild conditions in our setting (see Thm. 2, Thm. 3 and Prop. 7); (*iii*) they provide a common selection criterion for comparing FFT and LoRA; and (*iv*) they can generalize well even in overparameterized regimes, making their analysis practically relevant. The solution to FFT is classical and is recalled below.

Proposition 1 (FFT solution)**.** *The solution of FFT is given by*

$$\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{FFT}} \qquad \textit{such that} \qquad \widehat{\mathbf{\Delta}}_{\mathrm{FFT}} \coloneqq \big(\mathbf{Y} - \mathbf{A}_0\mathbf{X}\big)\mathbf{X}^{\dagger}, \tag{6}$$

where the superscript † *denotes the Moore–Penrose inverse.*

This expression unifies the FFT solution across the overdetermined and underdetermined regimes. In the overdetermined regime, the empirical second moment $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is invertible, we have $\mathbf{X}^{\dagger} = \mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}$. In this case, the empirical risk admits a unique minimizer that coincides with the ordinary least squares estimator. In contrast, in the underdetermined regime $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is rank deficient, and the empirical risk has infinitely many minimizers. Among these, $\widehat{\mathbf{A}}_{\mathrm{FFT}}$ minimizes the distance to $\mathbf{A}_0$ in every unitarily invariant norm, including the Frobenius, nuclear, and spectral norms. When $\mathbf{X}^{\top}\mathbf{X}$ is invertible, the pseudoinverse simplifies to $\mathbf{X}^{\dagger} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$.

2.2 LoRA FINE-TUNING

LoRA keeps the pre-trained model fixed and optimizes a low-rank additive update. We write $\mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta}$, where $\mathbf{\Delta} \in \mathbb{R}^{d_y\times d_x}$ has rank at most r. Thus, for linear regression, LoRA solves

$$\min_{\mathbf{A}} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 \qquad \text{subject to} \qquad \mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta} \quad \text{with} \quad \operatorname{rank}(\mathbf{\Delta}) \le r. \tag{LoRA}$$

The rank constraint does not fully regularize the problem, *i.e.*, ensure uniqueness. As in FFT, we distinguish two cases according to the rank of $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$. The full-rank case is analogous to the overdetermined FFT regime, although LoRA may admit a family of minimizers. The rank-deficient case parallels the underdetermined FFT regime, where we focus on min-norm solutions. The following theorem characterizes the minimizers in both cases (see proof in App. IV.2).

Theorem 1 (LoRA solution characterization, Sondermann (1986))**.** *Let*

$$\widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} \qquad \textit{such that} \qquad \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \operatorname{trunc}_r\Big\{\widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\Big\}\Big(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\Big)^{\dagger}, \tag{7}$$

where $\operatorname{trunc}_r\{\cdot\}$ *denotes a top-*r *singular-value truncation*[2].

- *If* $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ *has full rank, all minimizers of LoRA are given by* $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$.
- *If* $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ *is rank deficient, any min-Frobenius-norm minimizer of LoRA is given by* $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$.

This solution relates the LoRA and FFT updates: multiply $\widehat{\mathbf{\Delta}}_{\mathrm{FFT}}$ by $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$, truncate it to rank r, and map it back using $\big(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\big)^{\dagger}$. This top-$r$ truncation need not be unique. When it is nonunique, different admissible truncations may yield solutions with different Frobenius distances to $\mathbf{A}_0$. Selecting a minimum-norm solution therefore requires choosing a truncation that attains the smallest such distance. When the truncation is unique, we have the following.

Proposition 2. *Suppose the truncation in* (7) *is unique*[3]*, and let* $p \in (0, \infty)$*. Then:*

- $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ *minimizes the distance to* $\mathbf{A}_0$ *w.r.t. any unitarily invariant norm among LoRA's solutions.*
- $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ *is the unique LoRA solution that minimizes the distance to* $\mathbf{A}_0$ *w.r.t. the Schatten-*p *norm*[4].

[1] A min-norm solution $\mathbf{A}$ w.r.t. a norm $\|\cdot\|$ is a minimizer that also minimizes $\|\mathbf{A} - \mathbf{A}_0\|$.

[2] The expression ranges over all admissible top-r singular-value truncations when the truncation is nonunique.

[3] A rank-r truncation $\operatorname{trunc}_r\{\mathbf{M}\}$ is *not* unique if and only if $\sigma_r(\mathbf{M}) = \sigma_{r+1}(\mathbf{M}) > 0$.

[4] Note that for $p \in (0, 1)$ the Schatten-p is a quasi norm.

If the noise covariance $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \succ \mathbf{0}$, the top-$r$ singular-value truncation is unique almost surely. Under this condition, $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ minimizes the distance to $\mathbf{A}_0$ in every unitarily invariant norm. We next show that $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ can be obtained naturally using GD, even if the truncation is not unique.

Theorem 2. *Let $\mathbf{C} \in \mathbb{R}^{d_y \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_x}$. Consider GD iterates $\{(\mathbf{B}_t, \mathbf{C}_t)\}$ on the reparametrized LoRA objective $\|\mathbf{CBX} - (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\|_{\mathrm{F}}^2$, with $\mathbf{B}_0 = \mathbf{0}$ and $\mathbf{C}_0 \neq \mathbf{0}$. If GD converges to a global minimizer, then $\lim_{t\to\infty} \mathbf{C}_t\mathbf{B}_t$ is given by $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ for some admissible top-r singular-value truncation.*

This result identifies an implicit bias of GD induced by zero initialization of the input-side factor: at any iteration, the update vanishes on directions orthogonal to the span of the training samples. Conditional on convergence to a global minimizer, GD therefore selects a member of the family given by $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$. When the top-$r$ truncation is nonunique, the selected member need not minimize the Frobenius distance to $\mathbf{A}_0$. When the truncation is unique, however, the resulting model is a min-norm solution w.r.t. every unitarily invariant norm by Prop. 2. This phenomenon extends to deeper linear factorizations, with the input-side factor initialized at zero and the remaining factors initialized at nonzero values (see Thm. 8 and its proof in App. IV.2). We next show that vanishing weight decay provides another mechanism for obtaining solutions from this family, independently of the optimizer used, provided that the regularized objective is globally minimized.

Theorem 3. *Let $\mathbf{A} = \mathbf{A}_0 + \mathbf{CB}$ be a LoRA parameterization, where $\mathbf{C} \in \mathbb{R}^{d_y \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_x}$. For every $\lambda > 0$, let $(\widehat{\mathbf{B}}_\lambda, \widehat{\mathbf{C}}_\lambda)$ be a global minimizer of the $L2$ regularized LoRA objective, i.e.,*

$$\min_{\mathbf{B}\in\mathbb{R}^{r\times d_x}, \mathbf{C}\in\mathbb{R}^{d_y\times r}} \left\{ \|\mathbf{CBX} - (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\|_{\mathrm{F}}^2 + \lambda\big(\|\mathbf{B}\|_{\mathrm{F}}^2 + \|\mathbf{C}\|_{\mathrm{F}}^2 \big) \right\}, \tag{8}$$

and denote $\widehat{\mathbf{A}}_\lambda = \mathbf{A}_0 + \widehat{\mathbf{C}}_\lambda\widehat{\mathbf{B}}_\lambda$. Any accumulation point of $\widehat{\mathbf{A}}_\lambda$ as $\lambda \to 0^+$ has the form of $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ for some admissible top-r singular-value truncation. If $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ is unique, then $\widehat{\mathbf{A}}_\lambda \underset{\lambda\to 0^+}{\longrightarrow} \widehat{\mathbf{A}}_{\mathrm{LoRA}}$.

Theorem 3 connects our LoRA solution characterization to the known relationship between $L2$ regularization of the LoRA factors and nuclear-norm regularization of the adaptation. For the standard two-factor parameterization, minimizing the sum of squared Frobenius norms over all factorizations of a fixed adaptation yields twice its nuclear norm. Thus, $L2$ regularization of the factors induces a nuclear-norm penalty on their adaptation under the same rank constraint (Jang et al., 2024). As the regularization strength vanishes, every accumulation point of the regularized solutions has minimum nuclear norm among the minimizers of LoRA. Our theorem identifies the resulting models as members of the family represented by $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$. This characterization extends to depth-m linear factorizations, for which the minimum factor penalty is $m\|\mathbf{\Delta}\|_{S_{2/m}}^{2/m}$ (Dai et al., 2021), where $\|\cdot\|_{S_p}$ denotes the Schatten-p norm. Accordingly, vanishing regularization selects adaptations of minimum Schatten-$2/m$ norm (see Thm. 9 and its proof in App. IV.2). When the top-r truncation is unique, the regularized models converge to the same model $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$, independently of the factorization depth, provided that the rank constraint remains unchanged.

2.3 Expected excess risk

We compare the solutions found by LoRA and FFT in terms of *expected excess risk*, a metric that measures the generalization performance. Given an estimator $\widehat{\mathbf{A}}$, we evaluate its risk (error) over the population distribution. Mathematically, let $(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_{\mathbf{x},\mathbf{y}}$ be a new unseen data pair drawn independently of $\widehat{\mathbf{A}}$. Then the population risk of $\widehat{\mathbf{A}}$ is defined as

$$\mathcal{R}(\widehat{\mathbf{A}}) \coloneqq \mathbb{E}\big[\|\widehat{\mathbf{A}}\mathbf{x} - \mathbf{y}\|^2 \mid \widehat{\mathbf{A}}\big], \tag{9}$$

where the expectation is taken over $\mathbf{x}$ and $\mathbf{y}$. It is often more informative to consider the *excess risk*, which measures the error relative to that of the ground-truth parameter $\mathbf{A}^\star$. In our setting, the excess risk of an estimator $\widehat{\mathbf{A}}$, denoted by $\mathcal{E}(\widehat{\mathbf{A}})$, satisfies (see Lemma 6 in App. VII)

$$\mathcal{E}(\widehat{\mathbf{A}}) \coloneqq \mathcal{R}(\widehat{\mathbf{A}}) - \mathcal{R}(\mathbf{A}^\star) = \big\|(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\big\|_{\mathrm{F}}^2. \tag{10}$$

Note that $\widehat{\mathbf{A}}$ depends on the particular realization of its training data. While the resulting error may be small for some training sets, it can be large for others. To capture the average behavior, we take the expectation with respect to the training set $\mathcal{D}$. Our performance metric is thus the expected excess risk, $\mathbb{E}[\mathcal{E}(\widehat{\mathbf{A}})]$, where the expectation is taken over the randomness in the training set.

3 EXCESS RISK ANALYSIS

In this section, we analyze the excess risk of FFT and LoRA in both overdetermined and underdetermined regimes and compare their performance. Since the behavior of FFT is well understood, we include it only for comparison; our main contribution lies in deriving excess risk bounds for LoRA. Our results hold for every admissible top-r singular-value truncation used to define $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$, without requiring the truncation to be unique.

3.1 OVERDETERMINED (UNDERPARAMETERIZED) REGIME

We start with the overdetermined regime, where $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ has full rank. In this case, the FFT solution is independent of the pre-trained model $\mathbf{A}_0$, and is determined solely by the training data, taking the form $\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{Y}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$. A well-known result, typically stated in the multiple linear regression setting, provides an exact analytic expression for the excess risk (see standard proof in Bach (2024, Prop. 3.10), and App. V.1 for our setting).

Theorem 4 (FFT excess risk). *Assume* $\operatorname{rank}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\right) = d_x$ *almost surely. Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] = \frac{1}{n}\operatorname{Tr}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\operatorname{Tr}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \geq \bar{\sigma}_{\boldsymbol{\varepsilon}}^2\frac{d_x d_y}{n}, \tag{11}$$

where $\bar{\sigma}_{\boldsymbol{\varepsilon}}^2 \coloneqq \mathbb{E}[\|\boldsymbol{\varepsilon}\|^2]/d_y$ *(average noise variance per coordinate).*

From the classical bias-variance perspective, the FFT solution has only variance. Specifically, in the overdetermined regime, whenever $\widehat{\mathbf{A}}_{\mathrm{FFT}}$ is integrable it is unbiased, *i.e.*, $\mathbb{E}[\widehat{\mathbf{A}}_{\mathrm{FFT}}] = \mathbf{A}^\star$. Although the resulting expression for the excess risk is generally intractable, in certain cases we can derive a closed form (see proof in Bach (2024, Sec. 12.2.3) for the standard setting, and App. V.1 for ours).

Proposition 3 (Gaussian case). *Assume* $n > d_x+1$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{xx}})$ *i.i.d.,* $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \succeq \mathbf{0}$, $\mathbf{\Sigma}_{\mathbf{xx}} \succ \mathbf{0}$. *Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] = \operatorname{Tr}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\frac{d_x}{n-d_x-1}. \tag{12}$$

This result shows that, in the overdetermined regime, increasing the number of samples reduces excess risk by lowering variance. As n decreases towards $d_x + 1$, the excess risk blows up, reflecting the instability of the min-norm estimator near the interpolation threshold. From Thm. 4 we see that as long as the average noise variance per coordinate is non-vanishing, the excess risk scales as $\Omega\big(d_x d_y/n\big)$. This indicates that the generalization error of FFT can be quite large in high-dimensional settings. Let us compare this behavior with that of LoRA (see proof in App. VI.1).

Theorem 5 (LoRA excess risk). *Assume* $\operatorname{rank}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\right) = d_x$ *almost surely. Then*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\right)\right] \lesssim\ & \frac{r\max\{d_x,d_y\}}{n}\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \\ & + r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] + \sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\mathbf{\Delta}^\star\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right).\end{aligned} \tag{13}$$

The first term captures the *noise contribution* to the excess-risk bound; we call it the *noise term*. The last two terms capture the *approximation contribution* of the rank constraint. The population-weighted tail of $\mathbf{\Delta}^\star$, given by $\sum_{i>r}\sigma_i^2(\mathbf{\Delta}^\star\mathbf{\Sigma}_{\mathbf{xx}}^{1/2})$, is unavoidable for any rank-r adaptation (Prop. 8). Rapid spectral decay of $\mathbf{\Delta}^\star$ can make these approximation terms small, provided that the feature scale and empirical conditioning factors in the bound are controlled. Thus, let us focus on the behavior of the noise term. The next result present its high-dimensional behavior in the Gaussian setting of Prop. 3.

Proposition 4. *Assume* $n > d_x$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{xx}})$ *i.i.d., and* $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \succeq \mathbf{0}$. *As* $n, d_x \to \infty$ *with* $d_x/n \to c \in [0,1)$, *the first term in* (13) *is asymptotically equal to* $r\max\{d_x,d_y\}(\sqrt{n}-\sqrt{d_x})^{-2}\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})$.

We observe that LoRA may also exhibit instabilities near the interpolation threshold ($n = d_x$). Yet, far from this threshold ($n \gg d_x$), its noise term can be significantly smaller than FFT's variance. In particular, when $\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})$ remains bounded with respect to the dimension d_y, while $\bar{\sigma}_{\boldsymbol{\varepsilon}}^2 = \operatorname{Tr}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})/d_y$ does not vanish, *e.g.*, isotropic noise, LoRA's noise term scales as $\mathcal{O}\big(r\max\{d_x,d_y\}/n\big)$, compared

to the $\Omega(d_x d_y/n)$ scaling of FFT. Since the rank r used in practice is typically much smaller than the ambient dimensions d_x and d_y, LoRA noise term has favorable dimension dependence. In contrast, when the noise distribution is concentrated in a low-dimensional subspace around the ground-truth labels, LoRA is not expected to have a meaningful advantage. Increasing r enlarges the noise term and can reduce the approximation error, reflecting a tradeoff between noise and approximation.

Matrices arising in machine learning often concentrate most of their energy in a few leading singular values (Udell & Townsend, 2019). When the singular values of $\mathbf{\Delta}^\star$ decay rapidly and the feature scale and empirical conditioning are controlled, even a small adaptation rank can yield a small approximation contribution. This can allow LoRA's smaller noise-dependent estimation term to translate into lower excess risk than FFT. Conversely, a substantial spectral tail of $\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{1/2}$ imposes an unavoidable approximation error, limiting the benefit of low-rank adaptation. When $\operatorname{rank}(\mathbf{\Delta}^\star) \le r$, both approximation terms vanish, yielding the following corollary.

Corollary 1. *Under the assumptions of Thm. 5, suppose additionally that* $\operatorname{rank}(\mathbf{\Delta}^\star) \le r$. *Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\right)\right] \lesssim \frac{r \max\{d_x, d_y\}}{n} \lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right) \mathbb{E}\left[\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right]. \tag{14}$$

This corollary shows that when $\operatorname{rank}(\mathbf{\Delta}^\star) \le r$, the approximation terms vanish, leaving only the noise-dependent estimation term in the excess-risk bound. As discussed above, under a mild assumption on the noise covariance, this term scales better than the excess risk of FFT. This formally demonstrates the advantage of LoRA over FFT whenever the underlying update has low rank.

3.2 Underdetermined (overparameterized) regime

We now consider the underdetermined regime, where $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is rank deficient. This reflects the common fine-tuning setting of adapting a large pre-trained model using a small, domain-specific dataset. The FFT solution (6) takes the form $\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{Y}\mathbf{X}^\dagger + \mathbf{A}_0 \mathbf{P}_{\mathbf{X}}^{\perp}$, where $\mathbf{P}_{\mathbf{X}}^{\perp}$ projects onto the orthogonal complement of the column space of $\mathbf{X}$. Thus, FFT preserves the pre-trained model along directions unobserved in the training data. The following result gives its exact expected excess risk (see Bach, 2024, Sec. 12.2.3 and App. V.2 for our setting).

Theorem 6 (FFT excess risk)**.** *Let* $\widehat{\mathbf{A}}_{\mathrm{FFT}}$ *be the FFT solution* (6)*, then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] = \underbrace{\mathbb{E}\left[\left\|\mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}}^{\perp} \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right]}_{\textit{FFT Bias}} + \underbrace{\frac{1}{n} \operatorname{Tr}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right) \mathbb{E}\left[\operatorname{Tr}\left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right]}_{\textit{FFT Variance}}. \tag{15}$$

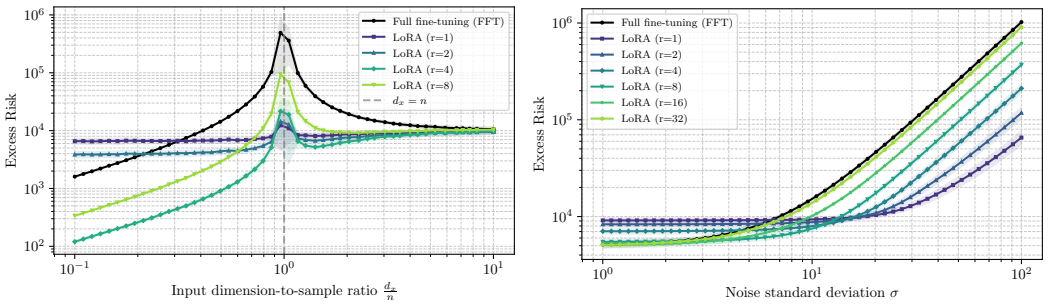
This result generalizes Thm. 4 to the underdetermined regime. Unlike the overdetermined case, the FFT estimator this time is biased, as reflected in the first term of this expression. To gain insight into these terms, we return to the Gaussian case. Specifically, we can obtain a closed-form expression when the features are distributed isotropically (see proof (Bach, 2024, Sec. 12.2.3) and App. V.2).

Proposition 5 (Gaussian case)**.** *Assume* $n < d_x - 1$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{x}}^2 \mathbf{I})$ *i.i.d., and* $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} \succeq \mathbf{0}$. *Then*

$$\textit{FFT Bias} = \sigma_{\mathbf{x}}^2 \left\|\mathbf{\Delta}^\star\right\|_{\mathrm{F}}^2 \frac{d_x - n}{d_x}, \qquad \textit{and} \qquad \textit{FFT Variance} = \operatorname{Tr}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right) \frac{n}{d_x - n - 1}. \tag{16}$$

We see that the bias term (left) decreases linearly with the number of samples n, whereas the variance term (right) increases with n. This behavior arises because the estimator is constructed from observed samples: each additional sample effectively adds a noisy dimension to the estimation. As a result, in the underdetermined regime, increasing the number of samples does not necessarily improve performance. When $\operatorname{Tr}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) \gg \sigma_{\mathbf{x}}^2 \left\|\mathbf{\Delta}^\star\right\|_{\mathrm{F}}^2$, the problem lies in the strong-noise regime[5], and the excess risk is dominated by the variance term. In this regime, adding more samples is expected to increase the risk. Conversely, when $\operatorname{Tr}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) \ll \sigma_{\mathbf{x}}^2 \left\|\mathbf{\Delta}^\star\right\|_{\mathrm{F}}^2$, the problem is in the weak-noise regime, and if n is not too close to d_x, the excess risk is governed by the bias term and the total risk decays with n. We now compare this behavior with that of LoRA (see proof in App. VI.2)

[5]Note that $\operatorname{Tr}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) = \mathbb{E}[\|\boldsymbol{\varepsilon}\|^2]$ is the expected noise magnitude.



(a) Excess risk vs. the dimension-to-sample ratio

(b) Excess risk vs. noise level

Figure 1: **Linear regression experiments.** Panel (a) presents the excess risk of FFT and LoRA under varying sample size n (decreasing from left to right), fixed dimensions $d_x = d_y = 100$, noise magnitude $\sigma = 12$, and true task-shift rank $\mathrm{rank}(\mathbf{\Delta}^\star) = 4$. Panel (b) plots the excess risk as a function of noise level $\sigma \in [1, 100]$ with $d_x = d_y = 100$, $n = 50$, and $\mathrm{rank}(\mathbf{\Delta}^\star) = 10$. Results are averaged over 100 random seeds; shaded regions indicate ± 1 standard deviation.

Theorem 7 (LoRA excess risk). *Let* $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ *be the LoRA solution* (7)*. If* $n < d_x$*, then*

$$\mathbb{E}\left[\mathcal{E}\big(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\big)\right] \lesssim \mathbb{E}\left[\left\|\mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}}^{\perp} \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] + \frac{r \max\{n, d_y\}}{n} \lambda_{\max}\big(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\big) \mathbb{E}\left[\lambda_{\max}\big(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\dagger}\big)\right] + r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right) \lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right] + \sum_{i=r+1}^{\min\{d_x, d_y\}} \sigma_i^2\left(\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right). \quad (17)$$

The first term captures the error from *unobserved directions*, orthogonal to the span of the training inputs. The remaining terms comprise a noise term and two approximation terms. Let us compare this bound with the excess risk of FFT. The error contribution from unobserved directions is identical to FFT's bias. We next evaluate the noise term in the Gaussian asymptotics (see proof in App. VI.2)

Proposition 6. *Assume* $d_x > n$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{x}}^2 \mathbf{I})$ *i.i.d. and* $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} \succeq \mathbf{0}$*. As* $n, d_x \to \infty$ *with* $n/d_x \to c \in [0, 1)$*, the second term in* (17) *is asymptotically equal to* $r \max\{n, d_y\}(\sqrt{d_x} - \sqrt{n})^{-2} \lambda_{\max}\big(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\big)$.

The comparison between LoRA's noise term and FFT's variance in the underdetermined regime is similar to that in the overdetermined regime. Again, away from the interpolation threshold ($d_x \gg n$), under mild assumptions on the noise covariance[6], LoRA's noise term scales as $\mathcal{O}\big(r \max\{n, d_y\}/d_x\big)$, compared to the $\Theta\big(n d_y / d_x\big)$ scaling of the corresponding variance term in FFT. Since the rank r used in practice is typically much smaller than the ambient dimension d_y and the number of samples n, LoRA's noise term has more favorable dimension dependence than FFT's variance. Therefore, just like in the overdetermined regime, LoRA can outperform FFT if the approximation error is negligible. As in the overdetermined regime, rapid spectral decay of $\mathbf{\Delta}^\star$ can make this contribution small, provided that the feature scale and pseudoinverse conditioning factors are controlled.

Corollary 2. *Under the assumptions of Thm. 7, suppose additionally that* $\mathrm{rank}(\mathbf{\Delta}^\star) \le r$*. Then*

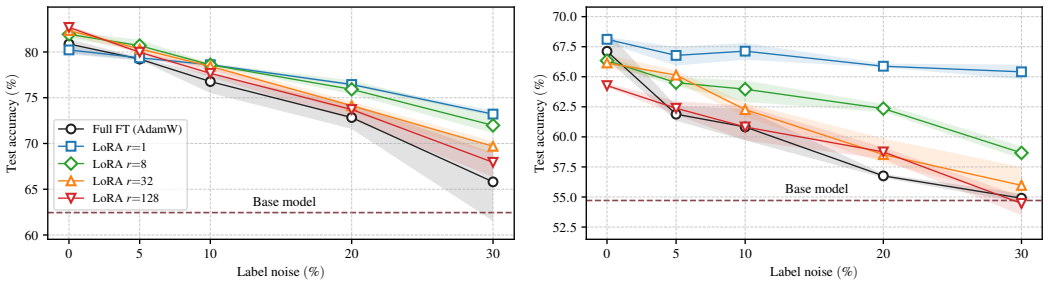
$$\mathbb{E}\left[\mathcal{E}\big(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\big)\right] \lesssim \mathbb{E}\left[\left\|\mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}}^{\perp} \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] + \frac{r \max\{n, d_y\}}{n} \lambda_{\max}\big(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\big) \mathbb{E}\left[\lambda_{\max}\big(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\dagger}\big)\right]. \quad (18)$$

Overall, our results show that when $\mathbf{\Delta}^\star$ is effectively low rank, *i.e.*, rapidly decaying spectrum, LoRA can achieve lower generalization error than FFT. The smaller noise term is especially relevant when FFT's excess risk is dominated by noise.

4 Experiments

In this section, we empirically validate our theory. We first examine a controlled linear regression setting to isolate the effects of rank, sample size, and noise, and then conduct large-scale fine-tuning experiments with LLMs to determine whether similar behaviors arise in real-world fine-tuning.

[6]The same condition from the overdetermined case, see discussion below Prop. 4



(a) Qwen2.5-0.5B on BoolQ

(b) Qwen2.5-0.5B on CommonsenseQA

Figure 2: **Effect of label noise in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different levels of label noise. For each configuration, we trained 3 times using random seeds. Panels (a) and (b) show the mean and the range of the results for the 0.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. Here, in the strong-noise regime, LoRA outperforms FFT as predicted by our theory. For other model sizes, see App. VIII.12.

Linear regression experiment. For this experiment, we use isotropic Gaussian feature vectors, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{d_x})$, where the labels are generated as in (1) with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_{d_y})$. The underlying linear map is created as $\mathbf{A}^\star = \mathbf{A}_0 + \boldsymbol{\Delta}^\star$, where the pre-trained model $\mathbf{A}_0$ has i.i.d. standard Gaussian entries, and $\boldsymbol{\Delta}^\star = \frac{1}{\sqrt{r^\star}}\mathbf{U}\mathbf{V}$ is a low-rank adaptation constructed by sampling $\mathbf{U} \in \mathbb{R}^{d_y \times r^\star}$ and $\mathbf{V} \in \mathbb{R}^{r^\star \times d_x}$ with i.i.d. normal Gaussian entries; the scaling ensures bounded perturbation magnitude. We use the excess risk $\mathcal{R}(\widehat{\mathbf{A}})$ from Sec. 2.3 to evaluate the performance of the obtained estimators. Figure 1 plots the excess risk of FFT and LoRA against sample size and the noise level.

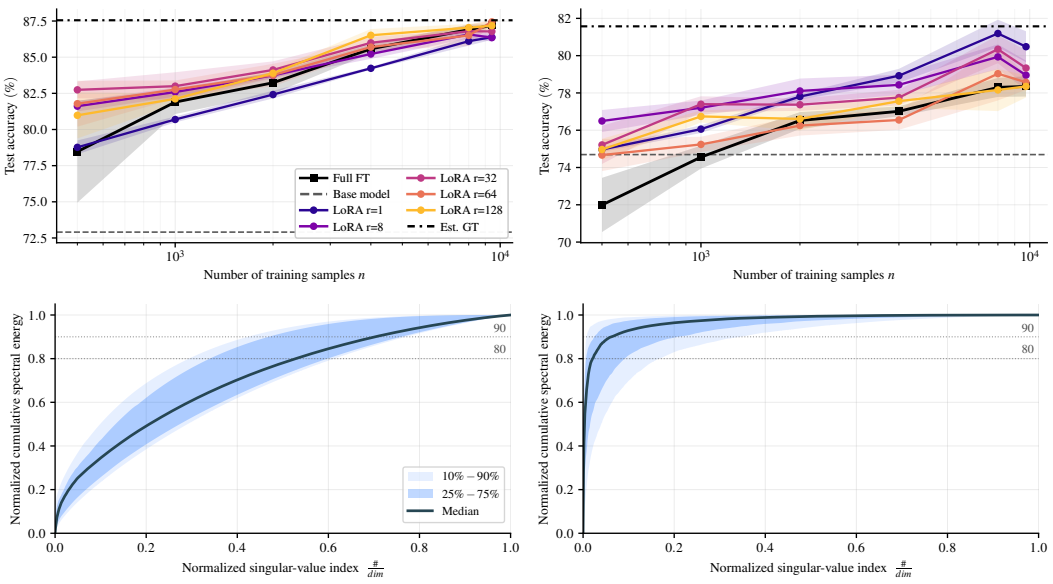
The gray line at $d_x = n$ in Fig. 1(a) marks the interpolation threshold, separating the overdetermined and underdetermined regimes. In the overdetermined regime ($n > d_x$), smaller sample sizes lead to higher excess risk for all methods, consistent with our theoretical predictions. Moreover, when r matches or exceeds the rank of the true shift $r^\star$, LoRA outperforms FFT, with the best performance attained at $r = r^\star = 4$. Using a larger rank than the ground-truth only slightly degrades performance, as it introduces unnecessary degrees of freedom. For smaller ranks, *e.g.*, $r \in \{1, 2\}$, the excess risk remains almost constant, reflecting the dominance of the approximation error imposed by the rank constraint. Since in this case, the spectral tail of $\boldsymbol{\Delta}^\star$ is significant, and as predicted by our theory, LoRA performs poorly.

In the underdetermined regime, most configurations exhibit a decrease in excess risk as the number of training samples gets smaller. As explained above (see discussion below Prop. 5), this behavior arises because the estimators are constructed from the observed samples, while each additional sample effectively introduces a new noisy dimension into the estimation. This affects both LoRA and FFT, where the effect's strength on LoRA depends on the rank. Specifically, just after the interpolation threshold $d_x = n$, the noise term dominates the excess risk, and decreasing n reduces it until the error from rank approximation and unobserved directions becomes dominant and the curves stabilize.

Figure 1(b) shows the effect of noise magnitude in the underdetermined regime. When the noise level is low, FFT performs best and is comparable to high-rank LoRA. As the noise increases, FFT degrades rapidly and becomes much less robust than low-rank LoRA. This behavior is expected, since updating many unnecessary directions amplifies noise, whereas LoRA reduces this effect.

LLM fine-tuning. We next examine whether similar trends arise in practical LLM fine-tuning. We fine-tune Qwen2.5 models (Yang et al., 2025) (0.5B, 1.5B, and 3B) using FFT and LoRA with various ranks across different levels of label noise/sample sizes. For each configuration, we run 3 random seeds on BoolQ (Clark et al., 2019) and CommonsenseQA (Talmor et al., 2019) datasets.

Figure 2 depicts the test accuracy of the trained 0.5B model under increasing percentage of label noise. In the weak-noise regime, no consistent ordering across ranks is observed. Whereas in the strong-noise regime, performance decreases with increasing rank. This trend is consistent with the noise-reduction mechanism predicted by our theory: when noise dominates, restricting the number of trainable directions can improve generalization (see discussion below Corollary 1).



(a) Qwen2.5-1.5B on BoolQ

(b) Qwen2.5-1.5B on CommonsenseQA

Figure 3: **Effect of sample size in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different sample sizes. For each configuration, we trained 3 times using random seeds. The top of Panels (a) and (b) shows the mean and the range of the results for the 1.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. For each task, we estimated $\mathbf{\Delta}^\star$ and computed its singular values per layer. The bottom part shows the statistics of the spectral cumulative energy of $\mathbf{\Delta}^\star$ across layers. When the effective rank of the adaptation is low (CommonsenseQA), LoRA outperforms FFT, as predicted by our theory. For other model sizes, see App. VIII.12.

Finally, Fig. 3 shows the test accuracy of the trained 1.5B model on varying training set sizes. To estimate the spectral structure of the ground-truth adaptation, we train a high-performing reference model using additional training samples and model averaging across multiple runs; details are provided in App. VIII. We then compute the per-layer normalized cumulative spectral energy of the corresponding weight difference $\mathbf{\Delta}^\star$, given by $s(r) = \sum_{i=1}^{r} \sigma_i^2(\mathbf{W}) / \|\mathbf{W}\|_{\mathrm{F}}^2$. The lower panels visualize the distribution of this quantity across layers. For CommonsenseQA, the reference update exhibits a rapidly decaying spectrum, suggesting that an optimal adaptation is well captured by a low-rank update. Correspondingly, LoRA significantly outperforms FFT, consistent with our theoretical prediction that low-rank adaptations favor LoRA. In contrast, for BoolQ, the spectral tail is heavier, suggesting that the underlying adaptation is less well approximated by a low-rank structure. In this case, LoRA provides no significant advantage over FFT, aligning with our theoretical expectations.

5 Conclusion

In this paper, we provided a rigorous theoretical analysis of the generalization behavior of LoRA compared to FFT by deriving excess risk bounds for linear models in both overdetermined and underdetermined regimes. Our bounds identify a smaller noise term as a potential advantage of LoRA, particularly when the task shift is effectively low rank and in the strong noise regime. Fine-tuning experiments on Qwen2.5 models across reasoning tasks reveal patterns consistent with these insights.

AI USE STATEMENT

In this work, we used generative AI tools for <assist in the writing of proofs, implement methods, and generate synthetic data sets>. We have not used generative AI tools for <propose or refine hypotheses, design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments> and <assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results> are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, UK, 1 edition, 2013. ISBN 978-0-19-953525-5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*. Association for Computational Linguistics, 2019.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. *Advances in Neural Information Processing Systems*, 34:26884–26896, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2021.

Yehoram Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6151–6159, 2017.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing LM adaptation with Tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

Uijeong Jang, Jason D Lee, and Ernest K Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21306–21328. PMLR, 2024.

Junsu Kim, Jaeyeon Kim, and Ernest K Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Anastasis Kratsios, Tin Sum Cheng, Aurelien Lucchi, and Haitz Sáez de Ocáriz Borde. Sharp generalization bounds for foundation models with asymmetric randomized low-rank adapters. *arXiv preprint arXiv:2506.14530*, 2025.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025.

Jack W Silverstein. The smallest eigenvalue of a large dimensional Wishart matrix. *The Annals of Probability*, 13(4):1364–1368, 1985.

Dieter Sondermann. Best approximate solutions to matrix equations under rank restrictions. *Statistische Hefte*, 27(1):57–66, 1986.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*. Association for Computational Linguistics, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5. Cambridge University Press, 2012.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Appendix

I SUMMARY OF NOTATION

Table 1: Summary of notation used in the paper

Symbol	Definition	Mathematical definition / short note
$\mathbf{x}_i, \mathbf{x}$	Fine-tuning input vector / test data	$\mathbf{x}_i, \mathbf{x} \in \mathbb{R}^{d_x}$
$\mathbf{y}_i, \mathbf{y}$	Fine-tuning output label / test label	$\mathbf{y}_i, \mathbf{y} \in \mathbb{R}^{d_y}, \mathbf{y}_i = \mathbf{A}^\star \mathbf{x}_i + \boldsymbol{\varepsilon}_i$
$\mathbf{X}, \mathbf{Y}$	Pre-training data matrices	$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n],\ \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$
$\varepsilon,\ \boldsymbol{\xi}$	Noise vector / matrix	$\mathbb{E}[\boldsymbol{\varepsilon}] = 0,\ \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$
$\boldsymbol{\Sigma}_{\mathbf{xx}},\ \boldsymbol{\Sigma}_{\mathbf{yx}}$	Population input second moment / cross moment	$\boldsymbol{\Sigma}_{\mathbf{xx}} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top],\ \boldsymbol{\Sigma}_{\mathbf{yx}} = \mathbb{E}[\mathbf{y}\mathbf{x}^\top]$
$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}},\ \widehat{\boldsymbol{\Sigma}}_{\mathbf{yx}}$	Empirical second moment / cross moment	$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top,\ \widehat{\boldsymbol{\Sigma}}_{\mathbf{yx}} = \frac{1}{n}\mathbf{Y}\mathbf{X}^\top$
d_x, d_y	Input and output dimensions	positive integers (dimensions)
$\mathcal{D}$	Fine-tuning dataset	collection of $(\mathbf{x}_i, \mathbf{y}_i)$
$\mathbf{A}^\star$	True map for FT dataset	$\mathbf{y} = \mathbf{A}^\star \mathbf{x} + \boldsymbol{\varepsilon}$ (true map)
$\mathbf{A}_0$	Pre-trained model (nonzero)	fixed initializer used before FT
$\widehat{\mathbf{A}}_{\mathrm{FFT}}$	Solution found by Full Fine-tuning	$\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{A}_0 + (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\mathbf{X}^\dagger$
$\widehat{\mathbf{A}}_{\mathrm{LoRA}}$	Solution found by LoRA	$\widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \mathrm{trunc}_r\left\{\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^\dagger$
$\boldsymbol{\Delta}^\star$	True fine-tuning update	$\boldsymbol{\Delta}^\star \coloneqq \mathbf{A}^\star - \mathbf{A}_0$
$\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}, \widehat{\boldsymbol{\Delta}}_{\mathrm{LoRA}} \in \mathbb{R}^{d_y \times d_x}$	FFT / LoRA update matrices	$\mathrm{rank}(\widehat{\boldsymbol{\Delta}}_{\mathrm{LoRA}}) \le r$
r	LoRA (adaptation) rank	chosen rank for low-rank adaptation
$\mathbf{X}^\dagger$	Moore–Penrose pseudoinverse of $\mathbf{X}$	$\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ when full column rank
$\mathbf{P}_{\mathbf{X}}$	Orthogonal projection onto the column space of $\mathbf{X}$	$\mathbf{P}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\dagger \in \mathbb{R}^{d_x \times d_x}$
$\boldsymbol{\Pi}_{\mathbf{X}}$	Orthogonal projection onto the row space of $\mathbf{X}$	$\boldsymbol{\Pi}_{\mathbf{X}} = \mathbf{X}^\dagger\mathbf{X} \in \mathbb{R}^{n \times n}$
$\mathbf{P}_{\mathbf{X}}^\perp$	Orthogonal projection onto $\ker(\mathbf{X}^\top)$	$\mathbf{P}_{\mathbf{X}}^\perp = \mathbf{I}_{d_x} - \mathbf{P}_{\mathbf{X}}$
$\mathcal{R}(\mathbf{A})$	Population risk of estimator $\mathbf{A}$	$\mathcal{R}(\mathbf{A}) = \mathbb{E}\left[\lVert\mathbf{y} - \mathbf{A}\mathbf{x}\rVert_2^2\right]$
$\sigma_i(\mathbf{M})$	i-th singular value of $\mathbf{M}$	ordered $\sigma_1 \ge \cdots \ge \sigma_{\min}$
$\lVert\cdot\rVert_{\mathrm{F}},\ \lVert\cdot\rVert_2$	Frobenius and operator norms	$\lVert\mathbf{M}\rVert_{\mathrm{F}} = \sqrt{\mathrm{Tr}(\mathbf{M}^\top\mathbf{M})},\ \lVert\mathbf{M}\rVert_2 = \sigma_{\max}(\mathbf{M})$
$\mathrm{Tr}(\cdot)$	Trace operator	sum of diagonal entries
$\mathrm{rank}(\cdot)$	Matrix rank	dimension of column space
$\lambda_{\min}(\cdot),\ \lambda_{\max}(\cdot)$	Smallest / largest eigenvalue	for symmetric PSD matrices

II Related work

The effectiveness of LoRA and its comparison with FFT have been the focus of several recent papers. On the empirical side, Shuttleworth et al. (2025) examined the spectral properties of weight matrices learned by LoRA and FFT. They found that these properties are very different, and LoRA training creates new high-ranking singular vectors called *intruder dimensions*. Experiments by Biderman et al. (2024) showed that LoRA often underperforms FFT on tasks such as coding, math, and instruction fine-tuning, but it suffers less from catastrophic forgetting and performs better than common regularization methods for reducing forgetting. Similarly, Ivison et al. (2023) found that QLoRA (Dettmers et al., 2023) falls short of FFT on long-form generation tasks. Additionally, Zhuo et al. (2024) empirically studied fine-tuning across various model sizes and methods and found that FFT typically achieves the best performance, while LoRA offers the best balance between performance and efficiency.

On the theoretical side, Zeng & Lee (2024) examined the expressive power of LoRA, focusing on its ability to represent smaller target models. They considered fully connected and transformer architectures and showed that, with sufficiently large rank, LoRA can adapt to these small models. Likewise, Zhang et al. (2025) demonstrated that LoRA adapters align with specific singular subspaces of a single full fine-tuning gradient step, leading to the LoRA-One algorithm that achieves linear convergence and better handles ill-conditioning through preconditioners. Finally, Kratsios et al. (2025) derived generalization bounds for asymmetric randomized LoRA, showing a sample complexity of $\tilde{\mathcal{O}}(\sqrt{r/N})$ for rank r and N samples, along with matching lower bounds that highlight its efficiency.

A separate line of work studies the implicit bias of factorized optimization. Gunasekar et al. (2017) proved minimum-nuclear-norm selection for commuting measurements and conjectured it for general matrix sensing. Li et al. (2021) later gave counterexamples, showing that minimum-nuclear-norm selection need not hold in general matrix sensing and connecting infinitesimal gradient flow to greedy low-rank learning. For LoRA, Kim et al. (2025) showed that zero initialization and weight decay favor low-rank, small-magnitude global minima. For our structured LoRA problem, when the top-r truncation is unique, Proposition 2 shows that the resulting solution minimizes every Schatten norm. Theorem 2 shows that GD with input-side zero initialization selects a basic LoRA solution, conditional on convergence to a global minimizer; under the same uniqueness condition, this is the minimum-norm solution.

Our excess risk analysis for LoRA fine-tuning is closely related to the literature on multivariate linear regression with rank constraints. However, our setting differs from the existing literature in three main ways. First, we study generalization through the population risk, *i.e.*, performance on unseen test data, rather than the empirical risk on the training set used by prior work. This makes our setting substantially more challenging. Moreover, bounds on empirical risk do not directly translate into guarantees on test error and, therefore, are insufficient for assessing generalization. Second, our results do not require any rank assumption on the underlying linear map that generates the data. In contrast, much of the existing work assumes that the true regression operator is low-rank. Third, our analysis applies to both the underdetermined and overdetermined regimes. For example, Giraud (2021) studied rank-constrained regression in the underdetermined regime and derived an upper bound on the excess risk over the *training set*. Similarly, Rigollet & Hütter (2023) analyzed estimators based on singular value thresholding and least squares with rank penalization in the overdetermined low-rank regime, establishing training error bounds of the form $r \max\{d_x, d_y\}/n$.

III Limitations and future work

Our theoretical results apply to the simple setting of linear models with the MSE loss. A natural way to extend our analysis to nonlinear models is to use the Neural Tangent Kernel (NTK) regime, in which nonlinear models behave as linear. However, showing that LLMs also exhibit an NTK regime is beyond the scope of this paper. Additionally, to fully reduce the comparison between LoRA and FFT from nonlinear models to linear regression, as in our setting, one needs the linearization to be the same for both methods. This is nontrivial, and we leave it for future work. Finally, the qualitative empirical observations from our LLM experiments are only suggestive of the patterns identified in our linear theory; further empirical investigation is required.

IV DERIVATION OF THE EXPLICIT SOLUTIONS

In this section, we derive the explicit solution found by LoRA and FFT during the fine-tuning. We note that these results are well known and do not claim any novelty in this regard. We include these derivations for the completeness of this work.

IV.1 EXPLICIT SOLUTION OF FULL FINE-TUNING

Proposition 7 (FFT solution under GD)**.** *Let $\mathbf{X} \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times n}$. Consider the minimization problem*

$$\min_{\mathbf{A} \in \mathbb{R}^{d_y \times d_x}} \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2, \tag{19}$$

Let $\{\mathbf{A}_t\}$ be the iterates of GD on this objective with step size $\eta > 0$ initialized at $\mathbf{A}_0 \in \mathbb{R}^{d_y \times d_x}$. If $\eta \lambda_{\max}(\mathbf{X}\mathbf{X}^\top) < 2$, then

$$\lim_{t \to \infty} \mathbf{A}_t = \widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{A}_0 + (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\,\mathbf{X}^\dagger. \tag{20}$$

Proof. Recall that our objective function is

$$\mathcal{L}(\mathbf{A}) = \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2, \tag{21}$$

and therefore its gradient in matrix form is given by

$$\nabla \mathcal{L}(\mathbf{A}) = (\mathbf{A}\mathbf{X} - \mathbf{Y})\,\mathbf{X}^\top. \tag{22}$$

Let us write one step of GD.

$$\mathbf{A}_{t+1} = \mathbf{A}_t - \eta \nabla \mathcal{L}(\mathbf{A}_t) = \mathbf{A}_t - \eta\,(\mathbf{A}_t\mathbf{X} - \mathbf{Y})\,\mathbf{X}^\top. \tag{23}$$

Subtracting $\mathbf{Y}\mathbf{X}^\dagger$ from both sides gives

$$\Rightarrow \quad \mathbf{A}_{t+1} - \mathbf{Y}\mathbf{X}^\dagger = \mathbf{A}_t - \mathbf{Y}\mathbf{X}^\dagger - \eta\,(\mathbf{A}_t\mathbf{X} - \mathbf{Y})\,\mathbf{X}^\top. \tag{24}$$

Recall that $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}^\dagger\mathbf{X}$ is the orthogonal projection onto the row space of $\mathbf{X}$, and therefore $\mathbf{\Pi}_{\mathbf{X}}\mathbf{X}^\top = \mathbf{X}^\top$. Thus,

$$(\mathbf{A}_t\mathbf{X} - \mathbf{Y})\,\mathbf{X}^\top = \mathbf{A}_t\mathbf{X}\mathbf{X}^\top - \mathbf{Y}\mathbf{X}^\top = \mathbf{A}_t\mathbf{X}\mathbf{X}^\top - \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}\mathbf{X}^\top = \left(\mathbf{A}_t - \mathbf{Y}\mathbf{X}^\dagger\right)\mathbf{X}\mathbf{X}^\top \tag{25}$$

Therefore,

$$\Rightarrow \mathbf{A}_{t+1} - \mathbf{Y}\mathbf{X}^\dagger = \mathbf{A}_t - \mathbf{Y}\mathbf{X}^\dagger - \eta\left(\mathbf{A}_t - \mathbf{Y}\mathbf{X}^\dagger\right)\mathbf{X}\mathbf{X}^\top = \left(\mathbf{A}_t - \mathbf{Y}\mathbf{X}^\dagger\right)\left(\mathbf{I} - \eta\mathbf{X}\mathbf{X}^\top\right). \tag{26}$$

Unrolling the recursion for T steps gives

$$\mathbf{A}_T - \mathbf{Y}\mathbf{X}^\dagger = \left(\mathbf{A}_0 - \mathbf{Y}\mathbf{X}^\dagger\right)\left(\mathbf{I} - \eta\mathbf{X}\mathbf{X}^\top\right)^T. \tag{27}$$

Since $\eta > 0$ and $\eta\lambda_{\max}(\mathbf{X}\mathbf{X}^\top) < 2$, every positive eigenvalue λ of $\mathbf{X}\mathbf{X}^\top$ satisfies $|1 - \eta\lambda| < 1$, while each zero eigenvalue corresponds to an eigenvalue 1 of $\mathbf{I} - \eta\mathbf{X}\mathbf{X}^\top$. Since $\ker(\mathbf{X}\mathbf{X}^\top) = \mathrm{col}(\mathbf{X})^\perp$, the spectral theorem gives

$$\lim_{T \to \infty} \left(\mathbf{I} - \eta\mathbf{X}\mathbf{X}^\top\right)^T = \mathbf{I} - \mathbf{P}_{\mathbf{X}}, \tag{28}$$

where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\dagger$ is the orthogonal projection onto the column space of $\mathbf{X}$. Consequently, the iterates converge, and their limit is

$$\begin{aligned}
\lim_{t \to \infty} \mathbf{A}_t &= \mathbf{Y}\mathbf{X}^\dagger + \left(\mathbf{A}_0 - \mathbf{Y}\mathbf{X}^\dagger\right)(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \\
&= \mathbf{Y}\mathbf{X}^\dagger + \left(\mathbf{A}_0 - \mathbf{Y}\mathbf{X}^\dagger\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger\right) \\
&= \mathbf{A}_0 + \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}\mathbf{X}^\dagger - \mathbf{A}_0\mathbf{X}\mathbf{X}^\dagger \\
&= \mathbf{A}_0 + (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\,\mathbf{X}^\dagger
\end{aligned} \tag{29}$$

where in the last step we used $\mathbf{X}^\dagger\mathbf{X}\mathbf{X}^\dagger = \mathbf{X}^\dagger$. □

IV.2 Explicit solution of rank constraint linear regression and LoRA

Theorem 1 (LoRA solution characterization, Sondermann (1986))**.** *Let*

$$\widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} \qquad \textit{such that} \qquad \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \mathrm{trunc}_r\Big\{\widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\Big\}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{\dagger}, \tag{7}$$

where $\mathrm{trunc}_r\{\cdot\}$ *denotes a top-*r *singular-value truncation*[7].

- *If* $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ *has full rank, all minimizers of LoRA are given by* $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$.
- *If* $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$ *is rank deficient, any min-Frobenius-norm minimizer of LoRA is given by* $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$.

Proof. Define the residual $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{A}_0\mathbf{X}$ and let $\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} = \bar{\mathbf{Y}}\mathbf{X}^{\dagger}$ be the min-norm adaptation of FFT, given in Prop. 1. Using $\mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta}$, let us reformulate the LoRA objective as

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 &= \|\mathbf{Y} - (\mathbf{A}_0 + \mathbf{\Delta})\mathbf{X}\|_{\mathrm{F}}^2 \\
&= \left\|\bar{\mathbf{Y}} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X} + \left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right\|_{\mathrm{F}}^2 \\
&= \left\|\bar{\mathbf{Y}} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X}\right\|_{\mathrm{F}}^2 + \left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right\|_{\mathrm{F}}^2 \\
&\qquad + 2\,\mathrm{Tr}\left(\left(\bar{\mathbf{Y}} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X}\right)\mathbf{X}^{\top}\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right).
\end{aligned} \tag{30}$$

Note that the third term above is, in fact, zero since

$$\begin{aligned}
\mathrm{Tr}\left(\left(\bar{\mathbf{Y}} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X}\right)\mathbf{X}^{\top}\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right) &= \mathrm{Tr}\left(\left(\bar{\mathbf{Y}}\mathbf{X}^{\top} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X}\mathbf{X}^{\top}\right)\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right) \\
&= \mathrm{Tr}\left(\left(\bar{\mathbf{Y}}\mathbf{X}^{\top} - \bar{\mathbf{Y}}\mathbf{X}^{\dagger}\mathbf{X}\mathbf{X}^{\top}\right)\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right) \\
&= 0.
\end{aligned} \tag{31}$$

Here, in the last step, we used the fact that $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}^{\dagger}\mathbf{X}$ is the orthogonal projection onto the row space of $\mathbf{X}$ and therefore $\mathbf{\Pi}_{\mathbf{X}}\mathbf{X}^{\top} = \mathbf{X}^{\top}$. We are left with the following optimization problem.

$$\min_{\mathbf{\Delta}\in\mathbb{R}^{d_y\times d_x}:\mathrm{rank}(\mathbf{\Delta})\le r} \left\{\left\|\bar{\mathbf{Y}} - \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\mathbf{X}\right\|_{\mathrm{F}}^2 + \left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right\|_{\mathrm{F}}^2\right\}. \tag{32}$$

Since the first term is independent of $\mathbf{\Delta}$, the minimizers of this optimization problem are the same as the minimizers of

$$\min_{\mathbf{\Delta}\in\mathbb{R}^{d_y\times d_x}:\mathrm{rank}(\mathbf{\Delta})\le r} \left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right\|_{\mathrm{F}}^2. \tag{33}$$

Let us rewrite the reformulated objective from (33) using trace

$$\begin{aligned}
\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right\|_{\mathrm{F}}^2 &= \mathrm{Tr}\left(\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right)\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\right)^{\top}\right) \\
&= \mathrm{Tr}\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\mathbf{X}\mathbf{X}^{\top}\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right) \\
&= \mathrm{Tr}\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)(\mathbf{X}\mathbf{X}^{\top})^{\frac{1}{2}}(\mathbf{X}\mathbf{X}^{\top})^{\frac{1}{2}}\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)^{\top}\right) \\
&= \mathrm{Tr}\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)(\mathbf{X}\mathbf{X}^{\top})^{\frac{1}{2}}\left(\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)(\mathbf{X}\mathbf{X}^{\top})^{\frac{1}{2}}\right)^{\top}\right) \\
&= \left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)(\mathbf{X}\mathbf{X}^{\top})^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\
&= n\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{FFT}} - \mathbf{\Delta}\right)\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{34}$$

[7]The expression ranges over all admissible top-r singular-value truncations when the truncation is nonunique.

Therefore, (33) can be further reduced to

$$\min_{\operatorname{rank}(\boldsymbol{\Delta})\leq r} \left\| \left(\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}} - \boldsymbol{\Delta} \right) \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 . \tag{35}$$

By a simple change of variable $\mathbf{M} = \widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$ and $\mathbf{Z} = \boldsymbol{\Delta} \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$, we have that the solutions of the optimization problem in (35) are equivalent to the solutions of

$$\min_{\mathbf{Z}\in\mathbb{R}^{d_y\times d_x}} \|\mathbf{M}-\mathbf{Z}\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \exists \boldsymbol{\Delta}\in\mathbb{R}^{d_y\times d_x} : \mathbf{Z} = \boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}},\ \operatorname{rank}(\boldsymbol{\Delta}) \leq r. \tag{36}$$

Let us also consider the solutions of the simplified optimization problem

$$\mathcal{Z}^\star \coloneqq \operatorname*{arg\,min}_{\mathbf{Z}\in\mathbb{R}^{d_y\times d_x} : \operatorname{rank}(\mathbf{Z})\leq r} \|\mathbf{M}-\mathbf{Z}\|_{\mathrm{F}}^2 . \tag{37}$$

Here $\mathcal{Z}^\star$ denotes the set of best rank-at-most-r approximations of $\mathbf{M}$. By the Eckart-Young theorem (Thm. 10), $\mathcal{Z}^\star$ consists of all the top-r singular-value truncations of $\mathbf{M}$. This set is a singleton if and only if the truncation of $\mathbf{M}$ is unique. Throughout the rest of the proof, $\mathbf{Z}^\star$ denotes an individual element of $\mathcal{Z}^\star$, and $\operatorname{trunc}_r\{\mathbf{M}\}$ denotes an individual truncation, with all admissible choices allowed when the truncation is non-unique.

Full-rank empirical second moment (overdetermined): If $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}$ has full rank, then the mapping $\mathbf{Z} = \boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$ is invertible, and therefore the optimization problem in (36) can be further reduced to

$$\min_{\operatorname{rank}(\mathbf{Z})\leq r} \|\mathbf{M}-\mathbf{Z}\|_{\mathrm{F}}^2 , \tag{38}$$

where we used the fact that

$$\operatorname{rank}(\mathbf{Z}) = \operatorname{rank}\left(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right) = \operatorname{rank}(\boldsymbol{\Delta}) \leq r. \tag{39}$$

By the Eckart–Young theorem (Thm. 10), all minimizers are given by

$$\mathbf{Z}^\star = \operatorname{trunc}_r\{\mathbf{M}\} \Longrightarrow \widehat{\boldsymbol{\Delta}}_{\mathrm{LoRA}} = \mathbf{Z}^\star \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{-1} = \operatorname{trunc}_r\left\{\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{-1}, \tag{40}$$

which proves the first point.

Rank-deficient empirical second moment (underdetermined): First, we prove that any $\mathbf{Z}^\star \in \mathcal{Z}^\star$ also satisfies the constraints of (36). This will show that $\mathcal{Z}^\star$ is a subset of the minimizers of (36). Fix any $\mathbf{Z}^\star \in \mathcal{Z}^\star$, and choose an SVD $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ such that $\mathbf{Z}^\star = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^\top$, where $\mathbf{U}_r, \mathbf{S}_r,$ and $\mathbf{V}_r^\top$ are the truncations of their corresponding full matrices. Then

$$\mathbf{Z}^\star = \operatorname{trunc}_r\{\mathbf{M}\} = \mathbf{U}_r\mathbf{U}_r^\top\mathbf{M} = \mathbf{U}_r\mathbf{U}_r^\top\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}. \tag{41}$$

Define $\tilde{\boldsymbol{\Delta}} \coloneqq \mathbf{U}_r\mathbf{U}_r^\top\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}} \in \mathbb{R}^{d_y\times d_x}$, then

$$\mathbf{Z}^\star = \tilde{\boldsymbol{\Delta}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \qquad \text{and} \qquad \operatorname{rank}(\tilde{\boldsymbol{\Delta}}) \leq \operatorname{rank}(\mathbf{U}_r\mathbf{U}_r^\top) = r. \tag{42}$$

Since every element of $\mathcal{Z}^\star$ is feasible for (36), $\mathcal{Z}^\star$ is also exactly the set of minimizers of that problem. The optimal transformed $\mathbf{Z}^\star$ matrix is unique whenever the truncation is unique. Note that

$$\tilde{\boldsymbol{\Delta}} \coloneqq \mathbf{U}_r\mathbf{U}_r^\top\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}} = \mathbf{U}_r\mathbf{U}_r^\top\bar{\mathbf{Y}}\mathbf{X}^\dagger = \mathbf{U}_r\mathbf{U}_r^\top\bar{\mathbf{Y}}\mathbf{X}^\dagger\mathbf{P}_{\mathbf{X}} = \mathbf{U}_r\mathbf{U}_r^\top\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\mathbf{P}_{\mathbf{X}}. \tag{43}$$

where $\mathbf{P}_{\mathbf{X}}$ is the orthogonal projection onto the column space of $\mathbf{X}$, and we used the identity $\mathbf{X}^\dagger = \mathbf{X}^\dagger\mathbf{P}_{\mathbf{X}}$. Moreover, we have that

$$\mathbf{P}_{\mathbf{X}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^\dagger . \tag{44}$$

Thus,

$$\tilde{\boldsymbol{\Delta}} = \mathbf{U}_r\mathbf{U}_r^\top\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^\dagger = \operatorname{trunc}_r\left\{\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^\dagger = \widehat{\boldsymbol{\Delta}}_{\mathrm{LoRA}}. \tag{45}$$

Therefore the family of *basic* solutions of LoRA adaptation is

$$\mathcal{B}_{\mathrm{LoRA}} \coloneqq \left\{ \mathbf{Z}^\star \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger : \mathbf{Z}^\star \in \mathcal{Z}^\star \right\}, \tag{46}$$

whereas the set of *all* solutions of LoRA adaptation is

$$\mathcal{D}^\star \coloneqq \left\{ \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} : \mathbf{\Delta} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \in \mathcal{Z}^\star, \quad \mathrm{rank}(\mathbf{\Delta}) \leq r \right\}. \tag{47}$$

The corresponding LoRA solutions are $\mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta}$ with $\mathbf{\Delta} \in \mathcal{D}^\star$.

Fix any $\mathbf{\Delta}_{\min} \in \mathcal{D}^\star$, and let

$$\mathbf{Z}_{\min} \coloneqq \mathbf{\Delta}_{\min} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \in \mathcal{Z}^\star. \tag{48}$$

Its corresponding basic update is

$$\tilde{\mathbf{\Delta}}_{\min} \coloneqq \mathbf{Z}_{\min} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \in \mathcal{B}_{\mathrm{LoRA}}. \tag{49}$$

By the construction leading to (45), we get that $\mathbf{Z}_{\min} = \tilde{\mathbf{\Delta}}_{\min} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$. Since both $\mathbf{\Delta}_{\min}$ and $\tilde{\mathbf{\Delta}}_{\min}$ map to $\mathbf{Z}_{\min}$ we get

$$\tilde{\mathbf{\Delta}}_{\min} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{Z}_{\min} = \mathbf{\Delta}_{\min} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}. \tag{50}$$

Let $\mathbf{Q} \coloneqq \mathbf{\Delta}_{\min} - \tilde{\mathbf{\Delta}}_{\min}$. Then

$$\mathbf{Q} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{0}. \tag{51}$$

Therefore, by optimality and the constraint of (35), we have the following conditions on $\mathbf{Q}$:

$$\mathbf{Q} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{0} \qquad \text{and} \qquad \mathrm{rank}\left(\tilde{\mathbf{\Delta}}_{\min} + \mathbf{Q} \right) \leq r. \tag{52}$$

Since $\mathbf{Q} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{0}$, we get $\mathrm{im}(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}) \subseteq \ker(\mathbf{Q})$. Note that

$$\ker^\perp\left(\tilde{\mathbf{\Delta}}_{\min} \right) = \ker^\perp\left(\mathrm{trunc}_r\{\mathbf{M}\} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \right) \subseteq \ker^\perp\left(\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \right) = \mathrm{im}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right). \tag{53}$$

Thus, $\ker^\perp(\tilde{\mathbf{\Delta}}_{\min}) \subseteq \ker(\mathbf{Q})$, namely, the row space of $\mathbf{Q}$ is orthogonal to the row space of $\tilde{\mathbf{\Delta}}_{\min}$. Therefore, for any solution $\mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta}_{\min}$

$$\begin{aligned} \|\mathbf{A} - \mathbf{A}_0\|_{\mathrm{F}}^2 = \|\mathbf{\Delta}_{\min}\|_{\mathrm{F}}^2 = \left\| \tilde{\mathbf{\Delta}}_{\min} + \mathbf{Q} \right\|_{\mathrm{F}}^2 &= \left\| \tilde{\mathbf{\Delta}}_{\min} \right\|_{\mathrm{F}}^2 + \|\mathbf{Q}\|_{\mathrm{F}}^2 + 2 \,\mathrm{Tr}\left(\mathbf{Q} \tilde{\mathbf{\Delta}}_{\min}^\top \right) \\ &= \left\| \tilde{\mathbf{\Delta}}_{\min} \right\|_{\mathrm{F}}^2 + \|\mathbf{Q}\|_{\mathrm{F}}^2 \\ &\geq \left\| \tilde{\mathbf{\Delta}}_{\min} \right\|_{\mathrm{F}}^2 \\ &= \|\widehat{\mathbf{A}}_{\mathrm{LoRA}} - \mathbf{A}_0\|_{\mathrm{F}}^2, \end{aligned} \tag{54}$$

where $\widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ is the LoRA solution given in (7) which corresponds to the update $\tilde{\mathbf{\Delta}}_{\min}$, *i.e.*, $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \tilde{\mathbf{\Delta}}_{\min}$. From here, we deduce that the basic (projected) LoRA solution which takes the form in (7) always has a smaller Euclidean distance to $\mathbf{A}_0$ than the solutions it spans. Thus, the min-Frobenius-norm solutions are given by $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$. Note that if the truncation is non-unique, then there are multiple values of $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$, where some of them must be with the minimal Euclidean distance to $\mathbf{A}_0$, while others may have larger distance.

More generally, any LoRA solution with minimal Schatten-p (quasi) norm for $p \in (0, \infty)$ is given by $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$. To show this, observe that

$$\mathbf{\Delta}_{\min} \mathbf{\Delta}_{\min}^\top = \left(\tilde{\mathbf{\Delta}}_{\min} + \mathbf{Q} \right) \left(\tilde{\mathbf{\Delta}}_{\min} + \mathbf{Q} \right)^\top = \tilde{\mathbf{\Delta}}_{\min} \tilde{\mathbf{\Delta}}_{\min}^\top + \mathbf{Q}\mathbf{Q}^\top \succeq \tilde{\mathbf{\Delta}}_{\min} \tilde{\mathbf{\Delta}}_{\min}^\top. \tag{55}$$

Therefore,

$$\sigma_i(\mathbf{\Delta}_{\min}) = \sqrt{\lambda_i\left(\mathbf{\Delta}_{\min} \mathbf{\Delta}_{\min}^\top \right)} \geq \sqrt{\lambda_i\left(\tilde{\mathbf{\Delta}}_{\min} \tilde{\mathbf{\Delta}}_{\min}^\top \right)} = \sigma_i\left(\tilde{\mathbf{\Delta}}_{\min} \right), \tag{56}$$

Thus,

$$\sum_{i=1}^{d}\sigma_i^p(\tilde{\mathbf{\Delta}}_{\min}) \le \sum_{i=1}^{d}\sigma_i^p(\mathbf{\Delta}_{\min}) \iff \left(\sum_{i=1}^{d}\sigma_i^p(\tilde{\mathbf{\Delta}}_{\min})\right)^{\frac{1}{p}} \le \left(\sum_{i=1}^{d}\sigma_i^p(\mathbf{\Delta}_{\min})\right)^{\frac{1}{p}}. \tag{57}$$

Equivalently,

$$\|\widehat{\mathbf{A}}_{\mathrm{LoRA}} - \mathbf{A}_0\|_{S_p} = \|\tilde{\mathbf{\Delta}}_{\min}\|_{S_p} \le \|\mathbf{\Delta}_{\min}\|_{S_p} = \|\mathbf{A} - \mathbf{A}_0\|_{S_p}, \tag{58}$$

This inequality is strict whenever $\mathbf{Q} \neq \mathbf{0}$, since $\mathrm{Tr}(\mathbf{Q}\mathbf{Q}^\top) = \|\mathbf{Q}\|_{\mathrm{F}}^2 > 0$, so at least one of the singular-value inequalities in (56) is strict. Any minimizer of (LoRA) with minimal Schatten-p distance to $\mathbf{A}_0$ must have $\mathbf{Q} = \mathbf{0}$, which means that it takes the form of $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$. □

Proposition 2. *Suppose the truncation in (7) is unique[8], and let $p \in (0, \infty)$. Then:*

- $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ *minimizes the distance to $\mathbf{A}_0$ w.r.t. any unitarily invariant norm among LoRA's solutions.*
- $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ *is the unique LoRA solution that minimizes the distance to $\mathbf{A}_0$ w.r.t. the Schatten-p norm[9].*

Proof. Let us assume that the rank-r truncation of $\mathbf{M} = \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$ is unique. We denote this truncation by $\mathbf{Z}^\star = \mathrm{trunc}_r\{\mathbf{M}\}$, and thus we get that $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \mathbf{Z}^\star(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}})^\dagger$. Under this assumption, the set of LoRA updates that minimize the objective, which we identify in the proof of Thm. 1 (see (47)), is

$$\mathcal{D}^\star := \left\{\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} : \mathbf{\Delta}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{Z}^\star, \quad \mathrm{rank}(\mathbf{\Delta}) \le r\right\}. \tag{59}$$

For every $\mathbf{\Delta}_{\min} \in \mathcal{D}^\star$ we can write $\mathbf{Q} := \mathbf{\Delta}_{\min} - \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$, and as we saw in the proof of Thm. 1 (see (53) and the discussion below it), we have $\mathbf{Q}\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}^\top = \mathbf{0}$. Therefore, for any solution $\mathbf{A} = \mathbf{A}_0 + \mathbf{\Delta}_{\min}$

$$\begin{aligned}\mathbf{\Delta}_{\min}\mathbf{\Delta}_{\min}^\top &= \left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} + \mathbf{Q}\right)\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} + \mathbf{Q}\right)^\top \\ &= \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}^\top + \mathbf{Q}\mathbf{Q}^\top \\ &\succeq \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}^\top,\end{aligned} \tag{60}$$

where the last inequality follows because $\mathbf{Q}\mathbf{Q}^\top$ is positive semi-definite. Therefore, by eigenvalue monotonicity for positive semi-definite matrices,

$$\sigma_i(\mathbf{\Delta}_{\min}) = \sqrt{\lambda_i\left(\mathbf{\Delta}_{\min}\mathbf{\Delta}_{\min}^\top\right)} \ge \sqrt{\lambda_i\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}^\top\right)} = \sigma_i\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\right) \tag{61}$$

for all $i \in [d]$, where $d = \min\{d_x, d_y\}$, with eigenvalues and singular values arranged in nonincreasing order. For every unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$ there exists a symmetric gauge norm ϕ on the vector of singular values. Specifically, given $\|\cdot\|_{\mathrm{ui}}$, for any $\mathbf{\Delta} = \mathbf{A} - \mathbf{A}_0 \in \mathbb{R}^{d_y \times d_x}$ define

$$\sigma(\mathbf{\Delta}) = [\sigma_1(\mathbf{\Delta}), \sigma_2(\mathbf{\Delta}), \ldots, \sigma_d(\mathbf{\Delta})]^\top \in \mathbb{R}^d. \tag{62}$$

Then, we have

$$\|\mathbf{\Delta}\|_{\mathrm{ui}} = \phi(\sigma(\mathbf{\Delta})), \tag{63}$$

where $\phi\colon \mathbb{R}^d \to \mathbb{R}_+$ is a *symmetric norm* on $\mathbb{R}^d$. Let $\mathbf{u} = \sigma(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}})$ and $\mathbf{v} = \sigma(\mathbf{\Delta}_{\min})$, and consider the dual norm ϕ^* of ϕ, for which

$$\phi(\mathbf{q}) = \max_{\mathbf{z}\colon\, \phi^*(\mathbf{z}) \le 1} \mathbf{z}^\top\mathbf{q}. \tag{64}$$

Using the fact that $0 \le \mathbf{u}_i \le \mathbf{v}_i$, for any $\mathbf{z}$ satisfying $\phi^*(\mathbf{z}) \le 1$ we have

$$\mathbf{z}^\top\mathbf{u} \le |\mathbf{z}|^\top\mathbf{u} \le |\mathbf{z}|^\top\mathbf{v} \le \phi^*(|\mathbf{z}|)\phi(\mathbf{v}) = \phi^*(\mathbf{z})\phi(\mathbf{v}) \le \phi(\mathbf{v}), \tag{65}$$

where:

[8] A rank-r truncation $\mathrm{trunc}_r\{\mathbf{M}\}$ is *not* unique if and only if $\sigma_r(\mathbf{M}) = \sigma_{r+1}(\mathbf{M}) > 0$.

[9] Note that for $p \in (0, 1)$ the Schatten-p is a quasi norm.

- In the first step we used $\mathbf{u}_i \geq 0$.
- In the second step we used $|\mathbf{z}_i| \geq 0$ and $\mathbf{u}_i \leq \mathbf{v}_i$, as $|\mathbf{z}_i|$ denotes the elementwise absolute value.
- In the third step we used Hölder's inequality for the norm ϕ and its dual.
- In the penultimate step we used the fact that the dual of a symmetric norm is again a symmetric norm: it is invariant under coordinatewise sign changes.

Maximizing over $\mathbf{z}$ such that $\phi^*(\mathbf{z}) \leq 1$ we get

$$\|\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\|_{\mathrm{ui}} = \phi(\sigma(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}})) = \max_{\mathbf{z}\colon \phi^*(\mathbf{z})\leq 1} \mathbf{z}^\top \mathbf{u} \leq \phi(\mathbf{v}) = \|\mathbf{\Delta}_{\min}\|_{\mathrm{ui}} \tag{66}$$

Overall, for any LoRA solution $\mathbf{A}$ we have

$$\|\widehat{\mathbf{A}}_{\mathrm{LoRA}} - \mathbf{A}_0\|_{\mathrm{ui}} = \|\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\|_{\mathrm{ui}} \leq \|\mathbf{\Delta}_{\min}\|_{\mathrm{ui}} = \|\mathbf{A} - \mathbf{A}_0\|_{\mathrm{ui}}, \tag{67}$$

which proves the first point.

For the second point, we denote the Schatten-p (quasi) norm for $p \in (0, \infty)$ as

$$\|\mathbf{\Delta}\|_{S_p} \coloneqq \left(\sum_{i=1}^{d} \sigma_i^p(\mathbf{\Delta})\right)^{\frac{1}{p}}, \tag{68}$$

while $\|\mathbf{\Delta}\|_{S_\infty} \coloneqq \sigma_1(\mathbf{\Delta})$. The coordinatewise inequality from (61) between the singular values gives

$$\forall i \in [d] \qquad \sigma_i^p\big(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\big) \leq \sigma_i^p\big(\mathbf{\Delta}_{\min}\big). \tag{69}$$

Therefore, by summing across all singular values, we get for any $p \in (0, \infty)$

$$\sum_{i=1}^{d} \sigma_i^p\big(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\big) \leq \sum_{i=1}^{d} \sigma_i^p\big(\mathbf{\Delta}_{\min}\big) \iff \left(\sum_{i=1}^{d} \sigma_i^p\big(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\big)\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{d} \sigma_i^p\big(\mathbf{\Delta}_{\min}\big)\right)^{\frac{1}{p}}. \tag{70}$$

Equivalently,

$$\|\widehat{\mathbf{A}}_{\mathrm{LoRA}} - \mathbf{A}_0\|_{S_p} = \|\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\|_{S_p} \leq \|\mathbf{\Delta}_{\min}\|_{S_p} = \|\mathbf{A} - \mathbf{A}_0\|_{S_p}, \tag{71}$$

This inequality is strict whenever $\mathbf{Q} \neq \mathbf{0}$, since $\mathrm{Tr}(\mathbf{Q}\mathbf{Q}^\top) = \|\mathbf{Q}\|_{\mathrm{F}}^2 > 0$, so at least one of the singular-value inequalities in (61) is strict. Hence $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ is the unique LoRA minimizer that minimizes the Schatten-p distance. Note that this inequality also holds for $p = \infty$, however uniqueness is not guaranteed in this case. □

Theorem 8. *Let $\mathbf{\Delta} = \prod_{i=1}^{m} \mathbf{W}_i$ be a depth $m \geq 1$ matrix factorization[10] with $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$, where $d_0 = d_x$, $d_m = d_y$ and $r = \min\{d_0, d_1, \ldots, d_m\}$. Consider GD iterates $\{(\mathbf{W}_1^t, \mathbf{W}_2^t, \ldots, \mathbf{W}_m^t)\}_t$ on the reparametrized LoRA objective*

$$\mathcal{L}\big(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m\big) = \Big\|\Big(\mathbf{A}_0 + \prod_{i=1}^{m} \mathbf{W}_i\Big)\mathbf{X} - \mathbf{Y}\Big\|_{\mathrm{F}}^2, \tag{72}$$

with $\mathbf{W}_1^{t=0} = \mathbf{0}$ and $\mathbf{W}_i^{t=0} \neq \mathbf{0}$ for all $i \in \{2, \ldots, m\}$. If GD converges to a global minimizer, then $\lim_{t\to\infty} \prod_{i=1}^{m} \mathbf{W}_i^t$ is given by $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$, for some top-$r$ singular-value truncation.

Proof. Let $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{A}_0\mathbf{X}$, then

$$\mathcal{L}\big(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m\big) = \Big\|\prod_{i=1}^{m} \mathbf{W}_i\mathbf{X} - \bar{\mathbf{Y}}\Big\|_{\mathrm{F}}^2. \tag{73}$$

At iteration $t \in \mathbb{N}$, the gradient with respect to the input-side factor $\mathbf{W}_1$ (first layer) is

$$\nabla_{\mathbf{W}_1} \mathcal{L}\big(\mathbf{W}_1^t, \mathbf{W}_2^t, \ldots, \mathbf{W}_m^t\big) = 2\Big(\prod_{i=2}^{m} \mathbf{W}_i^t\Big)^\top \Big(\prod_{i=1}^{m} \mathbf{W}_i^t\mathbf{X} - \bar{\mathbf{Y}}\Big)\mathbf{X}^\top. \tag{74}$$

[10]The product $\prod_{i=p}^{m} \mathbf{W}_i$ for $p \leq m$ is defined as $\mathbf{W}_m \cdots \mathbf{W}_{p+1}\mathbf{W}_p$, whereas for $p > m$, $\prod_{i=p}^{m} \mathbf{W}_i \coloneqq \mathbf{I}$.

Let $\mathbf{P}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^{\dagger}$ be the orthogonal projection onto the column space of $\mathbf{X}$, and $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I}_{d_x} - \mathbf{P}_{\mathbf{X}}$ be the orthogonal projection onto its complement. Thus, for any step size η_t,

$$\mathbf{W}_1^{t+1}\mathbf{P}_{\mathbf{X}}^{\perp} = \Big(\mathbf{W}_1^t - \eta_t \nabla_{\mathbf{W}_1}\mathcal{L}\big(\mathbf{W}_1^t, \mathbf{W}_2^t, \ldots, \mathbf{W}_m^t\big)\Big)\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{W}_1^t\mathbf{P}_{\mathbf{X}}^{\perp}. \tag{75}$$

where we used $\mathbf{X}^{\top}\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{0}$. Since $\mathbf{W}_1^{t=0} = \mathbf{0}$, we have that $\mathbf{W}_1^t\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{0}$, and hence $\mathbf{W}_1^t = \mathbf{W}_1^t\mathbf{P}_{\mathbf{X}}$ for every $t \in \mathbb{N}$. Suppose GD converges to a global minimizer, and denote the limiting adaptation by

$$\mathbf{\Delta}_{\infty} \coloneqq \lim_{t\to\infty}\prod_{i=1}^{m}\mathbf{W}_i^t. \tag{76}$$

Thus,

$$\mathbf{\Delta}_{\infty}\mathbf{P}_{\mathbf{X}} = \left(\lim_{t\to\infty}\prod_{i=1}^{m}\mathbf{W}_i^t\right)\mathbf{P}_{\mathbf{X}} = \lim_{t\to\infty}\prod_{i=2}^{m}\mathbf{W}_i^t\left(\mathbf{W}_1^t\mathbf{P}_{\mathbf{X}}\right) = \lim_{t\to\infty}\prod_{i=1}^{m}\mathbf{W}_i^t = \mathbf{\Delta}_{\infty}. \tag{77}$$

In this factorized setting, the product given by $\mathbf{\Delta}$ is an adaptation of rank at most r. Therefore, the feasible set of adaptations contains only matrices $\mathbf{\Delta}$ with rank at most r. At the same time, every matrix with rank at most r admits a factorization with the prescribed layer dimensions. Hence the factorized problem in (72) and LoRA have the same optimal adaptations. Recall from (47) that the set of optimal rank at most r adaptations of LoRA is given by

$$\mathcal{D}^{\star} \coloneqq \left\{\mathbf{\Delta} \in \mathbb{R}^{d_y\times d_x} \colon \mathbf{\Delta}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \in \mathcal{Z}^{\star}, \quad \operatorname{rank}(\mathbf{\Delta}) \le r\right\}, \tag{78}$$

where $\mathcal{Z}^{\star}$ is the set of all possible top-r SVD truncations of $\mathbf{M} = \widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$ (see (37)). Since GD converges to a global minimizer of the factorized problem, the preceding equivalence gives $\mathbf{\Delta}_{\infty} \in \mathcal{D}^{\star}$. Let $\mathbf{Z}^{\star}$ be the corresponding transformed solution such that $\mathbf{\Delta}_{\infty}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{Z}^{\star}$ and $\mathbf{Z}^{\star} = \operatorname{trunc}_r\{\mathbf{M}\}$ be its specific truncation. Therefore,

$$\mathbf{Z}^{\star}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{\dagger} = \mathbf{\Delta}_{\infty}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{\dagger} = \mathbf{\Delta}_{\infty}\mathbf{P}_{\mathbf{X}} = \mathbf{\Delta}_{\infty}. \tag{79}$$

Overall, we get that

$$\mathbf{\Delta}_{\infty} = \mathbf{Z}^{\star}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{\dagger} = \operatorname{trunc}_r\left\{\widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)^{\dagger} \tag{80}$$

Therefore, $\mathbf{\Delta}_{\infty}$ is given by $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ from (7) for some top-r singular-value truncation. □

Theorem 9. *Let $\mathbf{\Delta} = \prod_{i=1}^{m}\mathbf{W}_i$ be a depth $m \ge 1$ matrix factorization with $\mathbf{W}_i \in \mathbb{R}^{d_i\times d_{i-1}}$, where $d_0 = d_x,\ d_m = d_y$ and $r = \min\{d_0, d_1, \ldots, d_m\}$. For every $\lambda > 0$, let $(\hat{\mathbf{W}}_1^{\lambda}, \hat{\mathbf{W}}_2^{\lambda}, \ldots, \hat{\mathbf{W}}_m^{\lambda})$ be a global minimizer of the L2 regularized LoRA objective, i.e.,*

$$\min_{\mathbf{W}_1,\mathbf{W}_2,\ldots,\mathbf{W}_m}\left\{\left\|\Big(\mathbf{A}_0 + \prod_{i=1}^{m}\mathbf{W}_i\Big)\mathbf{X} - \mathbf{Y}\right\|_{\mathrm{F}}^2 + \lambda\sum_{i=1}^{m}\|\mathbf{W}_i\|_{\mathrm{F}}^2\right\}, \tag{81}$$

and denote $\widehat{\mathbf{A}}_{\lambda} = \mathbf{A}_0 + \prod_{i=1}^{m}\hat{\mathbf{W}}_i^{\lambda}$. Every accumulation point of $\widehat{\mathbf{A}}_{\lambda}$ as $\lambda \to 0^+$ has the form of $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ in (7) for some admissible top-r singular-value truncation. If $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ is unique, then

$$\lim_{\lambda\to 0^+}\hat{\mathbf{A}}_{\lambda} = \widehat{\mathbf{A}}_{\mathrm{LoRA}}. \tag{82}$$

Proof. The loss function

$$\mathcal{L}\big(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m\big) = \left\|\Big(\mathbf{A}_0 + \prod_{i=1}^{m}\mathbf{W}_i\Big)\mathbf{X} - \mathbf{Y}\right\|_{\mathrm{F}}^2 \tag{83}$$

is proper, continuous, and nonnegative. Moreover, in this factorized setting, the product given by $\mathbf{\Delta}$ is an adaptation of rank at most r. Therefore, the feasible set of adaptations contains only matrices $\mathbf{\Delta}$ with rank at most r. At the same time, every matrix with rank at most r admits a factorization with the prescribed layer dimensions. Hence the factorized problem in (83) and LoRA have the same

optimal adaptations. Recall from (47) that the set of optimal rank at most r adaptations of LoRA is given by

$$\mathcal{D}^\star := \left\{ \mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x} : \mathbf{\Delta} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \in \mathcal{Z}^\star, \quad \operatorname{rank}(\mathbf{\Delta}) \le r \right\}, \tag{84}$$

where $\mathcal{Z}^\star$ is the set of all possible top-r singular value truncations of $\mathbf{M} = \widehat{\mathbf{\Delta}}_{\mathrm{FFT}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}$ (see (37)). Therefore, the set of minimizers of (83) is given by

$$\mathcal{S} = \bigcup_{\mathbf{\Delta} \in \mathcal{D}^\star} \mathcal{S}_{\mathbf{\Delta}} \quad \text{such that} \quad \mathcal{S}_{\mathbf{\Delta}} = \left\{ (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m) : \prod_{i=1}^{m} \mathbf{W}_i = \mathbf{\Delta} \right\}. \tag{85}$$

It is easy to verify that $\mathcal{S}$ is nonempty. Therefore, by Thm. 11, any accumulation point of $(\hat{\mathbf{W}}_1^\lambda, \hat{\mathbf{W}}_2^\lambda, \ldots, \hat{\mathbf{W}}_m^\lambda)$ as $\lambda \to 0^+$ belongs to the set

$$\mathcal{S}^\star = \operatorname*{arg\,min}_{(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m) \in \mathcal{S}} \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{W}_i\|_{\mathrm{F}}^2 . \tag{86}$$

Using (85), the value of this optimization problem is given by

$$\min_{\mathbf{\Delta} \in \mathcal{D}^\star} \; \min_{\substack{\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}},\ i \in [m] \\ \mathbf{W}_m \cdots \mathbf{W}_1 = \mathbf{\Delta}}} \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{W}_i\|_{\mathrm{F}}^2 . \tag{87}$$

The minimal value of the inner minimization problem is given by Thm. 12, which is

$$\min_{\substack{\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}},\ i \in [m] \\ \mathbf{W}_m \cdots \mathbf{W}_1 = \mathbf{\Delta}}} \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{W}_i\|_{\mathrm{F}}^2 = \|\mathbf{\Delta}\|_{S_p}^p , \tag{88}$$

with $p = 2/m$. Thus, defining

$$\mathcal{T}_p := \operatorname*{arg\,min}_{\mathbf{\Delta} \in \mathcal{D}^\star} \|\mathbf{\Delta}\|_{S_p}^p, \tag{89}$$

we obtain

$$\left\{ \prod_{i=1}^{m} \mathbf{W}_i : (\mathbf{W}_1, \ldots, \mathbf{W}_m) \in \mathcal{S}^\star \right\} = \mathcal{T}_p. \tag{90}$$

Indeed, every tuple in $\mathcal{S}^\star$ must attain the minimum representation cost of its product, and its product must minimize that cost over $\mathcal{D}^\star$. Conversely, every element of $\mathcal{T}_p$ admits a factorization attaining its representation cost, and every such factorization belongs to $\mathcal{S}^\star$. Let

$$\widehat{\mathbf{\Delta}}_\lambda := \prod_{i=1}^{m} \hat{\mathbf{W}}_i^\lambda. \tag{91}$$

By Thm. 11, the family of factor tuples $(\hat{\mathbf{W}}_1^\lambda, \ldots, \hat{\mathbf{W}}_m^\lambda)$ is uniformly bounded. Hence $\widehat{\mathbf{\Delta}}_\lambda$ is also bounded. Let $\overline{\mathbf{\Delta}}$ be any accumulation point of $\widehat{\mathbf{\Delta}}_\lambda$ as $\lambda \to 0^+$, and choose a sequence $\lambda_k \to 0^+$ such that $\widehat{\mathbf{\Delta}}_{\lambda_k} \to \overline{\mathbf{\Delta}}$. By boundedness of the factor tuples, there exists a further subsequence along which

$$\left(\hat{\mathbf{W}}_1^{\lambda_{k_j}}, \ldots, \hat{\mathbf{W}}_m^{\lambda_{k_j}}\right) \longrightarrow \left(\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_m\right) \in \mathcal{S}^\star. \tag{92}$$

By continuity of matrix multiplication and (90),

$$\overline{\mathbf{\Delta}} = \prod_{i=1}^{m} \overline{\mathbf{W}}_i \in \mathcal{T}_p. \tag{93}$$

By the discussion following (58), every element of $\mathcal{T}_p$ is a basic LoRA update that is given by $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$. Consequently, every accumulation point of $\widehat{\mathbf{A}}_\lambda = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_\lambda$ as $\lambda \to 0^+$ takes the form of $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ for some admissible top-r singular-value truncation.

If $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ is unique, then according to Prop. 2, $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ is the unique minimizer of optimization problem in the definition of $\mathcal{T}_p$, so in this case

$$\left\{ \prod_{i=1}^{m} \mathbf{W}_i : (\mathbf{W}_1, \ldots, \mathbf{W}_m) \in \mathcal{S}^\star \right\} = \left\{ \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} \right\}. \tag{94}$$

Assume by contradiction that $\widehat{\mathbf{\Delta}}_\lambda$ does not converge to $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ as $\lambda \to 0^+$. Then there would exist $\varepsilon > 0$ and a sequence $\lambda_k \to 0^+$ such that

$$\|\widehat{\mathbf{\Delta}}_{\lambda_k} - \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\| \geq \varepsilon \qquad \text{for every } k. \tag{95}$$

Boundedness of $\widehat{\mathbf{\Delta}}_{\lambda_k}$ gives a convergent subsequence of this sequence. By the preceding argument, its limit belongs to $\{\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\}$. Since $\{\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\}$ is a singleton, the limit must equal $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$, contradicting $\|\widehat{\mathbf{\Delta}}_{\lambda_k} - \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}\| \geq \varepsilon$ along the convergent subsequence. Therefore, in this case

$$\lim_{\lambda \to 0^+} \widehat{\mathbf{\Delta}}_\lambda = \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}. \tag{96}$$

Consequently, $\widehat{\mathbf{A}}_\lambda = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_\lambda \longrightarrow \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \widehat{\mathbf{A}}_{\mathrm{LoRA}}$ as $\lambda \to 0^+$. □

V Excess risk analysis for full fine-tuning

In this section, we provide the analysis of the excess risk for FFT.

V.1 Overdetermined regime

Theorem 4 (FFT excess risk)**.** *Assume* $\operatorname{rank}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\right) = d_x$ *almost surely. Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] = \frac{1}{n}\operatorname{Tr}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\operatorname{Tr}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \geq \bar{\sigma}_{\boldsymbol{\varepsilon}}^2 \frac{d_x d_y}{n}, \tag{11}$$

where $\bar{\sigma}_{\boldsymbol{\varepsilon}}^2 \coloneqq \mathbb{E}[\|\boldsymbol{\varepsilon}\|^2]/d_y$ *(average noise variance per coordinate).*

Proof. Lemma 6 states that for any estimator $\mathbf{A}$ we have

$$\mathcal{R}(\mathbf{A}) - \mathcal{R}(\mathbf{A}^\star) = \operatorname{Tr}\left((\mathbf{A} - \mathbf{A}^\star)^\top (\mathbf{A} - \mathbf{A}^\star)\mathbf{\Sigma}_{\mathbf{xx}}\right). \tag{97}$$

From Prop. 1, the FFT is given by

$$\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{A}_0 + (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\,\mathbf{X}^\dagger. \tag{98}$$

In the overdetermined case, when $n \geq d_x$, $\operatorname{rank}(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}})$ can have a full rank with probability one. In such a case, $\mathbf{X}$ has full row rank, and $\mathbf{X}^\dagger$ has the closed-form expression

$$\mathbf{X}^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}. \tag{99}$$

Therefore,

$$\begin{aligned}
\widehat{\mathbf{A}}_{\mathrm{FFT}} &= \mathbf{A}_0 + \mathbf{Y}\mathbf{X}^\dagger - \mathbf{A}_0\mathbf{X}\mathbf{X}^\dagger \\
&= \mathbf{A}_0 + \mathbf{Y}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} - \mathbf{A}_0\mathbf{X}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} \\
&= \mathbf{Y}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}.
\end{aligned} \tag{100}$$

Note that in the overdetermined regime, the fine-tuned model does not depend on the pre-trained values. Namely, this estimator is the ordinary least squares (OLS) estimator. From the data generation model in (1), we have that $\mathbf{Y} = \mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi}$. Using this, we get

$$\begin{aligned}
\widehat{\mathbf{A}}_{\mathrm{FFT}} &= (\mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi})\,\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} \\
&= \mathbf{A}^\star\mathbf{X}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} + \boldsymbol{\xi}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} \\
&= \mathbf{A}^\star + \boldsymbol{\xi}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}.
\end{aligned} \tag{101}$$

Thus, $\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star = \boldsymbol{\xi}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$. Placing this in (97) yields

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{FFT}})\right] - \mathcal{R}(\mathbf{A}^\star) &= \mathbb{E}\left[\operatorname{Tr}\left(\left(\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star\right)^\top \left(\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}\right)\right] \\
&= \mathbb{E}\left[\operatorname{Tr}\left((\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbb{E}[\boldsymbol{\xi}^\top\boldsymbol{\xi} \mid \mathbf{X}]\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{\Sigma}_{\mathbf{xx}}\right)\right].
\end{aligned} \tag{102}$$

Computing $\mathbb{E}[\boldsymbol{\xi}^\top \boldsymbol{\xi} \mid \mathbf{X}]$, we have

$$\mathbb{E}\left[\left(\boldsymbol{\xi}^\top \boldsymbol{\xi}\right)_{ij} \,\middle|\, \mathbf{X}\right] = \mathbb{E}\left[\boldsymbol{\varepsilon}_i^\top \boldsymbol{\varepsilon}_j\right] = \begin{cases} \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\varepsilon}_i^\top \boldsymbol{\varepsilon}_i)\right] = \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top)\right] = \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right) & i = j \\ 0 & i \neq j \end{cases}, \tag{103}$$

where we used the fact that noise vectors of different samples $i \neq j$ are independent. Therefore $\mathbb{E}[\boldsymbol{\xi}^\top \boldsymbol{\xi} \mid \mathbf{X}] = \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\mathbf{I}$. Then

$$\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{FFT}})\right] - \mathcal{R}(\mathbf{A}^\star) = \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\mathbb{E}\left[\mathrm{Tr}\left((\mathbf{X}\mathbf{X}^\top)^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right]. \tag{104}$$

We can further simplify the term $\mathbf{X}\mathbf{X}^\top$,

$$\mathbf{X}\mathbf{X}^\top = n\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top\right) = n\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}. \tag{105}$$

Plugging this back into (102), we have

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{FFT}})\right] - \mathcal{R}(\mathbf{A}^\star) &= \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\mathrm{Tr}\left(\left(n\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right] \\ &= \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right]. \end{aligned} \tag{106}$$

Furthermore, note that using Jensen's inequality, we get

$$\begin{aligned} \mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{\frac{1}{2}}\right)\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\right)^{-1}\right)\right] \geq \mathrm{Tr}\left(\left(\mathbb{E}\left[\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\right]\right)^{-1}\right), \end{aligned} \tag{107}$$

where we used the fact that $f(\boldsymbol{\Sigma}) = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1})$ is convex over the set of positive-definite matrices. Then

$$\mathbb{E}\left[\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\right] = \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\mathbb{E}\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}\right]\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}} = \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}} = \mathbf{I}_{d_x}. \tag{108}$$

Therefore,

$$\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right] \geq \mathrm{Tr}(\mathbf{I}) = d_x. \tag{109}$$

Denote $\bar{\sigma}_{\boldsymbol{\varepsilon}}^2 = \mathbb{E}\left[\|\boldsymbol{\varepsilon}\|^2\right]/d_y = \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})/d_y$, then

$$\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{FFT}})\right] - \mathcal{R}(\mathbf{A}^\star) = \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right] \geq \bar{\sigma}_{\boldsymbol{\varepsilon}}^2\frac{d_y d_x}{n}. \tag{110}$$

□

Proposition 3 (Gaussian case)**.** *Assume* $n > d_x + 1$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}})$ *i.i.d.,* $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \succeq \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}} \succ \mathbf{0}$. *Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] = \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\frac{d_x}{n - d_x - 1}. \tag{12}$$

Proof. Recall from Theorem 4 that

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{\mathbf{A}}_{\mathrm{FFT}}\right)\right] - \mathcal{R}\left(\mathbf{A}^\star\right) = \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right]. \tag{111}$$

Since $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}})$ i.i.d., we can write $\mathbf{x}_i = \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{\frac{1}{2}}\mathbf{z}_i$ where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{d_x \times n}$. Then

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top = \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{\frac{1}{2}}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{\frac{1}{2}}. \tag{112}$$

Using the cyclicity of the trace, we obtain

$$\begin{aligned} \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right) &= \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right) \\ &= \mathrm{Tr}\left(\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^{-1}\right). \end{aligned} \tag{113}$$

Since $\mathbf{Z}\mathbf{Z}^\top \sim \mathcal{W}_{d_x}(\mathbf{I}, n)$ is a Wishart matrix and $n > d_x + 1$, it holds that (see Bach (2024, Sec. 3.8.1))

$$\mathbb{E}\left[\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^{-1}\right] = \frac{n}{n-d_x-1}\mathbf{I}, \tag{114}$$

and therefore

$$\mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}\right)\right] = \mathrm{Tr}\left(\frac{n}{n-d_x-1}\mathbf{I}\right) = \frac{nd_x}{n-d_x-1}. \tag{115}$$

Substituting this back yields

$$\mathbb{E}\left[\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big)\right] - \mathcal{R}\big(\mathbf{A}^\star\big) = \frac{1}{n}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}\right)\frac{nd_x}{n-d_x-1}, \tag{116}$$

which concludes the proof. □

V.2 Underdetermined regime

Theorem 6 (FFT excess risk)**.** *Let $\widehat{\mathbf{A}}_{\mathrm{FFT}}$ be the FFT solution (6), then*

$$\mathbb{E}\left[\mathcal{E}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big)\right] = \underbrace{\mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}_{\mathbf{X}}^{\perp}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right]}_{\textit{FFT Bias}} + \underbrace{\frac{1}{n}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}\right)\mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right]}_{\textit{FFT Variance}}. \tag{15}$$

Proof. By Lemma 6 we get

$$\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big) - \mathcal{R}\big(\mathbf{A}^\star\big) = \mathrm{Tr}\left(\big(\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star\big)^\top\big(\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star\big)\boldsymbol{\Sigma}_{\mathbf{xx}}\right). \tag{117}$$

From Prop. 1, the FFT is given by

$$\widehat{\mathbf{A}}_{\mathrm{FFT}} = \mathbf{A}_0 + \left(\mathbf{Y} - \mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger, \tag{118}$$

Using $\mathbf{Y} = \mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi}$ we get

$$\begin{aligned}
\widehat{\mathbf{A}}_{\mathrm{FFT}} - \mathbf{A}^\star &= \mathbf{A}_0 + \left(\mathbf{Y} - \mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger - \mathbf{A}^\star \\
&= \mathbf{A}_0 + \left(\mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi} - \mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger - \mathbf{A}^\star \\
&= \left(\boldsymbol{\Delta}^\star\mathbf{X} + \boldsymbol{\xi}\right)\mathbf{X}^\dagger - \boldsymbol{\Delta}^\star \\
&= \boldsymbol{\Delta}^\star\mathbf{X}\mathbf{X}^\dagger + \boldsymbol{\xi}\mathbf{X}^\dagger - \boldsymbol{\Delta}^\star \\
&= \boldsymbol{\Delta}^\star\mathbf{P}_{\mathbf{X}} - \boldsymbol{\Delta}^\star + \boldsymbol{\xi}\mathbf{X}^\dagger \\
&= -\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \boldsymbol{\xi}\mathbf{X}^\dagger,
\end{aligned} \tag{119}$$

where we used the relation $\boldsymbol{\Delta}^\star = \mathbf{A}^\star - \mathbf{A}_0$ from (3), and denoted the orthogonal projection matrix onto the column space of $\mathbf{X}$ by $\mathbf{P}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\dagger$. Plugging this into the excess-risk formula gives

$$\begin{aligned}
\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big) - \mathcal{R}\big(\mathbf{A}^\star\big) &= \mathrm{Tr}\left(\left(-\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \boldsymbol{\xi}\mathbf{X}^\dagger\right)^\top\left(-\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \boldsymbol{\xi}\mathbf{X}^\dagger\right)\boldsymbol{\Sigma}_{\mathbf{xx}}\right) \\
&= \mathrm{Tr}\left(\left(-\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \boldsymbol{\xi}\mathbf{X}^\dagger\right)\boldsymbol{\Sigma}_{\mathbf{xx}}\left(-\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) + \boldsymbol{\xi}\mathbf{X}^\dagger\right)^\top\right) \\
&= \mathrm{Tr}(\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}) - 2\,\mathrm{Tr}(\boldsymbol{\xi}\mathbf{X}^\dagger\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})^\top\boldsymbol{\Delta}^{\star\top}) \\
&\quad + \mathrm{Tr}\left(\boldsymbol{\xi}\mathbf{X}^\dagger\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^\dagger\right)^\top\boldsymbol{\xi}^\top\right).
\end{aligned} \tag{120}$$

Let us take a conditional expectation over the training noise matrix $\boldsymbol{\xi}$ given the training data $\mathbf{X}$. Since the noise has zero mean, and it is independant of $\mathbf{X}$,we get

$$\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\xi}\mathbf{X}^\dagger\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})^\top\boldsymbol{\Delta}^{\star\top}) \;\middle|\; \mathbf{X}\right] = \mathrm{Tr}(\mathbb{E}\left[\boldsymbol{\xi} \mid \mathbf{X}\right]\mathbf{X}^\dagger\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})^\top\boldsymbol{\Delta}^{\star\top}) = 0. \tag{121}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big) \;\middle|\; \mathbf{X}\right] - \mathcal{R}\big(\mathbf{A}^\star\big) &= \mathrm{Tr}\left(\boldsymbol{\Delta}^\star(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right) \\
&\quad + \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\xi}\mathbf{X}^\dagger\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^\dagger\right)^\top\boldsymbol{\xi}^\top\right) \;\middle|\; \mathbf{X}\right].
\end{aligned} \tag{122}$$

Now, for the second term, we have

$$\begin{aligned}\mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\boldsymbol{\xi}^{\top}\right)\,\middle|\,\mathbf{X}\right] &= \mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{\xi}^{\top}\boldsymbol{\xi}\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\right)\,\middle|\,\mathbf{X}\right]\\ &= \operatorname{Tr}\left(\mathbb{E}\left[\boldsymbol{\xi}^{\top}\boldsymbol{\xi}\mid\mathbf{X}\right]\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\right)\\ &= \operatorname{Tr}\left(\left(\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\mathbf{I}\right)\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\right)\\ &= \operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\operatorname{Tr}\left(\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\right),\end{aligned} \tag{123}$$

where we used the fact that $\mathbb{E}\left[\boldsymbol{\xi}^{\top}\boldsymbol{\xi}\mid\mathbf{X}\right]=\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\mathbf{I}$. Note that

$$\operatorname{Tr}\left(\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\right)=\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\mathbf{X}^{\dagger}\right)=\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\top}\right)^{\dagger}\mathbf{X}^{\dagger}\right). \tag{124}$$

Using $(\mathbf{X}^{\top})^{\dagger}\mathbf{X}^{\dagger}=(\mathbf{X}\mathbf{X}^{\top})^{\dagger}$, we have

$$\mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{X}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}^{\dagger}\right)^{\top}\boldsymbol{\xi}^{\top}\right)\,\middle|\,\mathbf{X}\right]=\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\left(\mathbf{X}\mathbf{X}^{\top}\right)^{\dagger}\right)=\frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right). \tag{125}$$

For now, we reduced (122) into

$$\mathbb{E}\left[\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big)\,\middle|\,\mathbf{X}\right]-\mathcal{R}\big(\mathbf{A}^{\star}\big)=\operatorname{Tr}\left(\boldsymbol{\Delta}^{\star}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right)+\frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right). \tag{126}$$

Let us proceed with the first term,

$$\begin{aligned}\operatorname{Tr}\left(\boldsymbol{\Delta}^{\star}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{xx}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right) &= \operatorname{Tr}\left(\boldsymbol{\Delta}^{\star}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right)\\ &= \operatorname{Tr}\left(\left(\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right)^{\top}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right)\\ &= \left\|\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right\|_{\mathrm{F}}^{2}.\end{aligned} \tag{127}$$

Therefore,

$$\mathbb{E}\left[\mathcal{R}\big(\widehat{\mathbf{A}}_{\mathrm{FFT}}\big)\right]-\mathcal{R}\big(\mathbf{A}^{\star}\big)=\mathbb{E}\left[\left\|\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\Delta}^{\star\top}\right\|_{\mathrm{F}}^{2}\right]+\frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right]. \tag{128}$$

□

Proposition 5 (Gaussian case)**.** *Assume* $n<d_x-1$, $\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\sigma_{\mathbf{x}}^2\mathbf{I})$ *i.i.d., and* $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\succeq\mathbf{0}$*. Then*

$$\textit{FFT Bias}=\sigma_{\mathbf{x}}^2\left\|\boldsymbol{\Delta}^{\star}\right\|_{\mathrm{F}}^2\frac{d_x-n}{d_x},\qquad\textit{and}\qquad\textit{FFT Variance}=\operatorname{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right)\frac{n}{d_x-n-1}. \tag{16}$$

Proof. Since $\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\sigma_{\mathbf{x}}^2\mathbf{I})$ i.i.d., define $\mathbf{z}_i:=\mathbf{x}_i/\sigma_{\mathbf{x}}$ so that $\mathbf{z}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})$. Let $\mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_n]\in\mathbb{R}^{d_x\times n}$. Then the empirical second moment can be written as

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}=\frac{1}{n}\mathbf{X}\mathbf{X}^{\top}=\sigma_{\mathbf{x}}^2\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\top}. \tag{129}$$

Since $n<d_x$, the matrix $\mathbf{Z}\mathbf{Z}^{\top}$ has rank n almost surely, and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}=\frac{1}{\sigma_{\mathbf{x}}^2}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\top}\right)^{\dagger}. \tag{130}$$

Using $\boldsymbol{\Sigma}_{\mathbf{xx}}=\sigma_{\mathbf{x}}^2\mathbf{I}$ and cyclicity of the trace,

$$\operatorname{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\boldsymbol{\Sigma}_{\mathbf{xx}}\right)=\operatorname{Tr}\left(\frac{1}{\sigma_{\mathbf{x}}^2}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\top}\right)^{\dagger}\sigma_{\mathbf{x}}^2\mathbf{I}\right)=\operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\top}\right)^{\dagger}\right). \tag{131}$$

Because $\mathbf{Z}$ has linearly independent columns with probability one, we have that $(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\dagger = (\mathbf{Z}\mathbf{Z}^\top)^\dagger$ and $\mathbf{Z}^\dagger = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$. Therefore,

$$\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^\dagger = n(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\dagger = n\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top. \tag{132}$$

Thus,

$$\operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^\dagger\right) = n\operatorname{Tr}\left(\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\right) = n\operatorname{Tr}\left((\mathbf{Z}^\top\mathbf{Z})^{-1}\right) = \operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right), \tag{133}$$

where we used the cyclic property of the trace. Since $\mathbf{Z}^\top\mathbf{Z} \sim \mathcal{W}_n(\mathbf{I}, d_x)$ with $n+1 < d_x$, it is known that

$$\mathbb{E}\left[\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right] = \frac{n}{d_x-n-1}\mathbf{I}. \tag{134}$$

Therefore,

$$\mathbb{E}\left[\operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^\dagger\right)\right] = \operatorname{Tr}\left(\mathbb{E}\left[\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right]\right) = \frac{n}{d_x-n-1}\operatorname{Tr}(\mathbf{I}) = \frac{n^2}{d_x-n-1}. \tag{135}$$

Substituting back, we obtain

$$\frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\mathbb{E}\left[\operatorname{Tr}\left(\widehat{\boldsymbol{\Sigma}}^\dagger_{\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\right)\right] = \frac{1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\frac{n^2}{d_x-n-1} = \operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\frac{n}{d_x-n-1}, \tag{136}$$

which concludes the first part of the proof. For the bias term, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}^\perp_{\mathbf{X}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathbf{x}\mathbf{x}}\right\|_{\mathrm{F}}^2\right] &= \sigma_{\mathbf{x}}^2\left(\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2 - \mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}_{\mathbf{X}}\right\|_{\mathrm{F}}^2\right]\right) \\
&= \sigma_{\mathbf{x}}^2\left(\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2 - \mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{\Delta}^\star\mathbf{P}_{\mathbf{X}}(\boldsymbol{\Delta}^\star)^\top\right)\right]\right) \\
&= \sigma_{\mathbf{x}}^2\left(\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2 - \operatorname{Tr}\left(\boldsymbol{\Delta}^\star\mathbb{E}[\mathbf{P}_{\mathbf{X}}](\boldsymbol{\Delta}^\star)^\top\right)\right),
\end{aligned} \tag{137}$$

where in the second step we used $\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}}^\top = \mathbf{P}_{\mathbf{X}}$. Using Lemma 8 we have that $\mathbb{E}[\mathbf{P}_{\mathbf{X}}] = \frac{n}{d_x}\mathbf{I}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}^\perp_{\mathbf{X}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathbf{x}\mathbf{x}}\right\|_{\mathrm{F}}^2\right] &= \sigma_{\mathbf{x}}^2\left(\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2 - \operatorname{Tr}\left(\boldsymbol{\Delta}^\star\mathbb{E}[\mathbf{P}_{\mathbf{X}}](\boldsymbol{\Delta}^\star)^\top\right)\right) \\
&= \sigma_{\mathbf{x}}^2\left(\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2 - \frac{n}{d_x}\operatorname{Tr}\left(\boldsymbol{\Delta}^\star(\boldsymbol{\Delta}^\star)^\top\right)\right) \\
&= \sigma_{\mathbf{x}}^2\left\|\boldsymbol{\Delta}^\star\right\|_{\mathrm{F}}^2\frac{d_x-n}{d_x}.
\end{aligned} \tag{138}$$

□

VI Excess risk analysis for LoRA

In this section, we provide detailed proofs for the LoRA solution in both overdetermined and underdetermined regimes. We start by introducing some useful lemmas used only for proofs related to LoRA.

Lemma 1. *We have shown in Lemma 1 that*

$$\widehat{\boldsymbol{\Delta}}_{\mathrm{LoRA}} = \operatorname{trunc}_r\left\{\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}}\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}_{\mathbf{x}\mathbf{x}}\right\}\left(\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}_{\mathbf{x}\mathbf{x}}\right)^\dagger. \tag{139}$$

Given the linear relation,

$$\mathbf{Y} = \mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi}, \tag{140}$$

and,

$$\widehat{\boldsymbol{\Delta}}_{\mathrm{FFT}} = (\mathbf{Y} - \mathbf{A}_0\mathbf{X})\,\mathbf{X}^\dagger. \tag{141}$$

The low-rank update $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ *can be written as,*

$$\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \mathrm{trunc}_r\Big\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}} + \boldsymbol{\xi}\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger. \tag{142}$$

In the overdetermined regime, we can further simplify the above equation to,

$$\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \mathrm{trunc}_r\left\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}} + \frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}^{-\frac{1}{2}}_{\mathbf{xx}}\right\}\widehat{\mathbf{\Sigma}}^{-\frac{1}{2}}_{\mathbf{xx}}. \tag{143}$$

Proof. Recall that,

$$\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} = \mathrm{trunc}_r\Big\{\widehat{\mathbf{\Delta}}_{\mathrm{FFT}}\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger = \mathrm{trunc}_r\Big\{\left(\mathbf{Y}-\mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger. \tag{144}$$

Given the the linear relation $\mathbf{Y} = \mathbf{A}^\star\mathbf{X} + \boldsymbol{\xi}$, we can simplify the term inside the truncation further,

$$\begin{aligned}
\mathrm{trunc}_r\Big\{\left(\mathbf{Y}-\mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger &= \mathrm{trunc}_r\Big\{\left(\mathbf{A}^\star\mathbf{X}+\boldsymbol{\xi}-\mathbf{A}_0\mathbf{X}\right)\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger \\
&= \mathrm{trunc}_r\Big\{\left(\mathbf{\Delta}^\star\mathbf{X}+\boldsymbol{\xi}\right)\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger \\
&= \mathrm{trunc}_r\Big\{\left(\mathbf{\Delta}^\star\mathbf{X}+\boldsymbol{\xi}\right)\mathbf{X}^\dagger\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}}\Big\}\left(\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}}\right)^\dagger.
\end{aligned} \tag{145}$$

Now we expand the term inside the truncation,

$$\begin{aligned}
\left(\mathbf{\Delta}^\star\mathbf{X}+\boldsymbol{\xi}\right)\mathbf{X}^\dagger\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} &= \mathbf{\Delta}^\star\mathbf{X}\mathbf{X}^\dagger\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} + \boldsymbol{\xi}\mathbf{X}^\dagger\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} \\
&= \mathbf{\Delta}^\star\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} + \boldsymbol{\xi}\mathbf{X}^\dagger\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} \\
&= \sqrt{n}\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}} + \sqrt{n}\boldsymbol{\xi}\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}.
\end{aligned} \tag{146}$$

Where in the second line we used the fact that $\mathbf{P}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\dagger$ is the orthogonal projection onto the column space of $\mathbf{X}$ and it leaves $(\mathbf{X}\mathbf{X}^\top)^{\frac{1}{2}}$ unchanged. Finally, we have this other explicit form for the $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$,

$$\begin{aligned}
\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} &= \mathrm{trunc}_r\Big\{\sqrt{n}\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}} + \sqrt{n}\boldsymbol{\xi}\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\left(\mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}}\right)^\dagger \\
&= \mathrm{trunc}_r\Big\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}} + \boldsymbol{\xi}\mathbf{X}^\dagger\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\Big\}\left(\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}_{\mathbf{xx}}\right)^\dagger.
\end{aligned} \tag{147}$$

Where we used the property that $(c\mathbf{D})^\dagger = c^{-1}\mathbf{D}^\dagger$ for some scaler c and matrix $\mathbf{D}$. For the more simplified case of overdetermined regime where $(\mathbf{X}\mathbf{X}^\top)$ is assumed to be invertible, we get the claimed result just by using the definition of $\mathbf{X}^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$. □

Lemma 2. *Let* $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})$ *for* $i \in [n]$ *i.i.d., and* $\boldsymbol{\xi} \in \mathbb{R}^{d_y\times n}$ *be the matrix containing* $\{\boldsymbol{\varepsilon}_i\}$ *as columns. Let* $\mathbf{X} \in \mathbb{R}^{d_x\times n}$ *be a random matrix independent of* $\boldsymbol{\xi}$ *and let* $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}^\dagger\mathbf{X} \in \mathbb{R}^{n\times n}$ *be the orthogonal projection onto its row space. Then*

$$\mathbb{E}\left[\|\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\|_2^2 \,\middle|\, \mathbf{X}\right] \le 5\max\{d_x, d_y\}\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}). \tag{148}$$

Proof. Since $\boldsymbol{\xi}$ has i.i.d. Gaussian columns and is independent of $\mathbf{X}$, we may represent $\boldsymbol{\xi} = \mathbf{\Sigma}^{\frac{1}{2}}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\mathbf{\Psi}$, where $\mathbf{\Psi} \in \mathbb{R}^{d_y\times n}$ has i.i.d. standard Gaussian entries and is independent of $\mathbf{X}$. Let $\mathbf{X} \in \mathbb{R}^{d_x\times n}$ a random matrix and let $\mathbf{\Pi}_{\mathbf{X}} \in \mathbb{R}^{n\times n}$ be the orthogonal projector onto the row space of $\mathbf{X}$. Define $\tilde{d} \coloneqq \mathrm{rank}(\mathbf{X}) \le d_x$, and note that given $\mathbf{X}$, there exists a matrix $\mathbf{Q} \in \mathbb{R}^{n\times\tilde{d}}$ with orthonormal columns ($\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_{\tilde{d}}$) such that

$$\mathbf{\Pi}_{\mathbf{X}} = \mathbf{Q}\mathbf{Q}^\top. \tag{149}$$

Write

$$\mathbf{\Psi}\mathbf{\Pi}_{\mathbf{X}} = \mathbf{\Psi}\mathbf{Q}\mathbf{Q}^\top = \mathbf{G}\mathbf{Q}^\top, \qquad \text{where} \qquad \mathbf{G} \coloneqq \mathbf{\Psi}\mathbf{Q} \in \mathbb{R}^{d_y\times\tilde{d}}. \tag{150}$$

Since $(\mathbf{G}\mathbf{Q}^\top)(\mathbf{G}\mathbf{Q}^\top)^\top = \mathbf{G}\mathbf{G}^\top$, we have that

$$\|\mathbf{\Psi}\mathbf{\Pi}_\mathbf{X}\|_2 = \left\|\mathbf{G}\mathbf{Q}^\top\right\|_2 = \|\mathbf{G}\|_2 . \tag{151}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\xi}\mathbf{\Pi}_\mathbf{X}\|_2^2 \,\middle|\, \mathbf{X}\right] &= \mathbb{E}\left[\left\|\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{\frac{1}{2}}\mathbf{\Psi}\mathbf{\Pi}_\mathbf{X}\right\|_2^2 \,\middle|\, \mathbf{X}\right] \\ &\le \mathbb{E}\left[\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\,\|\mathbf{\Psi}\mathbf{\Pi}_\mathbf{X}\|_2^2 \,\middle|\, \mathbf{X}\right] \\ &= \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})\mathbb{E}\left[\|\mathbf{G}\|_2^2 \,\middle|\, \mathbf{X}\right]. \end{aligned} \tag{152}$$

Now we show that $\mathbf{G}|\mathbf{X}$ is Gaussian with i.i.d. $\mathcal{N}(0,1)$ entries. To this end, let us look at the vectorization of $\mathbf{G}$, denoted by $\mathrm{vec}(\mathbf{G})$. First, note that $\mathrm{vec}(\mathbf{G}) = \mathrm{vec}(\mathbf{\Psi}\mathbf{Q}) = (\mathbf{Q}^\top \otimes \mathbf{I})\ \mathrm{vec}(\mathbf{\Psi})$, where $\otimes$ is the Kronecker product. Then,

$$\mathbb{E}[\mathrm{vec}(\mathbf{G}) \,|\, \mathbf{X}] = \mathbf{Q}^\top \otimes \mathbf{I}\,\mathbb{E}[\mathrm{vec}(\mathbf{\Psi}) \,|\, \mathbf{X}] = \mathbf{0}, \tag{153}$$

and

$$\begin{aligned}\mathrm{Cov}(\mathrm{vec}(\mathbf{G}) \,|\, \mathbf{X}) &= \mathbb{E}[\mathrm{vec}(\mathbf{G})\,\mathrm{vec}(\mathbf{G})^\top \,|\, \mathbf{X}] \\ &= \mathbb{E}\left[\left(\mathbf{Q}^\top \otimes \mathbf{I}\ \mathrm{vec}(\mathbf{\Psi})\right)\left(\mathbf{Q}^\top \otimes \mathbf{I}\ \mathrm{vec}(\mathbf{\Psi})\right)^\top \,\middle|\, \mathbf{X}\right] \\ &= \mathbb{E}\left[\mathbf{Q}^\top \otimes \mathbf{I}\ \mathrm{vec}(\mathbf{\Psi})\,\mathrm{vec}(\mathbf{\Psi})^\top\left(\mathbf{Q}^\top \otimes \mathbf{I}\right)^\top \,\middle|\, \mathbf{X}\right] \\ &= \mathbf{Q}^\top \otimes \mathbf{I}\,\mathbb{E}\left[\mathrm{vec}(\mathbf{\Psi})\,\mathrm{vec}(\mathbf{\Psi})^\top \,\middle|\, \mathbf{X}\right]\mathbf{Q} \otimes \mathbf{I}. \end{aligned} \tag{154}$$

Observe that $\mathbb{E}[\mathrm{vec}(\mathbf{\Psi})\,\mathrm{vec}(\mathbf{\Psi})^\top \,|\, \mathbf{X}] = \mathbf{I}$, then

$$\mathrm{Cov}(\mathrm{vec}(\mathbf{G}) \,|\, \mathbf{X}) = \left(\mathbf{Q}^\top \otimes \mathbf{I}\right)(\mathbf{Q} \otimes \mathbf{I}) = \left(\mathbf{Q}^\top\mathbf{Q}\right) \otimes \mathbf{I} = \mathbf{I} \otimes \mathbf{I} = \mathbf{I}. \tag{155}$$

Note that conditioning on $\mathbf{X}$, $\mathbf{G} = \mathbf{\Psi}\mathbf{Q}$ is a linear transformation of a Gaussian matrix, so these calculations shows that its entries are i.i.d. standard Gaussian conditioned on $\mathbf{X}$. Applying Lemma 5 under the conditioning on $\mathbf{X}$, we have

$$\mathbb{E}\left[\|\mathbf{G}\|_2^2 \,\middle|\, \mathbf{X}\right] \le 5\mathbb{E}\left[\max\{\tilde{d}, d_y\} \,\middle|\, \mathbf{X}\right] \le 5\max\{d_x, d_y\}. \tag{156}$$

Hence,

$$\mathbb{E}\left[\|\boldsymbol{\xi}\mathbf{\Pi}_\mathbf{X}\|_2^2 \,\middle|\, \mathbf{X}\right] \le 5\max\{d_x, d_y\}\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}). \tag{157}$$

□

VI.1 Overdetermined regime

Theorem 5 (LoRA excess risk)**.** *Assume* $\mathrm{rank}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}\right) = d_x$ *almost surely. Then*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\right)\right] \lesssim\ & \frac{r\max\{d_x, d_y\}}{n}\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right)\mathbb{E}\left[\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \\ &+ r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] + \sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\mathbf{\Delta}^\star\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right). \end{aligned} \tag{13}$$

Proof. Recall from Lemma 6 for any estimator $\mathbf{A}$,

$$\mathcal{R}(\mathbf{A}) - \mathcal{R}(\mathbf{A}^\star) = \left\|\left(\mathbf{A} - \mathbf{A}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2. \tag{158}$$

Let $\mathbf{A} = \widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ where $\mathrm{rank}(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}) \le r$. Then,

$$\begin{aligned}\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{LoRA}}) - \mathcal{R}(\mathbf{A}^\star) = \left\|\left(\mathbf{A} - \mathbf{A}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 &= \left\|\left(\mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{A}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\ &= \left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{159}$$

Using Lemma 9 with $\mathbf{C}^\star = \mathrm{trunc}_r\left\{\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\}$ we have,

$$\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \leq 2r\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 + \left\|\mathbf{\Delta}^\star\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{C}^\star\right\|_{\mathrm{F}}^2. \tag{160}$$

We first bound the operator-norm term using the representation of $\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ from Lemma 1 as follows.

$$\begin{aligned}
\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 &= \left\|\left(\mathrm{trunc}_r\left\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} + \frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 \\
&= \left\|\left(\mathrm{trunc}_r\left\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} + \frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\} - \mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 \\
&\leq \left\|\mathrm{trunc}_r\left\{\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} + \frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\} - \mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 \\
&\leq \left(8\left\|\frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\|_2^2 + 2\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\right)\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 \\
&= \left(\frac{8}{n}\left\|\frac{1}{\sqrt{n}}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\|_2^2 + 2\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\right)\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2.
\end{aligned} \tag{161}$$

Where in the fourth line we used Lemma 3. Note that

$$\begin{aligned}
\left\|\frac{1}{\sqrt{n}}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\right\|_2^2 &= \lambda_{\max}\left(\frac{1}{n}\boldsymbol{\xi}\mathbf{X}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\mathbf{X}\boldsymbol{\xi}^\top\right) \\
&= \lambda_{\max}\left(\boldsymbol{\xi}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\boldsymbol{\xi}^\top\right) \\
&= \lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\boldsymbol{\xi}^\top\right).
\end{aligned} \tag{162}$$

Here $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}$ is an orthogonal projection matrix onto the row space of $\mathbf{X}$. Moreover,

$$\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 = \lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right) = \lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right), \tag{163}$$

where in the second step, we use the fact that $\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}$ and $\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}$ are similar matrices. Overall,

$$2r\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 \leq \left(16\frac{r}{n}\lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\boldsymbol{\xi}^\top\right) + 4r\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right). \tag{164}$$

Now we proceed with upper bounding the term $\lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\boldsymbol{\xi}^\top\right)$.

$$\lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\boldsymbol{\xi}^\top\right) = \lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\mathbf{\Pi}_{\mathbf{X}}^\top\boldsymbol{\xi}^\top\right) = \left\|\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\right\|_2^2. \tag{165}$$

Now we proceed by taking expectation with respect to the randomness of the training dataset $\mathbf{X}, \boldsymbol{\xi}$:

$$\begin{aligned}
&\mathbb{E}\left[2r\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2\right] \qquad (166) \\
&\leq 16\frac{r}{n}\mathbb{E}\left[\lambda_{\max}\left(\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\boldsymbol{\xi}^\top\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] + 4r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \\
&= 16\frac{r}{n}\mathbb{E}\left[\left\|\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\right\|_2^2\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] + 4r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] \\
&= 16\frac{r}{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\right\|_2^2\middle|\mathbf{X}\right]\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right] + 4r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right].
\end{aligned}$$

Using Lemma 2 we have

$$\mathbb{E}\left[\left\|\boldsymbol{\xi}\mathbf{\Pi}_{\mathbf{X}}\right\|_2^2\middle|\mathbf{X}\right] \leq 5\max\{d_x, d_y\}\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right). \tag{167}$$

Putting back these results into (166) gives us:

$$\begin{aligned}
\mathbb{E}\left[2r\left\|\left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2\right] &\leq 80\frac{r\cdot\max\{d_x, d_y\}}{n}\mathbb{E}\left[\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right]\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\right) \\
&\quad + 4r\mathbb{E}\left[\sigma_{r+1}^2\left(\mathbf{\Delta}^\star\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\mathbf{\Sigma}_{\mathbf{xx}}\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right].
\end{aligned} \tag{168}$$

Next, using the Eckart-Young theorem (Thm. 10) we get

$$\left\|\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}-\mathbf{C}^\star\right\|_{\mathrm{F}}^2=\left\|\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}-\mathrm{trunc}_r\left\{\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\right\|_{\mathrm{F}}^2=\sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right) \tag{169}$$

Putting back all results into (159) gives us:

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{LoRA}})-\mathcal{R}(\mathbf{A}^\star)\right]\\&\leq 80\frac{r\max\{d_x,d_y\}}{n}\mathbb{E}\left[\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right]\lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\right)\\&\quad+4r\mathbb{E}\left[\sigma_{r+1}^2\left(\boldsymbol{\Delta}^\star\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\right)\right]+\sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right).\end{aligned} \tag{170}$$

□

Proposition 4. *Assume $n>d_x$, $\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{xx}})$ i.i.d., and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\succeq\mathbf{0}$. As $n,d_x\to\infty$ with $d_x/n\to c\in[0,1)$, the first term in (13) is asymptotically equal to $r\max\{d_x,d_y\}(\sqrt{n}-\sqrt{d_x})^{-2}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}})$.*

Proof. Write $\mathbf{x}_i=\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\mathbf{z}_i$, where $\mathbf{z}_1,\dots,\mathbf{z}_n$ are independent standard Gaussian vectors, and define

$$\mathbf{Z}=[\mathbf{z}_1,\dots,\mathbf{z}_n]\in\mathbb{R}^{d_x\times n},\qquad\mathbf{W}_n=\frac{1}{n}\mathbf{Z}\mathbf{Z}^\top. \tag{171}$$

The matrix $\mathbf{Z}$ has full row rank almost surely. Moreover,

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}=\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\mathbf{W}_n\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}},\qquad\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}=\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\mathbf{W}_n^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-\frac{1}{2}}. \tag{172}$$

Thus $\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}$ is similar to $\mathbf{W}_n^{-1}$, and hence

$$K_n\coloneqq\lambda_{\max}\big(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\big)=\frac{1}{\lambda_{\min}(\mathbf{W}_n)}=n\left\|\mathbf{Z}^\dagger\right\|_2^2. \tag{173}$$

By Lemma 7, applied with $m=d_x$ and $q=n$,

$$\mathbb{E}\left[K_n\right]\longrightarrow L\coloneqq(1-\sqrt{c})^{-2}. \tag{174}$$

Since $(1-\sqrt{d_x/n})^{-2}\to L>0$, it follows that

$$\mathbb{E}\left[K_n\right]=(1+o(1))\left(1-\sqrt{\frac{d_x}{n}}\right)^{-2}. \tag{175}$$

Substituting into the first term of (13), we obtain

$$\frac{r\max\{d_x,d_y\}}{n}\lambda_{\max}\big(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\big)\mathbb{E}\left[K_n\right]=(1+o(1))\frac{r\max\{d_x,d_y\}}{(\sqrt{n}-\sqrt{d_x})^2}\lambda_{\max}\big(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\big), \tag{176}$$

which proves the claim. □

VI.2 Underdetermined regime

Theorem 7 (LoRA excess risk)**.** *Let $\widehat{\mathbf{A}}_{\mathrm{LoRA}}$ be the LoRA solution (7). If $n<d_x$, then*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}\big(\widehat{\mathbf{A}}_{\mathrm{LoRA}}\big)\right]\lesssim\mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}_{\mathbf{X}}^{\perp}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right]+\frac{r\max\{n,d_y\}}{n}\lambda_{\max}\big(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\big)\mathbb{E}\left[\lambda_{\max}\big(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\big)\right]\\+r\mathbb{E}\left[\sigma_{r+1}^2\left(\boldsymbol{\Delta}^\star\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right]+\sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right).\end{aligned} \tag{17}$$

Proof. Recall from Lemma 6 for any estimator $\mathbf{A}$,

$$\mathcal{R}(\mathbf{A}) - \mathcal{R}(\mathbf{A}^\star) = \left\| (\mathbf{A} - \mathbf{A}^\star) \, \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 . \tag{177}$$

Recall that $\mathbf{A} = \widehat{\mathbf{A}}_{\mathrm{LoRA}} = \mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}$ where $\mathrm{rank}(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}}) \leq r$. Then,

$$\begin{aligned} \mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{LoRA}}) - \mathcal{R}(\mathbf{A}^\star) = \left\| (\mathbf{A} - \mathbf{A}^\star) \, \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 &= \left\| \left(\mathbf{A}_0 + \widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{A}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 \\ &= \left\| \left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 . \end{aligned} \tag{178}$$

By Lemma 10, we get

$$\begin{aligned} \left\| \left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 \leq 2 \left\| \mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}}^{\perp} \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 + 4r \left\| \left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}} \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 \\ + \sum_{i=r+1}^{\min\{d_x, d_y\}} \sigma_i^2 \left(\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right) . \end{aligned} \tag{179}$$

Denote ① as,

$$① = \left\| \left(\widehat{\mathbf{\Delta}}_{\mathrm{LoRA}} - \mathbf{\Delta}^\star \mathbf{P}_{\mathbf{X}} \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 . \tag{180}$$

Note that $\mathbf{P}_{\mathbf{X}} = \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger$, therefore

$$\begin{aligned} ① &= \left\| \left(\mathrm{trunc}_r \left\{ \mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} + \boldsymbol{\xi} \mathbf{X}^\dagger \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right\} - \mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right) \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 \\ &\leq \left(8 \left\| \boldsymbol{\xi} \mathbf{X}^\dagger \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 + 2\sigma_{r+1}^2 \left(\mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right) \right) \left\| \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 \\ &= \left(8 \left\| \boldsymbol{\xi} \mathbf{X}^\dagger \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 + 2\sigma_{r+1}^2 \left(\mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right) \right) \lambda_{\max} \left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^\dagger \right) . \end{aligned} \tag{181}$$

Where in the inequality line we used Lemma 3 and the fact that $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$, and in the last step we used

$$\left\| \left(\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right)^\dagger \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 = \lambda_{\max} \left(\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^\dagger \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right) = \lambda_{\max} \left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^\dagger \right) . \tag{182}$$

Here we used the fact that $\mathbf{C}$ and $\mathbf{DCD}^{-1}$ are similar matrices for $\mathbf{C} = \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^\dagger \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}$ and $\mathbf{D} = \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}$. We proceed with the upper bound of ①,

$$\begin{aligned} \left\| \boldsymbol{\xi} \mathbf{X}^\dagger \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 = \lambda_{\max} \left(\boldsymbol{\xi} \mathbf{X}^\dagger \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \left(\mathbf{X}^\dagger \right)^\top \boldsymbol{\xi}^\top \right) &= \lambda_{\max} \left(\frac{1}{n} \boldsymbol{\xi} \mathbf{X}^\dagger \mathbf{X} \mathbf{X}^\top \left(\mathbf{X}^\dagger \right)^\top \boldsymbol{\xi}^\top \right) \\ &= \frac{1}{n} \lambda_{\max} \left(\boldsymbol{\xi} \mathbf{\Pi}_{\mathbf{X}} \mathbf{\Pi}_{\mathbf{X}}^\top \boldsymbol{\xi}^\top \right) \\ &= \frac{1}{n} \left\| \boldsymbol{\xi} \mathbf{\Pi}_{\mathbf{X}} \right\|_2^2 \\ &\leq \frac{1}{n} \left\| \boldsymbol{\xi} \right\|_2^2 \\ &= \lambda_{\max} \left(\frac{1}{n} \boldsymbol{\xi} \boldsymbol{\xi}^\top \right) = \lambda_{\max} \left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \right) . \end{aligned} \tag{183}$$

Where $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} = \frac{1}{n} \boldsymbol{\xi} \boldsymbol{\xi}^\top$ is the empirical second moment matrix of the noise. Here we used the fact that $\mathbf{\Pi}_{\mathbf{X}} = \mathbf{X}^\dagger \mathbf{X}$ is an orthogonal projection onto the row space of $\mathbf{X}$. Putting back everything together we can bound ① as follows,

$$① \leq \left(8 \lambda_{\max} \left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \right) + 2\sigma_{r+1}^2 \left(\mathbf{\Delta}^\star \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}} \right) \right) \lambda_{\max} \left(\mathbf{\Sigma}_{\mathbf{xx}} \widehat{\mathbf{\Sigma}}_{\mathbf{xx}}^\dagger \right) . \tag{184}$$

Now we take the conditional expectation with respect to the randomness of noise while fixing the data, and get the bound

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{LoRA}})-\mathcal{R}(\mathbf{A}^\star)\middle|\mathbf{X}\right] \le\ & 2\left\|\boldsymbol{\Delta}^\star\mathbf{P}^{\perp}_{\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\ &+ 32r\mathbb{E}\left[\lambda_{\max}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}\right)\middle|\mathbf{X}\right]\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right) \\ &+ 8r\sigma_{r+1}^2\left(\boldsymbol{\Delta}^\star\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right) \\ &+ \sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right). \end{aligned} \tag{185}$$

To bound the term with the empirical second moment of the noise, we first write this Gaussian noise as $\boldsymbol{\xi}=\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}^{\frac{1}{2}}\mathbf{Z}$, where $\mathbf{Z}\in\mathbb{R}^{d_y\times n}$ is a Gaussian matrix with entries distributed $\mathcal{N}(0,1)$ i.i.d. Thus

$$\begin{aligned}\mathbb{E}\left[\lambda_{\max}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}\right)\middle|\mathbf{X}\right] &= \frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\xi}\|_2^2\middle|\mathbf{X}\right] = \frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}^{\frac{1}{2}}\mathbf{Z}\right\|_2^2\middle|\mathbf{X}\right] \\ &\le \lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right)\frac{1}{n}\mathbb{E}\left[\|\mathbf{Z}\|_2^2\middle|\mathbf{X}\right] \\ &\le \lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right)\frac{5\max\{d_y,n\}}{n}, \end{aligned} \tag{186}$$

where in the last line we used Lemma 5. Finally, take full expectation over the data and the noise,

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{\mathbf{A}}_{\mathrm{LoRA}})-\mathcal{R}(\mathbf{A}^\star)\right] \le\ & 2\mathbb{E}\left[\left\|\boldsymbol{\Delta}^\star\mathbf{P}^{\perp}_{\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] \\ &+ 160\frac{r\max\{d_y,n\}}{n}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})\mathbb{E}\left[\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right] \\ &+ 8r\mathbb{E}\left[\sigma_{r+1}^2\left(\boldsymbol{\Delta}^\star\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\frac{1}{2}}\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)\right] \\ &+ \sum_{i=r+1}^{\min\{d_x,d_y\}}\sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right). \end{aligned} \tag{187}$$

□

Proposition 6. *Assume $d_x > n$, $\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\sigma_{\mathbf{x}}^2\mathbf{I})$ i.i.d. and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\succeq\mathbf{0}$. As $n, d_x\to\infty$ with $n/d_x\to c\in[0,1)$, the second term in (17) is asymptotically equal to $r\max\{n,d_y\}(\sqrt{d_x}-\sqrt{n})^{-2}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})$.*

Proof. The argument follows the proof of Prop. 4, with the roles of the two Gaussian matrix dimensions interchanged.

Write $\mathbf{x}_i=\sigma_{\mathbf{x}}\mathbf{z}_i$, where $\mathbf{z}_1,\dots,\mathbf{z}_n$ are independent standard Gaussian vectors, and define

$$\mathbf{Z}=[\mathbf{z}_1,\dots,\mathbf{z}_n]\in\mathbb{R}^{d_x\times n},\qquad \mathbf{W}_n=\frac{1}{d_x}\mathbf{Z}^\top\mathbf{Z}. \tag{188}$$

The matrix $\mathbf{Z}$ has full column rank almost surely. Moreover,

$$\boldsymbol{\Sigma}_{\mathbf{xx}}=\sigma_{\mathbf{x}}^2\mathbf{I},\qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}=\frac{\sigma_{\mathbf{x}}^2}{n}\mathbf{Z}\mathbf{Z}^\top, \tag{189}$$

and therefore

$$\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}=n(\mathbf{Z}\mathbf{Z}^\top)^{\dagger}. \tag{190}$$

Since the nonzero eigenvalues of $\mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{Z}^\top\mathbf{Z}$ coincide, it follows that

$$\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{\dagger}\right)=\frac{n}{\lambda_{\min}(\mathbf{Z}^\top\mathbf{Z})}=\frac{n}{d_x}K_n, \tag{191}$$

where

$$K_n\coloneqq\frac{1}{\lambda_{\min}(\mathbf{W}_n)}=d_x\left\|\mathbf{Z}^{\dagger}\right\|_2^2. \tag{192}$$

By Lemma 7, applied to $\mathbf{Z}^\top \in \mathbb{R}^{n\times d_x}$ with $m = n$ and $q = d_x$, and using $\left\|(\mathbf{Z}^\top)^\dagger\right\|_2 = \left\|\mathbf{Z}^\dagger\right\|_2$,

$$\mathbb{E}\left[K_n\right] \longrightarrow L \coloneqq (1-\sqrt{c})^{-2}. \tag{193}$$

Since $(1-\sqrt{n/d_x})^{-2} \to L > 0$, it follows that

$$\mathbb{E}\left[K_n\right] = (1+o(1))\left(1-\sqrt{\frac{n}{d_x}}\right)^{-2}. \tag{194}$$

Consequently,

$$\mathbb{E}\left[\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}^\dagger_{\mathbf{xx}}\right)\right] = \frac{n}{d_x}\mathbb{E}\left[K_n\right] = (1+o(1))\frac{n}{(\sqrt{d_x}-\sqrt{n})^2}. \tag{195}$$

Substituting into the second term of (17), we obtain

$$\frac{r\max\{n,d_y\}}{n}\lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\right)\mathbb{E}\left[\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}\widehat{\boldsymbol{\Sigma}}^\dagger_{\mathbf{xx}}\right)\right] = (1+o(1))\frac{r\max\{n,d_y\}}{(\sqrt{d_x}-\sqrt{n})^2}\lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}\right), \tag{196}$$

which proves the claim. □

VI.3 Inevitable error of low-rank adaptations

Proposition 8. *Let $\widehat{\mathbf{A}} \in \mathbb{R}^{d_y\times d_x}$ be a low rank adaptation of $\mathbf{A}_0$, i.e., taking the form $\widehat{\mathbf{A}} = \mathbf{A}_0 + \widehat{\boldsymbol{\Delta}}$ with $\operatorname{rank}(\widehat{\boldsymbol{\Delta}}) \le r$. Then*

$$\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}\right)\right] \ge \sum_{i=r+1}^{\min\{d_x,d_y\}} \sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right). \tag{197}$$

Proof. From Lemma 6 we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}\left(\widehat{\mathbf{A}}\right)\right] = \mathbb{E}\left[\mathcal{R}\left(\widehat{\mathbf{A}}\right)\right] - \mathcal{R}\left(\mathbf{A}^\star\right) &= \mathbb{E}\left[\left\|\left(\widehat{\mathbf{A}}-\mathbf{A}^\star\right)\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\left\|\left(\widehat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}^\star\right)\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] \\
&\ge \min_{\substack{\boldsymbol{\Delta}\in\mathbb{R}^{d_y\times d_x}\\ \operatorname{rank}(\boldsymbol{\Delta})\le r}} \mathbb{E}\left[\left\|\left(\boldsymbol{\Delta}-\boldsymbol{\Delta}^\star\right)\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\right] \\
&= \sum_{i=r+1}^{\min\{d_x,d_y\}} \sigma_i^2\left(\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right),
\end{aligned} \tag{198}$$

where an optimal adaptation for the above minimization problem is attained, *e.g.*, for $\boldsymbol{\Delta} = \operatorname{trunc}_r\left\{\boldsymbol{\Delta}^\star\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-\frac{1}{2}}$, and the value is given by Eckart-Young theorem (Thm. 10). □

VII Auxiliary results

Theorem 10 (Eckart-Young theorem, Eckart & Young (1936))**.** *Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, any solution of*

$$\min_{\operatorname{rank}(\mathbf{B})\le r} \left\|\mathbf{B}-\mathbf{M}\right\|_{\mathrm{F}}, \tag{199}$$

is given by $\operatorname{trunc}_r\{\mathbf{M}\}$, where $\operatorname{trunc}_r\{\cdot\}$ denotes the top-r singular-value truncation. If the truncation is non-unique, then any truncation $\operatorname{trunc}_r\{\mathbf{M}\}$ is a solution. Furthermore, any truncation $\operatorname{trunc}_r\{\mathbf{M}\}$ is a solution of

$$\min_{\operatorname{rank}(\mathbf{B})\le r} \left\|\mathbf{B}-\mathbf{M}\right\|_2. \tag{200}$$

Finally,

$$\left\|\operatorname{trunc}_r\{\mathbf{M}\}-\mathbf{M}\right\|_2 = \sigma_{r+1}(\mathbf{M}), \qquad \left\|\operatorname{trunc}_r\{\mathbf{M}\}-\mathbf{M}\right\|_{\mathrm{F}}^2 = \sum_{i=r+1}^{\min\{m,n\}}\sigma_i^2(\mathbf{M}). \tag{201}$$

Lemma 3. *For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a perturbation matrix $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$ we have*

$$\|\mathrm{trunc}_r\{\mathbf{A}+\mathbf{\Delta}\} - \mathbf{A}\|_2 \le \sigma_{r+1}(\mathbf{A}) + 2\,\|\mathbf{\Delta}\|_2\,. \tag{202}$$

Proof. By adding and subtracting $\mathbf{\Delta}$ we have

$$\|\mathrm{trunc}_r\{\mathbf{A}+\mathbf{\Delta}\} - (\mathbf{A}+\mathbf{\Delta}) + \mathbf{\Delta}\|_2 \le \|\mathrm{trunc}_r\{\mathbf{A}+\mathbf{\Delta}\} - (\mathbf{A}+\mathbf{\Delta})\|_2 + \|\mathbf{\Delta}\|_2\,. \tag{203}$$

Now we bound the first term above as

$$\begin{aligned}\|\mathrm{trunc}_r\{\mathbf{A}+\mathbf{\Delta}\} - (\mathbf{A}+\mathbf{\Delta})\|_2 &= \min_{\mathrm{rank}(\mathbf{B})\le r} \|\mathbf{B} - (\mathbf{A}+\mathbf{\Delta})\|_2 \\ &\le \|\mathrm{trunc}_r\{\mathbf{A}\} - (\mathbf{A}+\mathbf{\Delta})\|_2 \\ &\le \|\mathrm{trunc}_r\{\mathbf{A}\} - \mathbf{A}\|_2 + \|\mathbf{\Delta}\|_2 \\ &\le \sigma_{r+1}(\mathbf{A}) + \|\mathbf{\Delta}\|_2\,,\end{aligned} \tag{204}$$

where in the first step we used Eckart-Young theorem (Thm. 10). Putting it together gives the result. □

Lemma 4 (Gaussian Poincaré inequality, Boucheron et al. (2013, Thm. 3.20))**.** *Let $X = (X_1, ..., X_n)$ be a vector of i.i.d. standard Gaussian random variables (i.e., X is a Gaussian vector with zero mean vector and identity covariance matrix), and $f : \mathbb{R}^d \to \mathbb{R}$ be 1-Lipschitz. Then*

$$\mathbb{V}\mathrm{ar}(f(X)) \le 1. \tag{205}$$

Lemma 5. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ with i.i.d. entries $\mathcal{N}(0,1)$. Then*

$$\mathbb{E}\left[\|\mathbf{A}\|_2^2\right] \le 5\max\{m,n\}. \tag{206}$$

Proof. We use two standard facts.

(i) Expectation bound. A classical bound for Gaussian matrices states that

$$\mathbb{E}\left[\|\mathbf{A}\|_2\right] \le \sqrt{m} + \sqrt{n}. \tag{207}$$

This can easily be obtained in various ways, *e.g.,* using Gordon's Theorem (Gordon, 1985, Thm. 2.5).

(ii) Variance bound via Gaussian Poincaré. The map $\mathbf{A} \mapsto \|\mathbf{A}\|$ is 1-Lipschitz with respect to the Frobenius (Euclidean) norm, since for all matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ we have

$$\left|\|\mathbf{A}\|_2 - \|\mathbf{B}\|_2\right| \le \|\mathbf{A}-\mathbf{B}\|_2 \le \|\mathbf{A}-\mathbf{B}\|_{\mathrm{F}}\,. \tag{208}$$

Hence, by the Gaussian Poincaré inequality (Lemma 4),

$$\mathbb{V}\mathrm{ar}\left(\|\mathbf{A}\|\right) \le 1. \tag{209}$$

Combining these two facts and using $\mathbb{E}[\|\mathbf{A}\|_2^2] = (\mathbb{E}\|\mathbf{A}\|)^2 + \mathbb{V}\mathrm{ar}(\|\mathbf{A}\|)$, we obtain

$$\mathbb{E}\left[\|\mathbf{A}\|_2^2\right] \le \left(\sqrt{m}+\sqrt{n}\right)^2 + 1. \tag{210}$$

Finally, using the elementary inequality $(\sqrt{a}+\sqrt{b})^2 \le 4\max\{a,b\}$ for $a, b \ge 0$, we have

$$\mathbb{E}\left[\|\mathbf{A}\|_2^2\right] \le 4\max\{m,n\} + 1 \le 5\max\{m,n\}, \tag{211}$$

which proves the lemma. □

Lemma 6. *Consider the regression task defined in Sec. 2, and let $\widehat{\mathbf{A}}$ be some estimator. Then*

$$\mathcal{R}(\widehat{\mathbf{A}}) - \mathcal{R}(\mathbf{A}^\star) = \left\|\left(\widehat{\mathbf{A}} - \mathbf{A}^\star\right)\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2. \tag{212}$$

Proof. The excess risk for a given estimator $\widehat{\mathbf{A}}$ is defined as

$$\mathcal{R}(\widehat{\mathbf{A}}) - \mathcal{R}(\mathbf{A}^\star) = \mathbb{E}\left[\|\mathbf{y} - \widehat{\mathbf{A}}\mathbf{x}\|^2 - \|\mathbf{y} - \mathbf{A}^\star\mathbf{x}\|^2 \;\middle|\; \widehat{\mathbf{A}}\right]. \tag{213}$$

Using $\mathbf{y} = \mathbf{A}^\star\mathbf{x} + \boldsymbol{\varepsilon}$, we compute for $\widehat{\mathbf{A}}$

$$\|\mathbf{y} - \widehat{\mathbf{A}}\mathbf{x}\|^2 = \|\boldsymbol{\varepsilon} - (\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x}\|^2 = \|\boldsymbol{\varepsilon}\|^2 - 2\boldsymbol{\varepsilon}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} + \mathbf{x}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x}, \tag{214}$$

while for $\mathbf{A}^\star$

$$\|\mathbf{y} - \mathbf{A}^\star\mathbf{x}\|^2 = \|\boldsymbol{\varepsilon}\|^2. \tag{215}$$

Thus,

$$\mathcal{R}(\widehat{\mathbf{A}}) - \mathcal{R}(\mathbf{A}^\star) = \mathbb{E}\left[-2\boldsymbol{\varepsilon}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} + \mathbf{x}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} \;\middle|\; \widehat{\mathbf{A}}\right]. \tag{216}$$

Since $\mathbf{x}$, $\boldsymbol{\varepsilon}$ and $\widehat{\mathbf{A}}$ are independent, and $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, the cross term vanishes, *i.e.*,

$$\mathbb{E}\left[\boldsymbol{\varepsilon}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} \;\middle|\; \widehat{\mathbf{A}}\right] = \mathbb{E}\left[\boldsymbol{\varepsilon}^\top \;\middle|\; \widehat{\mathbf{A}}\right]\mathbb{E}\left[(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} \;\middle|\; \widehat{\mathbf{A}}\right] = 0. \tag{217}$$

Hence, the excess risk is,

$$\begin{aligned}
\mathcal{R}(\widehat{\mathbf{A}}) - \mathcal{R}(\mathbf{A}^\star) &= \mathbb{E}\left[\mathbf{x}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x} \;\middle|\; \widehat{\mathbf{A}}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{x}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x}\right) \;\middle|\; \widehat{\mathbf{A}}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left((\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbf{x}\mathbf{x}^\top(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top\right) \;\middle|\; \widehat{\mathbf{A}}\right] \\
&= \mathrm{Tr}\left((\widehat{\mathbf{A}} - \mathbf{A}^\star)\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \;\middle|\; \widehat{\mathbf{A}}\right](\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top\right) \\
&= \mathrm{Tr}\left((\widehat{\mathbf{A}} - \mathbf{A}^\star)\boldsymbol{\Sigma}_{\mathbf{xx}}(\widehat{\mathbf{A}} - \mathbf{A}^\star)^\top\right) \\
&= \left\|(\widehat{\mathbf{A}} - \mathbf{A}^\star)\boldsymbol{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{218}$$

□

Theorem 11 (Vanishing quadratic regularization)**.** *Let $\mathcal{L}\colon \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lower semicontinuous, bounded below, and assume that $\mathcal{S} \coloneqq \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \neq \varnothing$. For $\lambda > 0$, let*

$$\boldsymbol{\theta}_\lambda \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{\mathcal{L}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|^2\right\}, \tag{219}$$

and define

$$\mathcal{S}^\star \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathcal{S}} \|\boldsymbol{\theta}\|^2. \tag{220}$$

Then $\boldsymbol{\theta}_\lambda$ is bounded, and every accumulation point of $\boldsymbol{\theta}_\lambda$ as $\lambda \to 0^+$ belongs to $\mathcal{S}^\star$. In particular, if $\mathcal{S}^\star = \{\boldsymbol{\theta}^\star\}$ is a singleton, then $\boldsymbol{\theta}_\lambda \underset{\lambda\to 0^+}{\longrightarrow} \boldsymbol{\theta}^\star$.

Proof. Assume that $\mathcal{L}$ is proper, lower semicontinuous, bounded below, and attains its minimum. Define

$$\mathcal{L}^\star \coloneqq \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \in \mathbb{R} \quad \text{and} \quad \mathcal{S} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}. \tag{221}$$

Since $\mathcal{L}$ is lower semicontinuous, the set $\mathcal{S} = \{\boldsymbol{\theta} : \mathcal{L}(\boldsymbol{\theta}) \leq \mathcal{L}^\star\}$ is closed. By the assumptions, it is also nonempty, so the continuous, coercive function $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|^2$ attains its minimum on $\mathcal{S}$. Thus $\mathcal{S}^\star \neq \varnothing$. Choose $\boldsymbol{\theta}^\star \in \mathcal{S}^\star$ and set $\alpha \coloneqq \|\boldsymbol{\theta}^\star\|$.

For every $\lambda > 0$, define the function

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) \coloneqq \mathcal{L}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|^2. \tag{222}$$

The function $\mathcal{L}_\lambda$ is proper and lower semicontinuous. Since $\mathcal{L}$ is bounded below, $\mathcal{L}_\lambda$ is also coercive. Consequently, it attains its minimum. Let $\boldsymbol{\theta}_\lambda$ be a global minimizer of $\mathcal{L}_\lambda$. Because $\boldsymbol{\theta}_\lambda$ minimizes $\mathcal{L}_\lambda$, comparing it with $\boldsymbol{\theta}^\star$ gives $\mathcal{L}_\lambda(\boldsymbol{\theta}_\lambda) \leq \mathcal{L}_\lambda(\boldsymbol{\theta}^\star)$, *i.e.*,

$$\mathcal{L}(\boldsymbol{\theta}_\lambda) + \lambda\|\boldsymbol{\theta}_\lambda\|^2 \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \lambda\|\boldsymbol{\theta}^\star\|^2 = \mathcal{L}^\star + \lambda\|\boldsymbol{\theta}^\star\|^2. \tag{223}$$

Because $\mathcal{L}^\star \le \mathcal{L}(\boldsymbol{\theta}_\lambda)$, (223) yields two bounds. First,

$$0 \le \mathcal{L}(\boldsymbol{\theta}_\lambda) - \mathcal{L}^\star \le \lambda \|\boldsymbol{\theta}^\star\|^2. \tag{224}$$

Therefore, by squeezing,

$$\lim_{\lambda \to 0^+} \mathcal{L}(\boldsymbol{\theta}_\lambda) = \mathcal{L}^\star. \tag{225}$$

Second, again from (223), $\lambda \|\boldsymbol{\theta}_\lambda\|^2 \le \lambda \|\boldsymbol{\theta}^\star\|^2$, so

$$\|\boldsymbol{\theta}_\lambda\| \le \|\boldsymbol{\theta}^\star\| = \alpha. \tag{226}$$

Thus $\boldsymbol{\theta}_\lambda$ is bounded as a function of λ. Let $\bar{\boldsymbol{\theta}}$ be any accumulation point as $\lambda \to 0^+$, and choose a sequence $\lambda_k \to 0^+$ such that $\boldsymbol{\theta}_{\lambda_k} \to \bar{\boldsymbol{\theta}}$. Lower semicontinuity yields

$$\mathcal{L}^\star \le \mathcal{L}(\bar{\boldsymbol{\theta}}) \le \liminf_{k \to \infty} \mathcal{L}(\boldsymbol{\theta}_{\lambda_k}) = \mathcal{L}^\star, \tag{227}$$

where at the last step we used (225). Therefore $\bar{\boldsymbol{\theta}} \in \mathcal{S}$. By continuity of the norm,

$$\|\bar{\boldsymbol{\theta}}\| = \lim_{k \to \infty} \|\boldsymbol{\theta}_{\lambda_k}\| \le \alpha, \tag{228}$$

where at the last step we used (226). On the other hand, because $\bar{\boldsymbol{\theta}} \in \mathcal{S}$ and α is the minimum norm on $\mathcal{S}$, we have $\alpha \le \|\bar{\boldsymbol{\theta}}\|$. Thus $\|\bar{\boldsymbol{\theta}}\| = \alpha = \|\boldsymbol{\theta}^\star\|$, which proves that $\bar{\boldsymbol{\theta}} \in \mathcal{S}^\star$. Therefore, every cluster point of $\boldsymbol{\theta}_\lambda$ as $\lambda \to 0^+$ belongs to $\mathcal{S}^\star$.

Finally, suppose that $\mathcal{S}^\star$ is a singleton, *i.e.*, $\mathcal{S}^\star = \{\boldsymbol{\theta}^\star\}$, and assume by contradiction that $\boldsymbol{\theta}_\lambda$ does not converge to $\boldsymbol{\theta}^\star$ as $\lambda \to 0^+$. Then there would exist $\varepsilon > 0$ and a sequence $\lambda_k \to 0^+$ such that

$$\|\boldsymbol{\theta}_{\lambda_k} - \boldsymbol{\theta}^\star\| \ge \varepsilon \qquad \text{for every } k. \tag{229}$$

Boundedness of $\boldsymbol{\theta}_{\lambda_k}$ gives a convergent subsequence of this sequence. By the preceding argument, its limit belongs to $\mathcal{S}^\star$. Since $\mathcal{S}^\star$ is a singleton, the limit must equal $\boldsymbol{\theta}^\star$, contradicting $\|\boldsymbol{\theta}_{\lambda_k} - \boldsymbol{\theta}^\star\| \ge \varepsilon$ along the convergent subsequence. Therefore, in this case

$$\lim_{\lambda \to 0^+} \boldsymbol{\theta}_\lambda = \boldsymbol{\theta}^\star. \tag{230}$$

□

Lemma 7 (Asymptotic Gaussian pseudoinverse moment)**.** *Let $m < q$ be positive integers and $\mathbf{Z} \in \mathbb{R}^{m \times q}$ have i.i.d. $\mathcal{N}(0,1)$ entries. Assume that $m, q \to \infty$ with $m/q \to c \in [0,1)$, and define $K_q \coloneqq q \left\|\mathbf{Z}^\dagger\right\|_2^2$. Then*

$$K_q \xrightarrow{\text{a.s.}} (1 - \sqrt{c})^{-2}, \qquad \textit{and} \qquad \mathbb{E}\left[K_q\right] \longrightarrow (1 - \sqrt{c})^{-2}. \tag{231}$$

Proof. Define

$$\mathbf{W}_q = \frac{1}{q} \mathbf{Z}\mathbf{Z}^\top. \tag{232}$$

Since $\mathbf{Z}$ has full row rank almost surely,

$$K_q = q \left\|\mathbf{Z}^\dagger\right\|_2^2 = \frac{q}{\sigma_m^2(\mathbf{Z})} = \frac{1}{\lambda_{\min}(\mathbf{W}_q)}. \tag{233}$$

For $c \in (0,1)$, the smallest-eigenvalue limit of Silverstein (1985) gives

$$\lambda_{\min}(\mathbf{W}_q) \xrightarrow{\text{a.s.}} (1 - \sqrt{c})^2. \tag{234}$$

The same conclusion holds for $c = 0$. Indeed, applying the Gaussian singular-value deviation bound (Vershynin, 2012, Cor. 5.35) to $\mathbf{Z}^\top$ with deviation parameter $q^{1/4}$, we obtain, with probability at least $1 - 2\exp(-\sqrt{q}/2)$,

$$\frac{\sigma_1(\mathbf{Z})}{\sqrt{q}} \le 1 + \sqrt{\frac{m}{q}} + q^{-1/4}, \qquad \text{and} \qquad \frac{\sigma_m(\mathbf{Z})}{\sqrt{q}} \ge 1 - \sqrt{\frac{m}{q}} - q^{-1/4}. \tag{235}$$

The failure probabilities are summable in q. Thus, by the Borel–Cantelli lemma and $m/q \to 0$, $\sigma_m(\mathbf{Z})/\sqrt{q} \to 1$ almost surely, which implies $\lambda_{\min}(\mathbf{W}_q) \to 1$ almost surely. Consequently, for every $c \in [0, 1)$,

$$K_q \xrightarrow{\text{a.s.}} L := (1 - \sqrt{c})^{-2} < \infty. \tag{236}$$

We next control the upper tail of K_q to justify passing to expectations. Put $p_q = q - m$. Since $p_q/q \to 1 - c > 0$, we may therefore restrict attention to sufficiently large indices for which $p_q \geq 4$, without affecting the asymptotic claim. The Gaussian pseudoinverse bound (Halko et al., 2011, Prop. 10.4, Eq. (10.6)) states that

$$\mathbb{P}\left\{\left\|\mathbf{Z}^\dagger\right\|_2 \geq \frac{e\sqrt{q}}{p_q + 1} t\right\} \leq t^{-(p_q+1)}, \qquad t \geq 1. \tag{237}$$

Define

$$T_q = \left\|\mathbf{Z}^\dagger\right\|_2, \qquad a_q = \frac{e\sqrt{q}}{p_q + 1}. \tag{238}$$

Equivalently, $\mathbb{P}\{T_q > u\} \leq (a_q/u)^{p_q+1}$ for $u \geq a_q$. Since

$$T_q^4 = \int_0^\infty 4u^3 \mathbf{1}_{\{T_q > u\}}\, du, \tag{239}$$

Tonelli's theorem, applied to this nonnegative integrand, gives

$$\mathbb{E}\left[T_q^4\right] = 4\int_0^\infty u^3 \mathbb{P}\{T_q > u\}\, du \tag{240}$$

$$\leq 4\int_0^{a_q} u^3\, du + 4a_q^{p_q+1} \int_{a_q}^\infty u^{2-p_q}\, du \tag{241}$$

$$= a_q^4 + \frac{4a_q^4}{p_q - 3} = a_q^4 \frac{p_q + 1}{p_q - 3}. \tag{242}$$

Here the last integral is finite because $p_q > 3$. It follows that

$$\mathbb{E}\left[K_q^2\right] = q^2 \mathbb{E}\left[T_q^4\right] \leq \frac{e^4 q^4}{(p_q + 1)^4} \frac{p_q + 1}{p_q - 3}. \tag{243}$$

Since $p_q/q \to 1 - c > 0$, there exist q_0 and a finite constant C, independent of q, such that

$$\sup_{q \geq q_0} \mathbb{E}\left[K_q^2\right] \leq C. \tag{244}$$

The uniform second-moment bound implies that $\{K_q : q \geq q_0\}$ is uniformly integrable. Together with $K_q \xrightarrow{\text{a.s.}} L$, this yields

$$\mathbb{E}[K_q] \longrightarrow L = (1 - \sqrt{c})^{-2}. \tag{245}$$

□

Theorem 12 (Representation cost of deep matrix factorization, adapted from Dai et al. (2021))**.** *Let $m \geq 1$ and $d_0, \ldots, d_m$ be positive integers, with $d_0 = d_x$ and $d_m = d_y$. Set*

$$r := \min_{0 \leq i \leq m} d_i, \qquad \textit{and} \qquad p := \frac{2}{m}.$$

For every $\mathbf{\Delta} \in \mathbb{R}^{d_y \times d_x}$ satisfying $\operatorname{rank}(\mathbf{\Delta}) \leq r$, we have

$$\min_{\substack{\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}},\ i \in [m] \\ \mathbf{W}_m \cdots \mathbf{W}_1 = \mathbf{\Delta}}} \frac{1}{m} \sum_{i=1}^m \|\mathbf{W}_i\|_{\mathrm{F}}^2 = \|\mathbf{\Delta}\|_{S_p}^p, \tag{246}$$

where $\|\cdot\|_{S_p}$ is the Schatten-p (quasi) norm.

Theorem 1 of Dai et al. (2021) is stated under the stronger assumption that every hidden width is at least $\min\{d_x, d_y\}$. Its proof in Appendix A.1 applies under the rank condition above: the lower-bound argument is independent of this width assumption, and the construction attaining the bound uses only $\operatorname{rank}(\mathbf{\Delta})$ neurons in each hidden layer.

Lemma 8. *Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be a random Gaussian matrix with entries $\mathcal{N}(0, \sigma_{\mathbf{x}}^2)$ i.i.d. such that $1 \le n \le d$ and $\sigma_{xx}^2 > 0$. Denote by $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ the projection matrix onto the column space of $\mathbf{X}$. Then $\mathbb{E}[\mathbf{P}_{\mathbf{X}}] = \frac{n}{d}\mathbf{I}$.*

Proof. Since $\mathbf{X}$ has full column rank almost surely, $\mathbf{P}_{\mathbf{X}}$ is a rank-n orthogonal projector almost surely. In particular, $\mathbf{0} \preceq \mathbf{P}_{\mathbf{X}} \preceq \mathbf{I}_d$, so it is bounded and its expectation exists. Let $\mathbf{R} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix. Note that $\mathbf{RX}$ is distributed identically to $\mathbf{X}$. Therefore,

$$\begin{aligned}\mathbb{E}[\mathbf{P}_{\mathbf{X}}] = \mathbb{E}[\mathbf{P}_{\mathbf{RX}}] &= \mathbb{E}\left[\mathbf{RX}\big((\mathbf{RX})^\top(\mathbf{RX})\big)^{-1}(\mathbf{RX})^\top\right] \\ &= \mathbf{R}\mathbb{E}\left[\mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\right]\mathbf{R}^\top = \mathbf{R}\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{R}^\top. \end{aligned} \tag{247}$$

Therefore, $\mathbb{E}[\mathbf{P}_{\mathbf{X}}]$ must satisfy $\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{R} = \mathbf{R}\mathbb{E}[\mathbf{P}_{\mathbf{X}}]$ for all orthogonal matrices. Namely, it commutes with any orthogonal matrix. Specifically for $\mathbf{R} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^\top$, with $\|\mathbf{v}\| = 1$. It is easy to validate that this is indeed an orthogonal matrix. Therefore,

$$\begin{aligned} & (\mathbf{I} - 2\mathbf{v}\mathbf{v}^\top)\mathbb{E}[\mathbf{P}_{\mathbf{X}}] = \mathbb{E}[\mathbf{P}_{\mathbf{X}}](\mathbf{I} - 2\mathbf{v}\mathbf{v}^\top) \qquad \forall \|\mathbf{v}\| = 1 \\ \Leftrightarrow \quad & \mathbf{v}(\mathbf{v}^\top\mathbb{E}[\mathbf{P}_{\mathbf{X}}]) = (\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{v})\mathbf{v}^\top \qquad \forall \|\mathbf{v}\| = 1 \\ \Rightarrow \quad & \mathbf{v}(\mathbf{v}^\top\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{v}) = \mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{v} \qquad \forall \|\mathbf{v}\| = 1. \end{aligned} \tag{248}$$

Where in the last step we multiplied by $\mathbf{v}$ from the right both sides of the equation. Therefore, we have that $\mathbb{E}[\mathbf{P}_{\mathbf{X}}]$ satisfies $\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{v} = \lambda_{\mathbf{v}}\mathbf{v}$, with $\lambda_{\mathbf{v}} = \mathbf{v}^\top\mathbb{E}[\mathbf{P}_{\mathbf{X}}]\mathbf{v}$, for all $\|\mathbf{v}\| = 1$. This means that any vector is an eigenvector for $\mathbb{E}[\mathbf{P}_{\mathbf{X}}]$. This holds only for matrices proportional to the identity matrix. Thus $\mathbb{E}[\mathbf{P}_{\mathbf{X}}] = \lambda\mathbf{I}$ for some positive λ. Using the fact that $\mathrm{Tr}(\mathbf{P}_{\mathbf{X}}) = n$ almost surely, we get

$$\lambda d = \mathrm{Tr}(\lambda\mathbf{I}) = \mathrm{Tr}(\mathbb{E}[\mathbf{P}_{\mathbf{X}}]) = \mathbb{E}[\mathrm{Tr}(\mathbf{P}_{\mathbf{X}})] = n. \tag{249}$$

Hence $\lambda = n/d$. □

Lemma 9. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_y\times d_x}$ and let $\mathbf{\Sigma}_{\mathbf{xx}} \in \mathbb{R}^{d_x\times d_x}$ be positive semidefinite. Assume that $\mathrm{rank}(\mathbf{A}) \le r$, and denote $\mathbf{C}^\star = \mathrm{trunc}_r\left\{\mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\}$. Then*

$$\left\|(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \le 2r\left\|(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 + \left\|\mathbf{C}^\star - \mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2. \tag{250}$$

Proof. Let $\mathbf{Q} \in \mathbb{R}^{d_y\times d_y}$ be the orthogonal projection onto the sum of the column spaces of $\mathbf{A}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}$ and $\mathbf{C}^\star$. Since both matrices have rank at most r, we have

$$\mathrm{rank}(\mathbf{Q}) \le \mathrm{rank}\left(\mathbf{A}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right) + \mathrm{rank}\left(\mathbf{C}^\star\right) \le 2r. \tag{251}$$

Moreover, $\mathbf{Q}\mathbf{A}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}$ and $\mathbf{Q}\mathbf{C}^\star = \mathbf{C}^\star$, so

$$(\mathbf{I} - \mathbf{Q})(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} = -(\mathbf{I} - \mathbf{Q})\left(\mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{C}^\star\right). \tag{252}$$

Using the orthogonality of $\mathbf{Q}$ and $\mathbf{I} - \mathbf{Q}$, we obtain

$$\begin{aligned}\left\|(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 &= \left\|\mathbf{Q}(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \left\|(\mathbf{I} - \mathbf{Q})(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\ &= \left\|\mathbf{Q}(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 + \left\|(\mathbf{I} - \mathbf{Q})\left(\mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{C}^\star\right)\right\|_{\mathrm{F}}^2 \\ &\le 2r\left\|\mathbf{Q}(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 + \left\|(\mathbf{I} - \mathbf{Q})\left(\mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{C}^\star\right)\right\|_{\mathrm{F}}^2 \\ &\le 2r\left\|(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right\|_2^2 + \left\|\mathbf{B}\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{C}^\star\right\|_{\mathrm{F}}^2 \end{aligned} \tag{253}$$

In the first inequality, we used $\mathrm{rank}\left(\mathbf{Q}(\mathbf{A} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}\right) \le 2r$ and the relation $\|\mathbf{M}\|_{\mathrm{F}}^2 \le \mathrm{rank}(\mathbf{M})\|\mathbf{M}\|_2^2$. The last inequality follows because orthogonal projections contract both the operator and Frobenius norms. □

Lemma 10 (Weighted low-rank decomposition)**.** *Let* $\widehat{\mathbf{\Delta}}, \mathbf{\Delta}^\star \in \mathbb{R}^{d_y \times d_x}$ *and let* $\mathbf{\Sigma_{xx}} \in \mathbb{R}^{d_x \times d_x}$ *be positive semidefinite. Assume that* $\operatorname{rank}(\widehat{\mathbf{\Delta}}) \le r$*, where* $0 \le r \le \min\{d_x, d_y\}$ *is an integer. For* $\mathbf{X} \in \mathbb{R}^{d_x \times n}$*, let* $\mathbf{P_X} = \mathbf{X}\mathbf{X}^\dagger$ *be the orthogonal projection onto its column space, and define* $\mathbf{P}_\mathbf{X}^\perp = \mathbf{I}_{d_x} - \mathbf{P_X}$*. Then*

$$\left\| \left(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 \le 2 \left\| \mathbf{\Delta}^\star \mathbf{P}_\mathbf{X}^\perp \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 + 4r \left\| \left(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}^\star \mathbf{P_X} \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_2^2 + \sum_{i=r+1}^{\min\{d_x, d_y\}} \sigma_i^2 \left(\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right). \tag{254}$$

Proof. Define

$$\mathbf{M} \coloneqq \mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}, \qquad \mathbf{M}_r \coloneqq \operatorname{trunc}_r\{\mathbf{M}\}, \tag{255}$$

and let $\mathbf{Q} \in \mathbb{R}^{d_y \times d_y}$ be the orthogonal projection onto the sum of the column spaces of $\widehat{\mathbf{\Delta}}$ and $\mathbf{M}_r$. Then

$$\operatorname{rank}(\mathbf{Q}) \le \operatorname{rank}(\widehat{\mathbf{\Delta}}) + \operatorname{rank}(\mathbf{M}_r) \le 2r, \qquad \mathbf{Q}\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Delta}}, \qquad \mathbf{Q}\mathbf{M}_r = \mathbf{M}_r. \tag{256}$$

We also define

$$\mathbf{D} \coloneqq \left(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}^\star \mathbf{P_X} \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}, \qquad \mathbf{B} \coloneqq \mathbf{\Delta}^\star \mathbf{P}_\mathbf{X}^\perp \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}}. \tag{257}$$

Thus,

$$\left(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{D} - \mathbf{B}. \tag{258}$$

Since $(\mathbf{I} - \mathbf{Q})\widehat{\mathbf{\Delta}} = \mathbf{0}$ and $(\mathbf{I} - \mathbf{Q})\mathbf{M}_r = \mathbf{0}$, we have

$$(\mathbf{I} - \mathbf{Q})(\mathbf{D} - \mathbf{B}) = -(\mathbf{I} - \mathbf{Q})\mathbf{M} = -(\mathbf{I} - \mathbf{Q})(\mathbf{M} - \mathbf{M}_r). \tag{259}$$

Using the orthogonality of $\mathbf{Q}$ and $\mathbf{I} - \mathbf{Q}$, we obtain

$$\begin{aligned} \left\| \left(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}^\star \right) \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 &= \|\mathbf{Q}(\mathbf{D} - \mathbf{B})\|_{\mathrm{F}}^2 + \|(\mathbf{I} - \mathbf{Q})(\mathbf{D} - \mathbf{B})\|_{\mathrm{F}}^2 \\ &= \|\mathbf{Q}\mathbf{D} - \mathbf{Q}\mathbf{B}\|_{\mathrm{F}}^2 + \|(\mathbf{I} - \mathbf{Q})(\mathbf{M} - \mathbf{M}_r)\|_{\mathrm{F}}^2 \\ &\le 2\|\mathbf{Q}\mathbf{D}\|_{\mathrm{F}}^2 + 2\|\mathbf{Q}\mathbf{B}\|_{\mathrm{F}}^2 + \|\mathbf{M} - \mathbf{M}_r\|_{\mathrm{F}}^2 \\ &\le 4r\|\mathbf{D}\|_2^2 + 2\|\mathbf{B}\|_{\mathrm{F}}^2 + \|\mathbf{M} - \mathbf{M}_r\|_{\mathrm{F}}^2. \end{aligned} \tag{260}$$

In the first inequality, we used $\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2 \le 2\|\mathbf{A}\|_{\mathrm{F}}^2 + 2\|\mathbf{B}\|_{\mathrm{F}}^2$ and the fact that orthogonal projections contract the Frobenius norm. For the last inequality, we used

$$\|\mathbf{Q}\mathbf{D}\|_{\mathrm{F}}^2 \le \operatorname{rank}(\mathbf{Q}\mathbf{D})\|\mathbf{Q}\mathbf{D}\|_2^2 \le 2r\|\mathbf{D}\|_2^2, \tag{261}$$

together with $\|\mathbf{Q}\mathbf{B}\|_{\mathrm{F}} \le \|\mathbf{B}\|_{\mathrm{F}}$.

Finally, the Eckart–Young theorem (Thm. 10) gives

$$\|\mathbf{M} - \mathbf{M}_r\|_{\mathrm{F}}^2 = \sum_{i=r+1}^{\min\{d_x, d_y\}} \sigma_i^2(\mathbf{M}) = \sum_{i=r+1}^{\min\{d_x, d_y\}} \sigma_i^2 \left(\mathbf{\Delta}^\star \mathbf{\Sigma}_{\mathbf{xx}}^{\frac{1}{2}} \right). \tag{262}$$

Substituting the definitions of $\mathbf{D}$ and $\mathbf{B}$ proves the claim. □

VIII EXPERIMENTS

VIII.1 MODELS AND TASKS

We use the Qwen2.5 family at three scales $\{0.5, 1.5, 3\}$B. For evaluation, we use two datasets: BoolQ and CommonsenseQA. All evaluation is zero-shot via `lm-evaluation-harness` [11].

VIII.2 INPUT/OUTPUT FORMAT

Training data is in Alpaca form $\{\texttt{instruction}, \texttt{input}, \texttt{output}\}$. The loss is cross-entropy on the answer tokens only; the prompt span is masked with -100. For BoolQ:

> `instruction`: "Answer the following yes/no question based on the passage."
> `input`: "Passage: ...\n\nQuestion: ..."
> `output`: "Yes" | "No"

For CSQA:

> `instruction`: "Answer the following multiple-choice question. Choose the correct option letter only (A, B, C, D, or E)."
> `input`: "Question: ...\n\nOptions:\nA. ...\nB. ...\nE. ..."
> `output`: "A" | ... | "E"

VIII.3 TRAINING HYPERPARAMETERS COMMON TO ALL RUNS

All runs use bf16 precision with SDPA attention and gradient checkpointing, a sequence-length cutoff of 512, and train for 3 epochs with an effective batch size of 32. We use a cosine learning-rate schedule with linear warmup and clip gradients at $\|g\|_2 \leq 1$. Weight decay is 0.01 on all parameters except biases and LayerNorm/RMSNorm scales, for which it is set to 0. Per-device batch sizes and gradient-accumulation counts depend on model size and are reported in Tab. 2. Each job runs on a single A100 GPU. We hold out 20% of the training data for validation ($\texttt{val_size} = 0.2$); the n reported throughout refers to the training-set size *before* this split.

Table 2: Per-device batch ($\mathrm{bs}_{\mathrm{dev}}$) $\times$ gradient accumulation (ga). Effective batch size is always 32.

	0.5B	1.5B	3B
Full FT (AdamW or SGD)	8×4	2×16	1×32
LoRA (any rank)	8×4	8×4	4×8

VIII.4 FULL FINE-TUNING

Both full-FT optimizers update every parameter in the model.

AdamW. We use the HuggingFace defaults ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) with a warmup ratio of 0.02. A learning rate of $5\mathrm{e}{-5}$ was optimal at every cell (Tab. 3): we scanned $\{2, 5, 10\} \times 10^{-5}$ across model sizes and $5\mathrm{e}{-5}$ won in each case.

SGD with momentum. We use momentum 0.9 and a warmup ratio of 0.20. The optimal learning rate is strongly cell-dependent (Tab. 3) and was tuned per cell over $\{1, 3, 5, 8\}\times10^{-3} \cup \{1, 1.5, 2, 3\}\times 10^{-2}$. For 3B SGD we further raise the per-device batch size from 1 to 4 (correspondingly dropping ga from 32 to 8); this reduces per-step gradient variance and eliminates the seed-level collapses we observed at $\mathrm{lr} \geq 5\mathrm{e}{-3}$.

VIII.5 LORA CONFIGURATION

All LoRA runs use AdamW with the same defaults as above and a warmup ratio of 0.02. We sweep over ranks $r \in \{1, 8, 32, 64, 128\}$ and fix the scaling to $\alpha = 2r$, so the effective scale $\alpha/r = 2$ is

[11] https://github.com/EleutherAI/lm-evaluation-harness

Table 3: Per-cell best learning rates and warmup ratios. Full FT AdamW uses $5e{-5}$ everywhere. SGD requires a much larger and size-dependent LR. LoRA values are reproduced for reference; LoRA at $r = 128$ uses a halved LR.

Cell	Full FT AdamW	Full FT SGD	LoRA $r \in \{1, 8, 32, 64\}$	LoRA $r = 128$
0.5B BoolQ	$5e{-5}$	$1e{-2}$	$2e{-4}$	$1e{-4}$
0.5B CSQA	$5e{-5}$	$8e{-3}$	$2e{-4}$	$1e{-4}$
1.5B BoolQ	$5e{-5}$	$1.5e{-2}$	$2e{-4}$	$1e{-4}$
1.5B CSQA	$5e{-5}$	$1.5e{-2}$	$2e{-4}$	$1e{-4}$
3B BoolQ	$5e{-5}$	$2e{-3}$	$2e{-4}$	$1e{-4}$
3B CSQA	$5e{-5}$	$2e{-3}$	$2e{-4}$	$1e{-4}$
warmup ratio	0.02	0.20	0.02	0.02

constant and the per-step adapter contribution $\frac{\alpha}{r} BA$ has the same magnitude across ranks. Adapter dropout is disabled.

Adapters are placed on `all-linear` targets—every linear projection in every transformer block, i.e. the attention projections $\{W_q, W_k, W_v, W_o\}$ together with the MLP projections $\{W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\}$. The base weights are frozen and only the low-rank factors A, B are trained. We follow the HuggingFace PEFT defaults: A is drawn from $\mathcal{N}(0, 1/r)$ and B is initialized to zero.

For $r \in \{1, 8, 32, 64\}$ we use a learning rate of $2e{-4}$; at $r = 128$ we lower it to $1e{-4}$. For each (cell, rank) configuration we report 3 training seeds, used for both the sample-size sweeps and the final accuracy means.

VIII.6 Model averaging (SGD only)

For SGD we report a uniform average of fully-trained checkpoints,

$$W_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} W_{\text{seed}_i},$$

computed as an arithmetic mean in fp32 and recast to bf16. We train $N = 4$ independent seeds per cell at the cell's best learning rate. We do *not* average AdamW seeds, instead, we report best-of-N for AdamW comparisons.

VIII.7 Data augmentation

We paraphrase BoolQ using GPT-5.4-mini. For each original (passage, question) pair, the model is asked to rewrite both the passage and the question while preserving the gold yes/no label as well as all named entities and numerical values. Every candidate paraphrase is then *validated* by re-asking the same model the rewritten question against the rewritten passage at temperature 0; pairs whose re-answer disagrees with the original gold label are discarded. This yields $8{,}510$ kept paraphrases out of $9{,}427$ originals—a retention rate of 90.3%. The augmented BoolQ training set is the concatenation of the originals and the validated paraphrases, for a total of $17{,}937$ examples.

VIII.8 Runs on the augmented dataset

The augmented BoolQ runs reuse the training configuration above without modification; the only difference is the training-set size.

VIII.9 Label-noise injection

We consider two flipping protocols, both using a fixed noise seed so that the corrupted subsets are nested across noise levels—the examples flipped at $p = 5\%$ are a strict subset of those flipped at $p = 10\%$, and so on.

- **BoolQ:** flip the binary label on $p\%$ of the training examples.
- **CSQA:** for each example selected for flipping, replace the correct answer letter with one of the four incorrect letters drawn uniformly at random.

In both cases the resulting label distribution remains close to balanced. We sweep $p \in \{5, 10, 20, 30\}\%$, and otherwise keep the per-cell hyperparameters of Tab. 3 unchanged—only the data file is swapped. We train 3 seeds per (cell, method, noise level) and report the mean together with the min–max band.

VIII.10 Compute

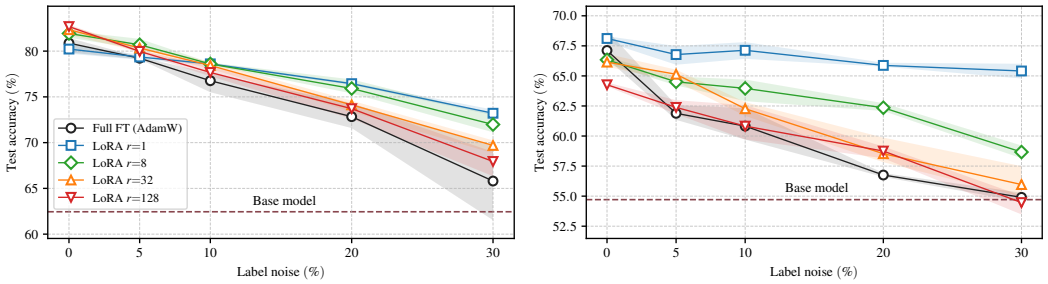
All training jobs run on a single NVIDIA A100 (40 GB or 80 GB). The 3B full-FT runs require the 80 GB variant; everything else fits comfortably on 40 GB. Per-seed wall-clock training time on the original training set is roughly 13 min at 0.5B, 36 min at 1.5B, and 80 min at 3B; on augmented BoolQ ($n = 17{,}937$) these times roughly double.

VIII.11 High-performing reference model

To obtain the high-performing reference model for each dataset, we first generated additional samples that match the same format as described above in Sec. VIII.7. Then, we applied two training techniques on the extended training set: (1) FFT with AdamW and (2) training a few times using SGD with different seeds, and averaging the resulting weights to get a better model, as explained in Sec. VIII.6. The reference model was chosen as the one with the better test performance.

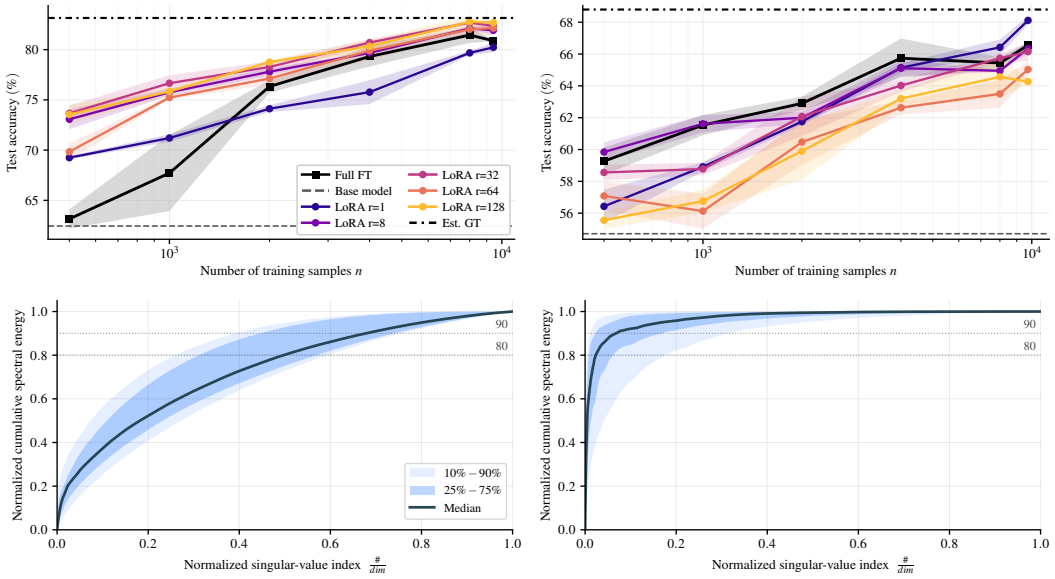
VIII.12 ADDITIONAL LLM EXPERIMENTS

In this section, we provide the plots for the LLM experiment form Sec. 4, including all model sizes (0.5B, 1.5B, and 3B). The behavior for all models is the same as described in the main text.



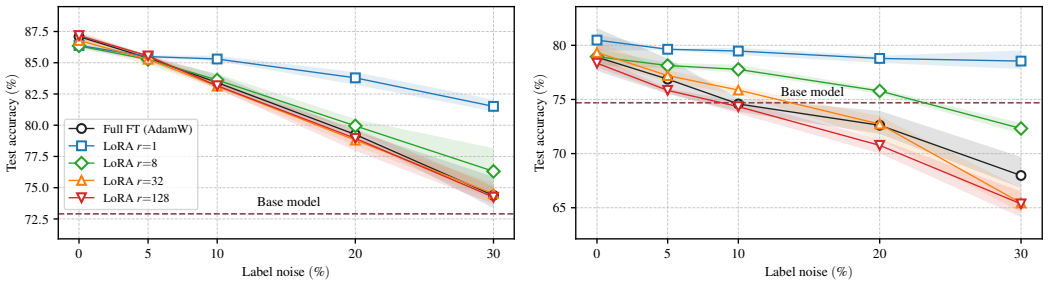
(a) Qwen2.5-0.5B on BoolQ (b) Qwen2.5-0.5B on CommonsenseQA

Figure 4: **Effect of label noise in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different levels of label noise. For each configuration, we trained 3 times using random seeds. Panels (a) and (b) show the mean and the range of the results for the 0.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. Here, in the strong-noise regime, LoRA outperforms FFT as predicted by our theory.



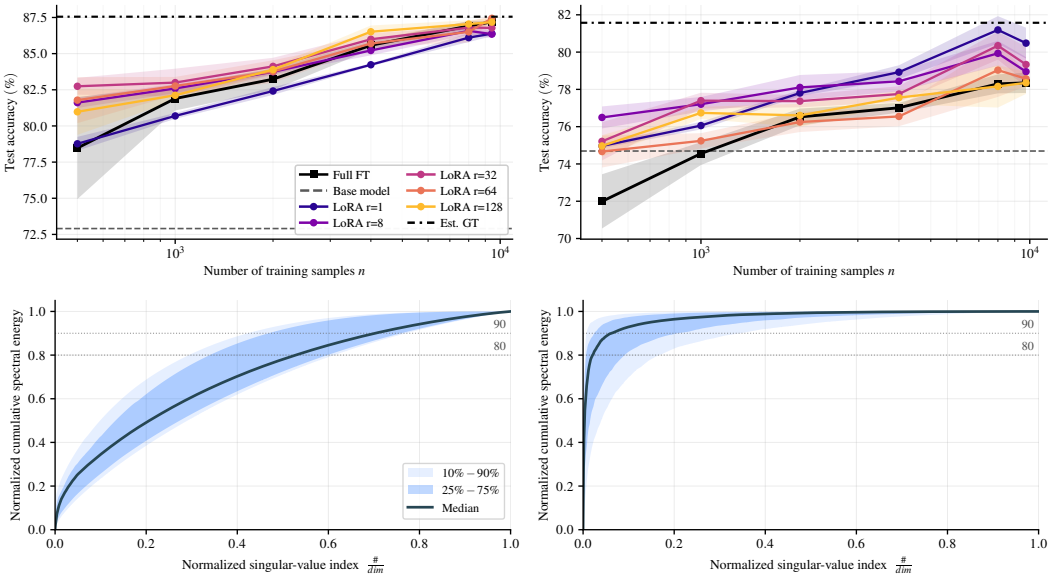
(a) Qwen2.5-0.5B on BoolQ (b) Qwen2.5-0.5B on CommonsenseQA

Figure 5: **Effect of sample size in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different sample sizes. For each configuration, we trained 3 times using random seeds. The top of Panels (a) and (b) shows the mean and the range of the results for the 0.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. For each task, we estimated $\mathbf{\Delta}^\star$ and computed its singular values per layer. The bottom part shows the statistics of the spectral cumulative energy of $\mathbf{\Delta}^\star$ across layers. When the effective rank of the adaptation is low (CommonsenseQA), LoRA outperforms FFT, as predicted by our theory.



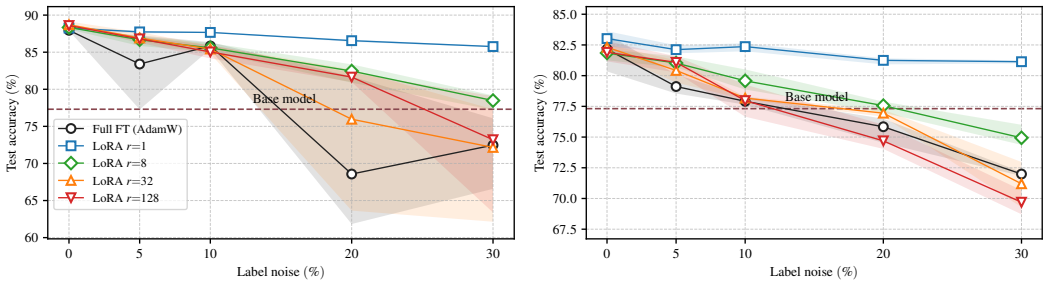
(a) Qwen2.5-1.5B on BoolQ (b) Qwen2.5-1.5B on CommonsenseQA

Figure 6: **Effect of label noise in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different levels of label noise. For each configuration, we trained 3 times using random seeds. Panels (a) and (b) show the mean and the range of the results for the 1.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. Here, in the strong-noise regime, LoRA outperforms FFT as predicted by our theory.



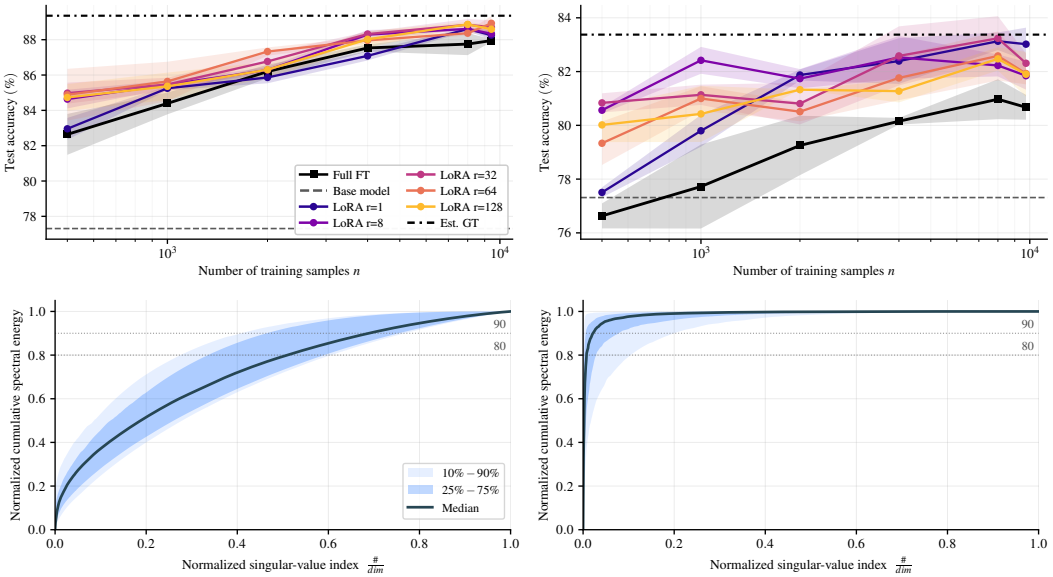
(a) Qwen2.5-1.5B on BoolQ (b) Qwen2.5-1.5B on CommonsenseQA

Figure 7: **Effect of sample size in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different sample sizes. For each configuration, we trained 3 times using random seeds. The top of Panels (a) and (b) shows the mean and the range of the results for the 1.5B model fine-tuned on BoolQ and CommonsenseQA, respectively. For each task, we estimated $\mathbf{\Delta}^\star$ and computed its singular values per layer. The bottom part shows the statistics of the spectral cumulative energy of $\mathbf{\Delta}^\star$ across layers. When the effective rank of the adaptation is low (CommonsenseQA), LoRA outperforms FFT, as predicted by our theory.



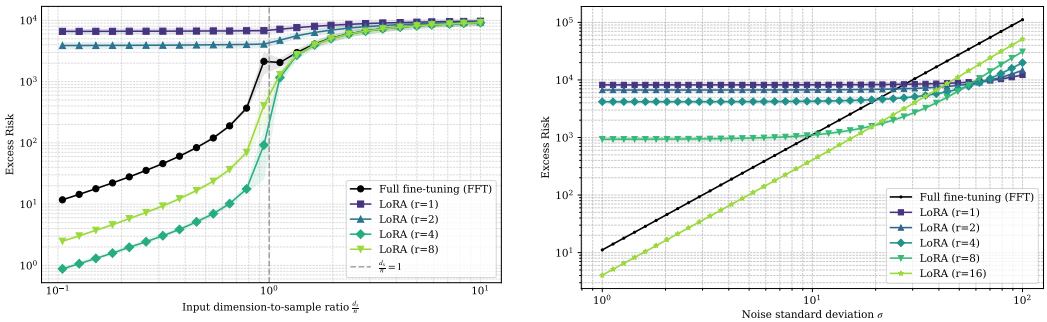
(a) Qwen2.5-3B on BoolQ (b) Qwen2.5-3B on CommonsenseQA

Figure 8: **Effect of label noise in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different levels of label noise. For each configuration, we trained 3 times using random seeds. Panels (a) and (b) show the mean and the range of the results for the 3B model fine-tuned on BoolQ and CommonsenseQA, respectively. Here, in the strong-noise regime, LoRA outperforms FFT as predicted by our theory.



(a) Qwen2.5-3B on BoolQ (b) Qwen2.5-3B on CommonsenseQA

Figure 9: **Effect of sample size in LLMs fine-tuning.** We fine-tuned Qwen2.5 models using LoRA with various ranks across different sample sizes. For each configuration, we trained 3 times using random seeds. The top of Panels (a) and (b) shows the mean and the range of the results for the 3B model fine-tuned on BoolQ and CommonsenseQA, respectively. For each task, we estimated $\mathbf{\Delta}^\star$ and computed its singular values per layer. The bottom part shows the statistics of the spectral cumulative energy of $\mathbf{\Delta}^\star$ across layers. When the effective rank of the adaptation is low (CommonsenseQA), LoRA outperforms FFT, as predicted by our theory.



(a) Excess risk vs. the dimension-to-sample ratio

(b) Excess risk vs. noise level

Figure 10: **Linear regression experiments.** Panel (a) presents the excess risk of FFT and LoRA under varying sample size n (decreasing from left to right), fixed dimensions $d_x = d_y = 100$, noise magnitude $\sigma = 1$, and true task-shift rank $\mathrm{rank}(\mathbf{\Delta}^\star) = 4$. Panel (b) plots the excess risk as a function of noise level $\sigma \in [1, 100]$ with $d_x = d_y = 100$, $n = 1000$, and $\mathrm{rank}(\mathbf{\Delta}^\star) = 10$. Results are averaged over 100 random seeds; shaded regions indicate ± 1 standard deviation.

VIII.13 Additional linear regression experiments

In this section, we present additional linear regression experiments to further illustrate the behavior of LoRA and FFT. We first study the effect of the rank of $\mathbf{\Delta}^\star$ on fine-tuning performance. Next, we repeat the experiment of Figure 1 in the low-noise and overdetermined regime (corresponding to Figure 1(a) and Figure 1(b)) and observe a similar qualitative pattern. Finally, we show that even when $\mathbf{\Delta}^\star$ is full-rank, the decay of its singular values plays a key role in determining whether LoRA or FFT performs better, highlighting the importance of the spectral tail of $\mathbf{\Delta}^\star$, which also appears explicitly in our theoretical bounds.

Figure 10 illustrates the behavior of FFT and LoRA in complementary experimental settings: the left panel shows the effect of varying dataset size in a low noise regime, while the right panel examines the effect of noise magnitude in the overdetermined regime.

Effect of dataset size in low-noise regime. The left panel (Figure 10(a)) displays excess risk as a function of the input dimension-to-sample ratio d_x/n in a low noise setting, with fixed dimensions, noise level, and $\mathrm{rank}(\mathbf{\Delta}^\star) = 4$. The gray dashed vertical line at $d_x/n = 1$ marks the interpolation threshold, separating the overdetermined regime (left, where $n > d_x$) from the underdetermined regime (right, where $n < d_x$).

In the overdetermined regime, LoRA with rank matching the true rank achieves the best performance: LoRA with $r = 4$ (matching $r^\star = 4$) exhibits the lowest excess risk, followed by LoRA with $r = 8$, and then FFT. This demonstrates that in low noise conditions, proper rank selection allows LoRA to outperform FFT even when sufficient samples are available. The superior performance of rank-matched LoRA arises because it correctly captures the low-rank structure of $\mathbf{\Delta}^\star$ without introducing unnecessary degrees of freedom. In contrast, FFT updates all directions, including irrelevant ones, leading to higher variance even in the overdetermined regime with low noise. For smaller ranks $r = 1$ and $r = 2$, which are insufficient to capture the true rank, the excess risk remains nearly constant and substantially higher throughout the overdetermined regime. This flat behavior reflects the dominance of the bias term induced by the rank constraint, as the spectral tail of $\mathbf{\Delta}^\star$ is significant and cannot be captured by such low ranks.

As the dimension-to-sample ratio increases toward the interpolation threshold, all methods experience increasing excess risk. At the threshold $d_x/n = 1$, FFT exhibits a sharp increase in risk as it transitions into the underdetermined regime. LoRA methods show more graceful degradation across the threshold, with the transition behavior depending on the rank. In the deeply underdetermined regime (large d_x/n), all methods eventually converge to similar excess risk levels determined by the bias of the initial model $\mathbf{A}_0$.

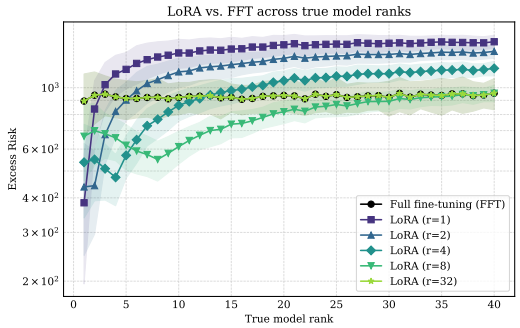


Figure 11: This figure illustrates the role of the central quantity $\mathrm{rank}(\mathbf{\Delta}^\star)$ in our analysis in underdetermined regime. We vary the rank of $\mathbf{\Delta}^\star$ from 1 to 40 (full rank) with a fixed number of training data $n = 30$, fixed dimension $d_x = d_y = 40$, and $\sigma = 2$. Results are reported across 100 random seeds, and shaded areas show the standard deviation.

Effect of noise magnitude in overdetermined regime. Figure 10(b) examines the effect of noise magnitude in the overdetermined regime with fixed dimensions and $\mathrm{rank}(\mathbf{\Delta}^\star)$. In this setting, when noise is low, FFT and high-rank LoRA (such as $r = 16$) achieve similar performance and outperform low-rank alternatives. However, as noise increases, FFT becomes increasingly sensitive, and its excess risk grows rapidly. In contrast, low-rank LoRA methods ($r = 1$, $r = 2$, and $r = 4$) demonstrate remarkable stability, with their excess risk remaining nearly constant across a wide range of noise levels. This robustness arises because the low-rank constraint prevents the estimator from fitting noise in irrelevant directions. At sufficiently high noise levels, low-rank LoRA outperforms FFT, highlighting the regularization benefit of the rank constraint even in the overdetermined regime.

Effect of the rank of true map on excess risk. Figure 11 shows how the excess risk depends on the rank of $\mathbf{\Delta}^\star$ in the underdetermined regime. When $\mathbf{\Delta}^\star$ is low-rank, LoRA with a suitable choice of rank r consistently achieves lower excess risk. As the rank of $\mathbf{\Delta}^\star$ increases and approaches the full-rank setting, FFT begins to outperform LoRA with small r, since low-rank updates do not provide enough flexibility to capture all relevant directions. Nevertheless, even when $\mathbf{\Delta}^\star$ is full-rank, choosing a sufficiently large r allows LoRA to achieve performance close to that of FFT. The exact behavior also depends on the noise level. Overall, these results indicate that the choice of fine-tuning method and LoRA rank should account for the intrinsic rank of the task shift, the data regime, and the noise level.

Full-rank $\mathbf{\Delta}^\star$ with varying spectral decay. We study how the singular value structure of $\mathbf{\Delta}^\star$ affects the relative performance of FFT and LoRA in the overdetermined regime. We fix $d_x = d_y = 40$ and construct $\mathbf{\Delta}^\star = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and the singular values follow

$$\sigma_i = 5\exp(-\lambda i), \tag{263}$$

with decay rate $\lambda \in [0, 2.5]$. When $\lambda = 0$, the spectrum is flat, and all directions are equally important; as λ increases, the spectrum becomes increasingly concentrated in the top singular directions.

Figure 12 reports the excess risk as a function of λ for $n = 200$ samples and noise level $\sigma_{\boldsymbol{\varepsilon}} = 0.5$. When the spectrum is flat ($\lambda = 0$), FFT significantly outperforms LoRA, since the task effectively requires modifying many directions and low-rank updates are too restrictive. As λ increases, the effective rank of $\mathbf{\Delta}^\star$ decreases, and LoRA with moderate rank begins to outperform FFT. In this regime, constraining the update to a low-dimensional subspace helps avoid fitting noise in weak directions while still capturing the dominant structure.

Overall, this experiment shows that the benefit of LoRA is governed not by the formal rank of $\mathbf{\Delta}^\star$, but by its spectral decay. When the update is spread across many directions, full fine-tuning is preferable; when it is concentrated in a few directions, low-rank adaptation yields better generalization. This is indeed aligned with our excess risk bounds for LoRA, which has an error term depending on the tail of $\mathbf{\Delta}^\star$. When this tail decays quickly, this error becomes negligible, while a flat tail could cause a large error.

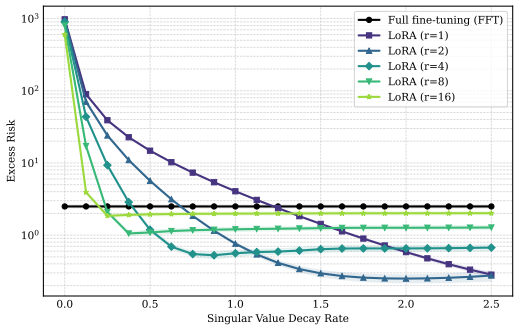


Figure 12: **Effect of spectral decay on FFT vs. LoRA.** Excess risk as a function of the singular value decay rate λ for a **full-rank** $\boldsymbol{\Delta}^\star$ with $\sigma_i = 5\exp(-\lambda i)$. A flat spectrum ($\lambda = 0$) favors FFT, while increasing spectral concentration leads to superior performance of LoRA with moderate rank. Settings: $d_x = d_y = 40$, $n = 200$, $\sigma_{\boldsymbol{\varepsilon}} = 0.5$, averaged over 100 runs.