LoRA Fine-Tuning as Reduced-Rank Regression in Frozen-Feature Models

Terry Ma
Carnegie Mellon University
Pittsburgh, PA
terryma@cs.cmu.edu

Abstract

Low-Rank Adaptation (LoRA) is a widely used parameter-efficient fine-tuning method, but its statistical role relative to full fine-tuning is not yet fully understood. This paper studies LoRA in a frozen-feature setting, where a pretrained backbone is fixed and only the final linear head is adapted. In this setting, we show that LoRA on the linear head is exactly multivariate reduced-rank regression (RRR) applied to the residuals of the pretrained model. This equivalence has two implications. First, it gives a precise statistical interpretation of LoRA: LoRA restricts the task-specific deviation from the pretrained head to lie in a low-dimensional subspace, yielding a bias-variance trade-off relative to unrestricted full fine-tuning. Under correct rank specification, the reduced-rank estimator has smaller first-order asymptotic variance than unrestricted least squares; under approximate low-rank structure, we obtain a non-asymptotic oracle inequality with an estimation term of order $\sigma^2 r(p+q)/n$ plus the approximation error from the neglected singular values. Second, the equivalence connects LoRA to the rich literature on reduced-rank regression. This connection suggests that classical RRR tools—including rank-selection criteria, spectral thresholding, nuclear-norm regularization, and adaptive reduced-rank methods—can inform practical choices in LoRA, such as selecting the rank r. Synthetic experiments illustrate the predicted bias-variance behavior and show that RRR rank-selection criteria can recover effective LoRA ranks; an exploratory DistilBERT experiment suggests that similar low-rank spectral structure can appear in standard transformer fine-tuning. Our results provide a statistical bridge between LoRA and classical low-rank multivariate regression, while leaving the analysis of simultaneous LoRA updates in deep nonlinear networks as an important direction for future work.

1 Introduction

The rise of large pre-trained foundation models—spanning language, vision, and multi-modal tasks—has fundamentally changed the landscape of machine learning [2, 6, 3, 7]. These models are typically pre-trained on large corpora and then adapted to downstream tasks through fine-tuning [2, 6]. Full fine-tuning, which updates all model parameters, is conceptually simple but often expensive in memory and computation, and in low-data regimes it may also overfit. Parameter-efficient fine-tuning (PEFT) methods have therefore become a practical alternative. Among them, **Low-Rank Adaptation (LoRA)** [12] is one of the most widely used: it freezes pre-trained weights and learns low-rank updates, often maintaining strong downstream performance while training far fewer parameters.

LoRA is usually motivated by efficiency, but its effectiveness also raises a statistical question: when can restricting an update to be low-rank improve estimation relative to an unrestricted update? This question is difficult to answer in full generality, because practical LoRA is often applied simultaneously to many nonlinear layers of a large model. We therefore study a controlled setting

STAI-X 2026 (Full Paper Track). Do not distribute.

in which an exact answer is possible: a frozen feature extractor followed by a trainable linear head. This setting includes a frozen multi-layer backbone with an adapted final head, but does not cover simultaneous LoRA updates throughout a deep nonlinear network. Our goal is not to prove that LoRA universally dominates full fine-tuning; rather, it is to identify a regime in which the statistical role of LoRA can be characterized exactly.

The key observation is that, in this frozen-feature linear-head setting, LoRA is exactly multivariate reduced-rank regression (RRR) [13, 22]. After subtracting the prediction of the frozen pre-trained head, full head fine-tuning is ordinary least squares, whereas LoRA solves the corresponding rank-constrained least-squares problem. Equivalently, LoRA estimates a low-rank task-specific deviation from the pre-trained head, while full head fine-tuning estimates this deviation without a rank constraint. Thus, in this setting, LoRA is not merely a parameter-saving reparameterization: it is a classical reduced-complexity estimator. This equivalence has two consequences that motivate the rest of the paper. First, it gives a precise statistical lens for comparing LoRA with unrestricted head fine-tuning. When the downstream correction is concentrated in a small number of singular directions, the rank constraint can reduce estimation variance; when important directions are discarded, it incurs approximation bias. The LoRA–full-tuning comparison therefore becomes a familiar bias–variance trade-off, and reduced-rank regression provides the mathematical language for quantifying it. Second, the equivalence brings the mature methodology of RRR to LoRA. A central practical question in LoRA is how to choose the rank r. In current practice, r is often selected by heuristic convention, grid search, or validation. By contrast, rank determination has been studied extensively in reduced-rank regression through information criteria, sequential testing, spectral thresholding, nuclear-norm penalization, and adaptive spectral regularization [22, 4, 8, 5, 20]. Thus, the LoRA–RRR connection is not only an analytical device for proving risk bounds; it also suggests concrete routes toward data-driven and adaptive rank selection for LoRA.

Building on this reduction, we establish two complementary statistical results. Under exact low-rank adaptation, we use the classical RRR viewpoint to obtain an asymptotic efficiency comparison between the rank-constrained estimator and unrestricted least squares (Theorem 3.2). This result formalizes the variance reduction obtained by restricting estimation to the tangent space of the rank-r matrix manifold under correct rank specification. Under approximate low-rank adaptation, we prove a non-asymptotic oracle inequality (Theorem 3.3). The bound separates the error into an approximation term, given by the spectral tail beyond rank r, and an estimation term of order $\sigma^2 r(p+q)/n$, reflecting the effective dimension of the low-rank update. Together, these results quantify when low-rank adaptation can be statistically favorable: namely, when the variance reduction from the rank constraint outweighs the bias from neglected singular directions. The scope of our theory is intentionally specific. The exact RRR equivalence holds whenever the features entering the adapted linear head are fixed. Therefore, the feature extractor may be a two-layer network, a deep pre-trained backbone, or any fixed representation map. However, the equivalence does not directly cover the standard setting in which LoRA updates are inserted simultaneously into multiple nonlinear layers of a transformer. We view the frozen-head setting as a clean statistical model that isolates one mechanism behind LoRA, rather than as a complete theory of all LoRA fine-tuning. Our experiments are designed to illustrate this statistical picture. Synthetic experiments in the exact regression model verify the predicted bias–variance behavior under both exact and approximate low-rank structure. We also show that classical RRR rank-selection criteria can recover effective LoRA ranks in the same setting. Finally, an exploratory DistilBERT experiment on AG News suggests that full fine-tuning updates in a standard transformer can exhibit rapidly decaying singular spectra, providing empirical motivation for low-rank adaptation beyond the exact frozen-head regression model. These experiments are not intended as large-scale benchmarking; rather, they show that the RRR viewpoint is informative in the regime covered by the theory and suggest where it may remain useful beyond that regime.

In summary, our contributions are fourfold. First, we show that LoRA fine-tuning of a frozen linear head is exactly multivariate RRR applied to the residuals of the pre-trained model. Second, we use this equivalence to interpret the difference between LoRA and full fine-tuning as a bias–variance trade-off. Under exact low-rank structure, we obtain an asymptotic efficiency comparison with unrestricted least squares; under approximate low-rank structure, we prove a non-asymptotic oracle inequality. Third, we argue that the LoRA–RRR equivalence opens a methodological path from reduced-rank regression to LoRA. In particular, classical tools for rank selection, nuclear-norm penalization, and adaptive reduced-rank estimation provide principled candidates for choosing or adapting the LoRA

rank r. Fourth, we provide synthetic experiments, RRR-based rank-selection experiments, and an exploratory DistilBERT study that illustrate the predicted bias–variance behavior and the relevance of low-rank spectral structure beyond the exact frozen-head model.

2 Problem Setup

We study LoRA in the simplest setting where its connection to reduced-rank regression is exact: a frozen feature map followed by a trainable linear head under squared loss. This setup should be viewed as a tractable model of *head-only adaptation*, not as an exact model of simultaneous LoRA updates across many nonlinear layers of a modern transformer. Let $\phi : \mathcal{X} \to \mathbb{R}^p$ denote the frozen feature map produced by a pre-trained backbone. For the motivating two-layer ReLU network, one may take $\phi(x) = \mathrm{ReLU}(W_1 x)$, but the analysis below uses only the fact that ϕ is fixed. Thus the feature extractor may be a single hidden layer or a deep frozen backbone; what matters for the theory is that only the final linear head is adapted.

To match standard multivariate regression notation, write $C = W^\top \in \mathbb{R}^{p\times q}$ for the coefficient matrix of the linear head. We assume the target task obeys

$$y = C^{*\top}\phi(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_q). \tag{1}$$

Let $C_0 = W_0^\top$ denote the pre-trained head, and decompose $C^* = C_0 + \Delta^*$, where $\Delta^* \in \mathbb{R}^{p\times q}$ is the task-specific deviation. The central structural question is whether this deviation is exactly or approximately low-rank. When it is, a low-rank update can reduce estimation variance relative to unrestricted head fine-tuning, at the cost of approximation bias if the deviation is not exactly rank r.

The theory in Section 3 is a Gaussian squared-loss regression analysis. The classification experiments in Section 4 are therefore intended as qualitative illustrations of the low-rank viewpoint, rather than literal instantiations of all assumptions below.

Approximate low-rank structure. Let $d_1(\Delta^*) \geq d_2(\Delta^*) \geq \cdots \geq d_{\min(p,q)}(\Delta^*) \geq 0$ denote the singular values of Δ^*. For any rank budget $r \leq \min(p, q)$, let $\Delta_r \in \arg\min_{\mathrm{rank}(\Delta)\leq r} \|\Delta - \Delta^*\|_F$ be a best rank-r approximation to Δ^*. By the Eckart–Young–Mirsky theorem [9], $\|\Delta^* - \Delta_r\|_F^2 = \sum_{j>r} d_j(\Delta^*)^2$. In the exact low-rank regime, $\mathrm{rank}(\Delta^*) \leq r$, so $\Delta_r = \Delta^*$. In the approximate regime, the tail energy $\sum_{j>r} d_j(\Delta^*)^2$ measures the approximation error incurred by restricting the update to rank r.

Fixed-design regression formulation. Given n samples (x_i, y_i), define the design matrix $X \in \mathbb{R}^{n\times p}$ and response matrix $Y \in \mathbb{R}^{n\times q}$ by

$$X = \begin{bmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_n)^\top \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{bmatrix}.$$

Then

$$Y = X(C_0 + \Delta^*) + E, \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \tag{2}$$

Throughout, we work conditionally on the realized design matrix X, assume $n > p$, and assume that X has full column rank. Define the residualized response $\widetilde{Y} := Y - XC_0$. Then the fine-tuning problem reduces to the multivariate linear model $\widetilde{Y} = X\Delta^* + E$. For any estimator $\widehat{\Delta}$, we compare estimators through the conditional Frobenius risk $R(\widehat{\Delta} \mid X) := \mathbb{E}\Big[\|\widehat{\Delta} - \Delta^*\|_F^2 \mid X\Big]$.

Full fine-tuning of the head. In this paper, "full fine-tuning" refers to unrestricted fine-tuning of the final linear head, not to updating all layers of the backbone. When the final head is updated without a rank constraint, the estimator is the ordinary least-squares solution $\widehat{\Delta}_{\mathrm{full}} = (X^\top X)^{-1} X^\top \widetilde{Y}$, $\widehat{C}_{\mathrm{full}} = C_0 + \widehat{\Delta}_{\mathrm{full}}$.

LoRA fine-tuning of the head. In standard LoRA, the update to $W \in \mathbb{R}^{q\times p}$ is parameterized as $W = W_0 + BA$, with $B \in \mathbb{R}^{q\times r}$, $A \in \mathbb{R}^{r\times p}$. The usual scalar factor α/r is omitted here because it

can be absorbed into A or B. Equivalently, the induced update in $C = W^\top$ is $\Delta = (BA)^\top = A^\top B^\top$, $\mathrm{rank}(\Delta) \le r$. Since only the product BA matters, the statistical analysis is naturally stated in terms of the induced update matrix Δ, rather than the non-identifiable factors A and B. Hence LoRA solves

$$\widehat{\Delta}_{\mathrm{LoRA}} \in \arg\min_{\mathrm{rank}(\Delta)\le r} \|\widetilde{Y} - X\Delta\|_F^2, \qquad \widehat{C}_{\mathrm{LoRA}} = C_0 + \widehat{\Delta}_{\mathrm{LoRA}}. \tag{3}$$

Equation (3) is the point at which LoRA becomes RRR: in the frozen-feature squared-loss model, LoRA is exactly rank-constrained multivariate least squares. Throughout the main theoretical analysis, r is treated as a fixed user-specified tuning parameter. One reason this equivalence is useful is that the RRR literature provides a rich set of tools for choosing or adapting r, including information criteria, spectral thresholding, and penalized low-rank estimators.

Closed-form reduced-rank solution. Let $\widehat{Y}_{\mathrm{OLS}} := X\widehat{\Delta}_{\mathrm{full}} = P_X\widetilde{Y}$, $P_X = X(X^\top X)^{-1}X^\top$. If $\widehat{Y}_{\mathrm{OLS}} = U\Sigma V^\top$ is a singular value decomposition, and V_r contains any orthonormal basis for a top-r right singular subspace, then one reduced-rank solution is

$$\widehat{\Delta}_{\mathrm{LoRA}} = \widehat{\Delta}_{\mathrm{full}} V_r V_r^\top. \tag{4}$$

If the r-th and $(r+1)$-st singular values tie, the minimizer need not be unique, but any such choice of V_r yields a valid rank-r solution. Equation (4) shows that LoRA keeps only the leading r response directions of the ordinary least-squares fitted values. This closed-form characterization is the formal link that allows the algorithms and theory of RRR to be brought into the study of LoRA.

3 Main Theoretical Results

In this section, we present three core results for the frozen-feature model of Section 2. The results serve two purposes. First, they make precise the equivalence between LoRA and reduced-rank regression (RRR), thereby giving a statistical interpretation of LoRA relative to unrestricted head fine-tuning. Second, they show how the RRR viewpoint brings classical tools—including bias–variance analysis, oracle inequalities, and rank-selection methodology—into the study of LoRA. Theorem 3.1 is an exact finite-sample equivalence result. Theorem 3.2 is an asymptotic comparison under correct rank specification. Theorem 3.3 is a non-asymptotic oracle inequality for approximate low-rank adaptation. We state the results in this order to distinguish clearly what is exact in finite samples from what requires asymptotics.

3.1 Equivalence to reduced-rank regression

We first make the connection between LoRA and reduced-rank regression completely explicit. This equivalence is the formal bridge that allows us to interpret LoRA using the statistical theory and methodology of RRR.

Theorem 3.1 (Equivalence to reduced-rank regression)**.** *Assume $X \in \mathbb{R}^{n\times p}$ has full column rank. Under the centered model associated with* (2)*, with $\widetilde{Y} = Y - XC_0$, the LoRA estimator*

$$\widehat{\Delta}_{\mathrm{LoRA}} = \arg\min_{\mathrm{rank}(\Delta)\le r} \|\widetilde{Y} - X\Delta\|_F^2$$

coincides with the rank-r reduced-rank regression estimator applied to $\widetilde{Y}$. In particular, let

$$\widehat{\Delta}_{\mathrm{full}} = (X^\top X)^{-1}X^\top\widetilde{Y}, \qquad \widehat{Y}_{\mathrm{OLS}} = X\widehat{\Delta}_{\mathrm{full}} = U\Sigma V^\top$$

be the unrestricted least-squares estimator and the singular value decomposition of its fitted values. If $V_r = V_{[:,1:r]}$ denotes any choice of leading r right singular vectors, then a rank-r LoRA minimizer is $\widehat{\Delta}_{\mathrm{LoRA}} = \widehat{\Delta}_{\mathrm{full}} V_r V_r^\top$. Equivalently, $X\widehat{\Delta}_{\mathrm{LoRA}}$ is a best rank-r Frobenius-norm approximation to $\widehat{Y}_{\mathrm{OLS}}$. If the r-th and $(r+1)$-st singular values are distinct, this fitted matrix is unique.

Theorem 3.1 says that, in the frozen-head squared-loss model, LoRA is exactly RRR. Consequently, full head fine-tuning corresponds to unrestricted least squares, while LoRA corresponds to imposing a rank constraint on the task-specific deviation from the pre-trained head.

3.2 Exact low-rank adaptation: asymptotic efficiency under correct rank specification

We now assume that the deviation is exactly of rank r and consider the *correctly specified* reduced-rank model: $\mathrm{rank}(\Delta^*) = r$. Because the reduced-rank estimator is a nonlinear projection of the unrestricted least-squares estimator, the comparison here is asymptotic rather than finite-sample.

Theorem 3.2 (Exact low-rank adaptation: asymptotic comparison)**.** *Suppose p, q, r are fixed, $r < \min(p, q)$, $n^{-1}X^\top X \to Q \succ 0$, and the errors E have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Assume $\mathrm{rank}(\Delta^*) = r$, and define $\Theta^* = Q^{1/2}\Delta^*$. Assume that the nonzero singular values of Θ^* are bounded away from zero, so that the rank-r model is locally identifiable. Let $\mathcal{T}$ denote the tangent space of the rank-r matrix manifold at Θ^*, and let $P_{\mathcal{T}}$ be the matrix representation of the orthogonal projection onto $\mathcal{T}$ under column-wise vectorization. Then:*

(a) ***Consistency.*** *Both $\widehat{\Delta}_{\mathrm{full}}$ and $\widehat{\Delta}_{\mathrm{LoRA}}$ are consistent for Δ^*.*

(b) ***Asymptotic law of full fine-tuning.***

$$\sqrt{n}\ \mathrm{vec}\Big(\widehat{\Delta}_{\mathrm{full}} - \Delta^*\Big) \Rightarrow \mathcal{N}\big(0,\ \sigma^2(I_q \otimes Q^{-1})\big)\,.$$

(c) ***Asymptotic law of LoRA.***

$$\sqrt{n}\ \mathrm{vec}\Big(\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^*\Big) \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{LoRA}}),$$

where

$$\Sigma_{\mathrm{LoRA}} = \sigma^2(I_q \otimes Q^{-1/2})P_{\mathcal{T}}(I_q \otimes Q^{-1/2}).$$

(d) ***First-order risk comparison.*** *Since $P_{\mathcal{T}} \preceq I_{pq}$,*

$$\Sigma_{\mathrm{LoRA}} \preceq \sigma^2(I_q \otimes Q^{-1}) = \Sigma_{\mathrm{full}}.$$

Consequently,

$$\mathrm{tr}(\Sigma_{\mathrm{LoRA}}) \le \sigma^2 q\ \mathrm{tr}(Q^{-1}).$$

Moreover, because $r < \min(p, q)$, the tangent space is a proper subspace, so the trace inequality is strict:

$$\mathrm{tr}(\Sigma_{\mathrm{LoRA}}) < \sigma^2 q\ \mathrm{tr}(Q^{-1}).$$

Thus LoRA has strictly smaller first-order asymptotic *Frobenius risk than unrestricted least squares under correct rank specification.*

This theorem is purely asymptotic. It does not assert finite-sample unbiasedness or finite-sample risk dominance of the nonlinear reduced-rank estimator. Its role is to show how the RRR equivalence explains the variance reduction of LoRA under correct rank specification: asymptotically, LoRA removes first-order noise directions orthogonal to the rank-r tangent space.

3.3 Approximate low-rank adaptation: oracle inequality

In practice, Δ^* need not have rank exactly r; it may only be *well approximated* by a rank-r matrix. The correct oracle comparator is therefore the best rank-r approximation to Δ^*, not Δ^* itself.

We use the following standard well-conditioned design assumption.

Assumption 3.1 (Well-conditioned design for low-rank estimation)**.** *There exist constants $0 < \kappa_0 \le \kappa_1 < \infty$ such that*

$$\kappa_0\|\Delta\|_F^2 \le \frac{1}{n}\|X\Delta\|_F^2 \qquad \textit{for all } \Delta \in \mathbb{R}^{p\times q} \textit{ with } \mathrm{rank}(\Delta) \le 2r,$$

and

$$\frac{1}{n}\|XA\|_F^2 \le \kappa_1\|A\|_F^2 \qquad \textit{for all } A \in \mathbb{R}^{p\times q}.$$

The lower restricted-isometry condition is needed only for differences of rank-r estimators, which have rank at most $2r$. The upper bound is stated for all matrices because the approximation residual $\Delta^* - \Delta_r$ need not be low rank. For fixed X, the second condition holds with $\kappa_1 = \lambda_{\max}(X^\top X/n)$.

Let Δ_r denote the best rank-r approximation to Δ^* in Frobenius norm, and let $d_1 \ge d_2 \ge \cdots \ge d_{\min(p,q)} \ge 0$ be the singular values of Δ^*, so that $\|\Delta^* - \Delta_r\|_F^2 = \sum_{j=r+1}^{\min(p,q)} d_j^2$.

Theorem 3.3 (Oracle inequality for approximate low-rank adaptation)**.** *Let the centered model $\widetilde{Y} = X\Delta^* + E$ hold with $E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Assume that X has full column rank and satisfies Assumption 3.1. Let Δ_r be the best rank-r approximation to Δ^* in Frobenius norm, and write $\|\Delta^* - \Delta_r\|_F^2 = \sum_{j=r+1}^{\min(p,q)} d_j^2$. Then, conditionally on X, there exist constants $c_1, c_2, c_3, c_4 > 0$, depending only on κ_0, κ_1, such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{\Delta}_{\mathrm{LoRA}} - \Delta_r\|_F^2 \le c_1\|\Delta^* - \Delta_r\|_F^2 + c_2\sigma^2\frac{r(p+q) + \log(1/\delta)}{n}.$$

Consequently, with the same probability,

$$\|\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^*\|_F^2 \le c_3 \sum_{j=r+1}^{\min(p,q)} d_j^2 + c_4\sigma^2\frac{r(p+q) + \log(1/\delta)}{n}. \tag{5}$$

Moreover,

$$\mathbb{E}\Big[\|\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^*\|_F^2 \,\Big|\, X\Big] \le c_3 \sum_{j=r+1}^{\min(p,q)} d_j^2 + C\sigma^2\frac{r(p+q)}{n},$$

for a constant $C > 0$ depending only on κ_0, κ_1.

For comparison, the unrestricted full head fine-tuning estimator satisfies the exact conditional risk identity $\mathbb{E}\Big[\|\widehat{\Delta}_{\mathrm{full}} - \Delta^*\|_F^2 \,\Big|\, X\Big] = \sigma^2 q\ \mathrm{tr}\big((X^\top X)^{-1}\big)$. Thus, when $n^{-1}X^\top X$ is well-conditioned and of constant order, full head fine-tuning has risk of order $\sigma^2 pq/n$, whereas the oracle bound for LoRA contains an estimation term of order $\sigma^2 r(p+q)/n$ plus the approximation error from the neglected singular values.

Theorem 3.3 makes the bias–variance trade-off explicit: in (5), the first term measures how well the true adaptation can be approximated by rank r, while the second term is the estimation cost of fitting a rank-r update. The bound therefore identifies the regime in which low-rank adaptation is statistically favorable: the spectral tail $\sum_{j>r} d_j^2$ must be small enough that the variance reduction relative to unrestricted head fitting outweighs the approximation bias. This is the same trade-off studied in reduced-rank regression, and it is precisely why RRR methodology is relevant for LoRA.

Remark on rank selection. Theorem 3.3 is an oracle bound for a *fixed* rank r. It does not claim that the optimal rank is known a priori. Rather, it characterizes the risk frontier as a function of r: increasing r reduces the approximation term but increases the estimation term. This observation provides a statistical rationale for importing rank-selection methods from RRR into LoRA. In particular, validation, information criteria, spectral thresholding, nuclear-norm penalization, and adaptive spectral regularization are natural candidates for choosing or adapting the LoRA rank. Developing such procedures for general deep LoRA is beyond the scope of this paper, but the exact LoRA–RRR equivalence gives a principled starting point in the frozen-head setting.

4 Experiments

The experiments are designed to support the two main messages of the paper. First, in the frozen-feature squared-loss setting, LoRA behaves exactly as RRR predicts: it trades approximation error against estimation variance. Second, because LoRA is equivalent to RRR in this setting, algorithms developed for RRR—especially rank-selection methods—can be used to choose the LoRA rank r.

We organize the experiments accordingly. The controlled synthetic study directly instantiates the regression model of Section 2 and is the primary empirical validation of the theory. We then add a data-driven rank-selection experiment, using classical RRR model-selection criteria to choose the LoRA rank.

Finally, we include an exploratory DistilBERT experiment on AG News to test whether similar low-rank spectral structure appears in a standard nonlinear LoRA setting. Because this experiment uses cross-entropy loss and applies LoRA inside a transformer block, it should be interpreted as a sanity check rather than as a literal test of the Gaussian squared-loss theory.

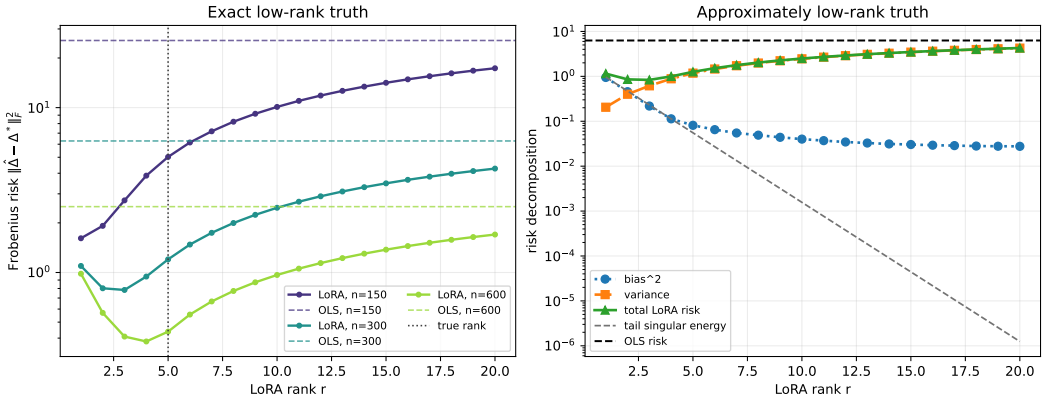

Figure 1: **Synthetic regression.** Mean Frobenius risk over 200 trials. Left: exact low-rank truth with $r^* = 5$, where the LoRA risk is minimized near the true rank. Right: approximately low-rank truth, showing the empirical bias–variance decomposition and the resulting U-shaped risk curve.

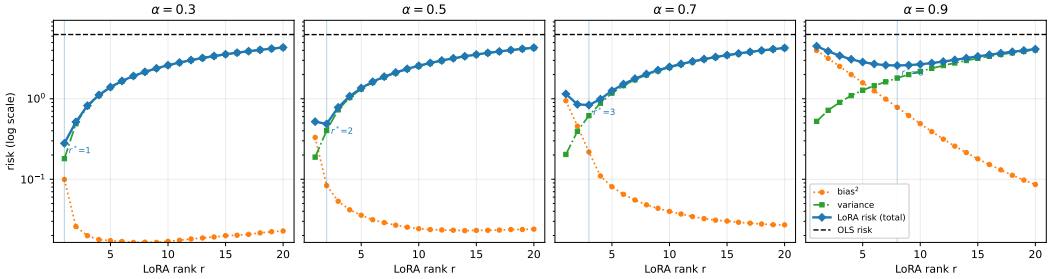

Figure 2: **Approximate low-rank regression with varying spectral decay.** As the spectrum $d_j = \alpha^{j-1}$ becomes flatter, the bias from truncation decays more slowly and the risk-minimizing rank shifts upward, matching the oracle trade-off in Theorem 3.3.

4.1 Synthetic regression: controlled bias–variance behavior

Setup. We generate data from (2) with $p = 100$, $q = 50$, $n \in \{150, 300, 600\}$, Gaussian design $X_{ij} \sim \mathcal{N}(0, 1)$, and noise variance $\sigma^2 = 0.25$. The pre-trained coefficient C_0 is fixed to a random Gaussian matrix, and the task-specific deviation is generated as $\Delta^* = UDV^\top$, where U, V have orthonormal columns and D has singular values $d_j = \alpha^{j-1}$. We consider an exact low-rank case, where D is truncated after $r^* = 5$ singular values, and an approximately low-rank case, where the full spectrum is retained with $\alpha = 0.7$. We compare unrestricted least squares $\widehat{\Delta}_{\text{full}}$ with LoRA-r, the reduced-rank estimator for $r = 1, \ldots, 20$, computed using (4). For each configuration, we report mean Frobenius error $\|\widehat{\Delta} - \Delta^*\|_F^2$ over 200 trials; in the approximately low-rank case, we also estimate empirical squared bias and variance.

Results. Figure 1 summarizes the synthetic results. In the exact low-rank case, the risk is minimized near the true rank $r^* = 5$, and the rank-constrained estimator attains lower mean risk than OLS in the sample sizes considered. In the approximately low-rank case, the LoRA risk has the predicted U-shape: small r underfits because of approximation error, while large r increases estimation variance. This reflects the two terms in Theorem 3.3: the spectral tail $\sum_{j>r} d_j^2$ and the estimation term of order $r(p+q)/n$. The advantage of LoRA over OLS is largest when n is small, consistent with the variance-reduction interpretation.

Figure 2 further tests the dependence on spectral decay by varying $d_j = \alpha^{j-1}$ with $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$. Fast decay leads to a small optimal rank, while flatter spectra shift the risk-minimizing rank upward. This supports the oracle trade-off in Theorem 3.3: increasing r reduces approximation error but increases estimation variance.

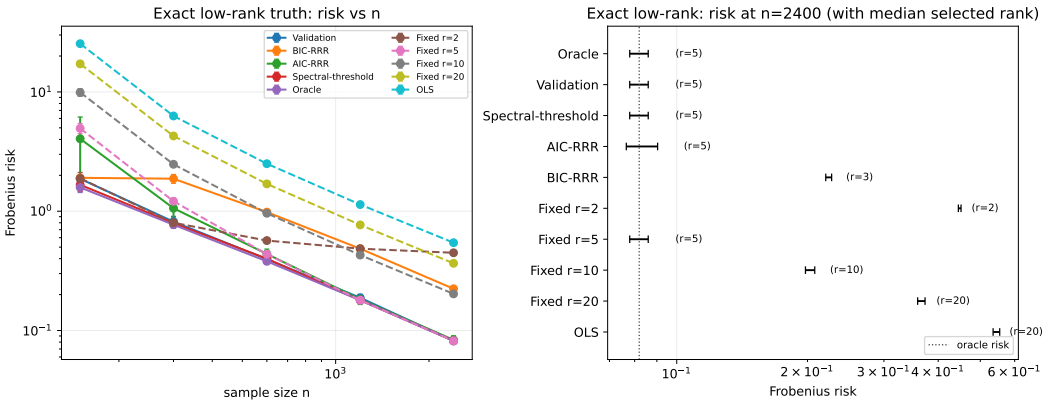

Figure 3: **RRR-based rank selection for LoRA under exact low-rank truth.** The true rank is $r^* = 5$. Left: Frobenius risk as a function of sample size. Validation selection, spectral thresholding, and AIC-RRR closely track the oracle risk and improve over unrestricted OLS and misspecified fixed ranks. Right: risk at $n = 2400$, with median selected ranks shown in parentheses. Validation, spectral thresholding, and AIC-RRR select median rank 5, while BIC-RRR is more conservative and underselects in this configuration.

4.2 RRR-based rank selection for LoRA

A key motivation for establishing the LoRA–RRR equivalence is that rank selection has been extensively studied in reduced-rank regression, whereas LoRA typically treats r as a manually chosen hyperparameter. We therefore test whether classical RRR rank-selection criteria can select effective LoRA ranks in the same frozen-feature regression model.

Rank-selection criteria. For each candidate rank r, let $\mathrm{RSS}(r) = \|\widetilde{Y} - X\widehat{\Delta}_r\|_F^2$, where $\widehat{\Delta}_r$ is the rank-r RRR/LoRA estimator. We consider the following RRR-inspired criteria:

1. **Validation selection:** choose r minimizing prediction error on a held-out validation set.
2. **BIC-RRR:** choose r minimizing $\mathrm{BIC}(r) = nq \log\left(\frac{\mathrm{RSS}(r)}{nq}\right) + \log(nq)\, r(p+q-r)$, where $r(p+q-r)$ is the dimension of the rank-r matrix manifold.
3. **AIC-RRR:** choose r minimizing $\mathrm{AIC}(r) = nq \log\left(\frac{\mathrm{RSS}(r)}{nq}\right) + 2r(p+q-r)$.
4. **Spectral thresholding:** choose r by thresholding the singular values of the whitened least-squares fit, using a noise-calibrated threshold inspired by optimal singular-value thresholding.

Evaluation. We compare the selected-rank estimators with oracle-rank LoRA, fixed-rank LoRA for $r \in \{2, 5, 10, 20\}$, and unrestricted OLS. We report selected rank, Frobenius risk, prediction error, and the frequency of selecting the true rank in the exact low-rank setting.

Results. Figure 3 reports the rank-selection experiment in the exact low-rank setting with $r^* = 5$. Validation selection, spectral thresholding, and AIC-RRR closely track the oracle risk across sample sizes. At $n = 2400$, all three methods have median selected rank 5, matching the true rank. Fixed ranks that are too small underfit, while ranks that are too large incur unnecessary variance. BIC-RRR is more conservative in this configuration and underselects the rank.

These results support the methodological implication of the LoRA–RRR equivalence: once LoRA is written as reduced-rank regression, classical rank-selection criteria can be used directly to choose the LoRA rank in the frozen-feature regression setting.

Table 1: **Exploratory DistilBERT/AG News experiment.** Test accuracy (%) for full fine-tuning and LoRA at different ranks. The best method for each sample size is boldfaced.

Method	$m = 16$	$m = 64$	$m = 256$
Full FT (last attn + head)	39.1	81.2	**88.6**
LoRA, $r = 1$	**49.1**	79.4	85.2
LoRA, $r = 4$	35.4	**82.4**	85.3
LoRA, $r = 16$	43.3	81.7	85.6

4.3 Exploratory transformer experiment

Finally, we include a small transformer experiment to test whether the low-rank spectral structure suggested by the regression theory also appears in a standard nonlinear LoRA setting. This experiment is outside the formal scope of Theorems 3.2–3.3: the model is nonlinear, the loss is cross-entropy, and LoRA is applied inside a transformer block rather than only to a frozen linear head. We therefore present this experiment as an exploratory sanity check, not as a direct validation of the theory.

Setup. We fine-tune DistilBERT on AG News using $m \in \{16, 64, 256\}$ labeled examples per class. We apply LoRA to the four attention projection matrices in the last transformer block and compare ranks $r \in \{1, 4, 16\}$ against full fine-tuning of the same attention projections and classification head. All methods are trained for four epochs with AdamW. We report test accuracy and inspect the singular-value spectrum of the corresponding full fine-tuning update.

Results. Table 1 shows that the best rank depends on the sample size. In the smallest-data setting, $m = 16$, rank-one LoRA obtains higher test accuracy than full fine-tuning, suggesting that the low-rank constraint can act as a useful regularizer. At $m = 64$, LoRA with $r = 4$ is slightly better than full fine-tuning. At $m = 256$, full fine-tuning performs best, consistent with the idea that the variance advantage of a low-rank constraint diminishes as more data become available.

These results should not be interpreted as showing that LoRA uniformly dominates full fine-tuning. Rather, they are consistent with the same qualitative bias–variance picture seen in the regression experiments: low ranks can help in very small-sample regimes, but may underfit when the available data are sufficient to estimate a richer update.

The full fine-tuning updates also have rapidly decaying singular spectra; see Figure 4 in Appendix B. This provides additional empirical motivation for low-rank adaptation beyond the exact frozen-head regression model, while remaining outside the formal scope of Theorems 3.2–3.3.

5 Discussion and Future Work

This paper develops a clean statistical benchmark for understanding LoRA through the lens of reduced-rank regression. In the frozen-feature, final-linear-head setting, LoRA is exactly RRR applied to the residuals of the pre-trained model. This equivalence gives two main insights: first, it interprets the comparison between LoRA and unrestricted head fine-tuning as a bias–variance trade-off; second, it makes the rich RRR literature directly relevant to LoRA, especially for rank selection, spectral regularization, and adaptive low-rank estimation. Our results should therefore be read as an exact theory for a deliberately restricted regime, not as a general proof that LoRA uniformly dominates full fine-tuning in deep nonlinear architectures.

Several directions remain open. The most immediate is to use RRR methodology to design practical LoRA rank-selection procedures, including information criteria, spectral thresholding, nuclear-norm penalties, and adaptive spectral regularization. A second direction is to extend the analysis beyond a single frozen head, for example through layerwise or linearized models of multi-layer LoRA where different layers may require different ranks. A third direction is empirical: larger regression-aligned benchmarks and modern LoRA studies could test whether the spectral decay and rank–risk trade-offs predicted by RRR remain useful in realistic architectures. These extensions would help determine how far the LoRA–RRR bridge can be pushed beyond the setting where the equivalence is exact.

Acknowledgments

We thank the STAI-X 2026 organizers for creating a venue at the interface of statistics and AI.

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2021.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 (NeurIPS)*, 2020.

[4] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, 39(2):1282–1309, 2011.

[5] Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[8] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[9] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. ISBN 978-1421407944.

[10] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019.

[12] 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 *International Conference on Learning Representations (ICLR)*, 2022.

[13] Alan J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[15] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[17] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 4582–4597, 2021.

[18] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

[19] Sadhika Malladi, Alexander Alexander Wettig, Dingli Yu, Danqi Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[20] Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[21] Eshaan Nichani, Alex Damian, and Jason D Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[22] Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, 2 edition, 2022.

[23] Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*, 2023.

[25] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[26] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[27] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

A Related Work

Parameter-efficient fine-tuning. Low-Rank Adaptation (LoRA) [12] is a widely used parameter-efficient fine-tuning method that freezes pre-trained weights and learns low-rank updates. It belongs to a broader family of parameter-efficient fine-tuning (PEFT) methods, including adapters [11], prompt tuning [16], and prefix tuning [17]; see He et al. [10] for a unified view. Most PEFT methods are motivated by reducing memory and computational cost, but their empirical success also raises a statistical question: why can a heavily constrained update often compete with full fine-tuning? Our paper addresses this question in a frozen-feature, linear-head setting by showing that the standard fixed-rank LoRA objective is exactly a classical reduced-rank regression estimator. Thus, our goal is complementary to algorithmic variants of LoRA: rather than proposing a new PEFT architecture, we identify a regime in which the statistical role of LoRA can be characterized precisely.

Adaptive and rank-varying LoRA methods. A practical issue in LoRA is the choice of the rank r. Several recent methods modify the original fixed-rank formulation to improve rank allocation or training stability, including dynamic rank search and adaptive budget allocation [24, 27], rank-stabilized LoRA [15], and weight-decomposed low-rank adaptation [18]. These works demonstrate that rank choice is important in practice, but the statistical principles underlying rank selection for LoRA remain less developed. One motivation for our LoRA–RRR equivalence is precisely to connect this issue to the reduced-rank regression literature, where rank determination has been studied extensively.

Low-dimensional views of fine-tuning. A recurring theme in the fine-tuning literature is that downstream adaptation may live in a much lower-dimensional space than the ambient parameter space. Aghajanyan et al. [1] argue that successful fine-tuning can have low intrinsic dimension, while Malladi et al. [19] study fine-tuning and PEFT through a kernelized, subspace-restricted lens in a linearized regime. More broadly, theoretical analyses of transfer learning and fine-tuning often rely on shared-representation or NTK-style approximations [14, 23, 21]. Our contribution is narrower in architectural scope but sharper at the estimator level: when the feature representation is frozen and only the final linear head is adapted, LoRA is not merely analogous to a low-dimensional estimator; it is exactly reduced-rank regression applied to the residuals of the pre-trained model. This exact identification allows us to interpret the difference between LoRA and full fine-tuning as a bias–variance trade-off.

Reduced-rank regression and low-rank estimation. Multivariate reduced-rank regression (RRR) constrains the coefficient matrix in multivariate regression to have low rank and admits a closed-form solution through truncated singular value decomposition [22]. Classical RRR theory studies efficiency gains over unrestricted least squares when the true coefficient matrix has low-rank structure, while modern low-rank estimation develops oracle inequalities, spectral thresholding rules, nuclear-norm relaxations, and adaptive spectral penalties [4, 8, 5, 20]. Our paper uses this literature in two ways. First, it uses RRR as a statistical lens for understanding when LoRA can improve over full fine-tuning: LoRA reduces variance by restricting the update to a low-rank matrix class, but incurs approximation bias when the task-specific deviation is not exactly low-rank. Second, it uses RRR as a methodological bridge: once LoRA is written as a rank-constrained regression problem, existing RRR tools become natural candidates for LoRA rank selection and adaptive regularization.

Rank selection and adaptive low-rank estimation. The rank r is a central hyperparameter in LoRA, and in current practice it is often chosen by heuristic convention, grid search, or validation. By contrast, the corresponding problem has a long history in reduced-rank regression and low-rank matrix estimation. Classical approaches include information criteria and sequential testing, while modern approaches include spectral thresholding, nuclear-norm penalization, and adaptive nuclear-norm or spectral regularization [22, 4, 8, 5, 20]. These methods are designed to balance approximation error against estimation variance, which is exactly the trade-off that appears in the LoRA–RRR formulation. Therefore, the equivalence established here suggests a principled route for importing rank-selection and adaptive low-rank methodology from RRR into LoRA. Our paper does not solve rank selection for general deep LoRA, but it identifies a setting in which these tools can be transferred exactly and studied rigorously.

Positioning of our contribution. We do not claim a general theory of LoRA for deep Transformers with multiple simultaneously adapted layers. Instead, we isolate a clean frozen-feature regime in which LoRA on the final linear head coincides exactly with a classical statistical estimator. Within this

regime, the contribution is twofold: the equivalence explains LoRA versus full fine-tuning through the bias–variance geometry of reduced-rank regression, and it opens a path for importing the rich algorithmic and methodological literature of RRR into LoRA.

B Additional transformer diagnostics

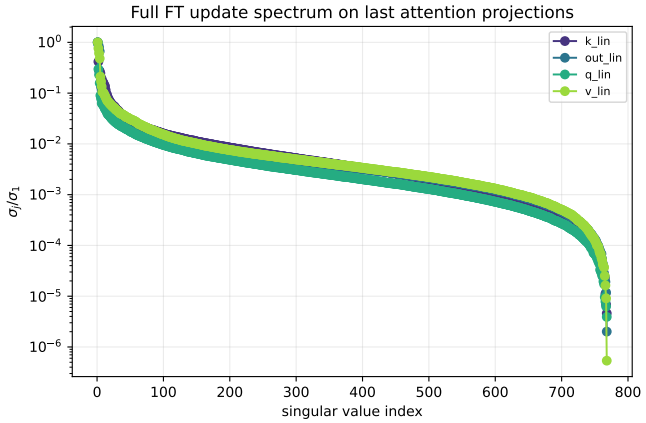

Figure 4: **Singular spectrum of transformer fine-tuning updates.** Normalized singular values of $\widehat{W}_{\text{full}} - W_0$ for the four attention projection matrices in the last DistilBERT block on AG News with $m = 256$. The spectra decay by several orders of magnitude, indicating that the full fine-tuning updates are highly compressible even in this nonlinear transformer setting.

C Additional MNIST frozen-feature experiment

As an additional qualitative diagnostic, we include a small frozen-feature transfer experiment on MNIST. This experiment is not a direct test of the Gaussian squared-loss theory, since it uses a nonlinear feature map and cross-entropy loss. Its purpose is to inspect whether the dense fine-tuning update has a rapidly decaying spectrum and whether moderate LoRA ranks recover most of the dense-head accuracy.

Setup. We use a two-layer network with $d = 784$ input pixels, $p = 256$ hidden units, and $q = 10$ output classes. The first layer $W_1 \in \mathbb{R}^{256\times 784}$ is pre-trained on digits 0–4, then frozen. The second-layer head is adapted on digits 5–9, with $m \in \{10, 50, 100\}$ target examples per class. We compare dense head tuning with LoRA ranks $r \in \{2, 5, 10, 20, 50\}$, all trained with cross-entropy loss.

Table 2: **MNIST frozen-feature transfer accuracy.** Test accuracy (%) on digits 5–9, averaged over repeated runs.

Method	$m = 10$	$m = 50$	$m = 100$
Full FT	80.7 ± 1.4	88.1 ± 0.2	90.8 ± 0.4
LoRA $r = 2$	49.3 ± 4.0	49.1 ± 2.8	67.5 ± 6.4
LoRA $r = 5$	74.4 ± 3.3	82.7 ± 1.9	88.9 ± 2.3
LoRA $r = 10$	78.4 ± 1.6	86.2 ± 0.6	90.3 ± 0.3
LoRA $r = 20$	79.7 ± 1.2	87.6 ± 0.4	90.5 ± 0.4
LoRA $r = 50$	79.3 ± 2.1	88.2 ± 0.4	90.9 ± 0.2

Discussion. The experiment supports the qualitative low-rank viewpoint but should be interpreted cautiously. At $m = 100$, LoRA with $r = 10$ nearly matches full head tuning, and larger factorization ranks are statistically comparable. However, because $q = 10$, the intrinsic rank of the head update is at most 10; ranks $r > 10$ are therefore overparameterized controls rather than stricter low-rank models. At $m = 10$, full head tuning performs best, showing that this nonlinear classification setting does not imply uniform LoRA dominance.

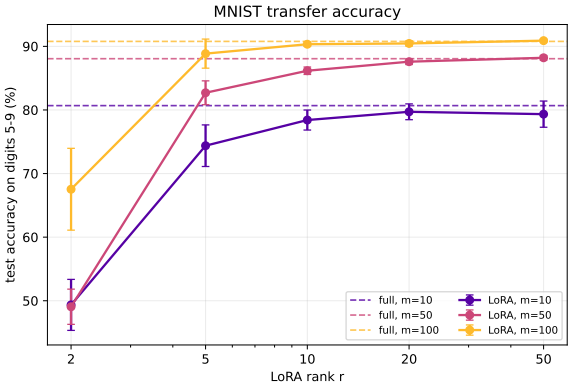

Figure 5: **MNIST transfer accuracy.** Moderate ranks recover much of the dense-head accuracy, while very small ranks underfit.

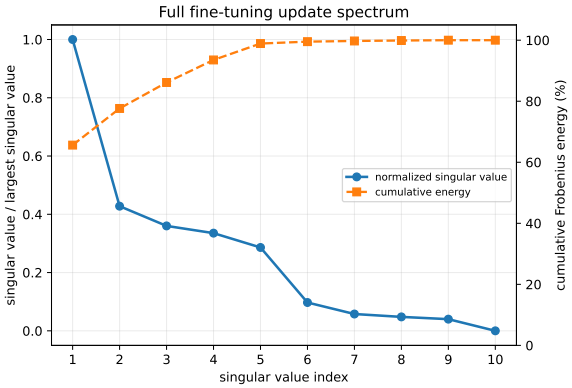

Figure 6: **Spectrum of the MNIST dense-head update.** The singular values of $W_{\text{full}} - W_0$ decay rapidly; the top five singular directions capture nearly all of the Frobenius energy.

D Proof of Theorem 3.1 (Equivalence to Reduced-Rank Regression)

We prove the equivalence by whitening the design. The centered response is

$$\widetilde{Y} = Y - XC_0.$$

For a candidate update Δ, the fitted residual response is $X\Delta$, and the LoRA estimator solves

$$\widehat{\Delta}_{\text{LoRA}} = \arg \min_{\text{rank}(\Delta) \le r} \|\widetilde{Y} - X\Delta\|_F^2.$$

Because X has full column rank, define

$$Z := X(X^\top X)^{-1/2} \in \mathbb{R}^{n \times p}, \qquad G := (X^\top X)^{1/2}\Delta \in \mathbb{R}^{p \times q}.$$

Then $Z^\top Z = I_p$, and

$$X\Delta = ZG, \qquad \text{rank}(G) = \text{rank}(\Delta).$$

Therefore the LoRA optimization problem is equivalent to

$$\widehat{G}_{\text{LoRA}} = \arg \min_{\text{rank}(G) \le r} \|\widetilde{Y} - ZG\|_F^2, \qquad \widehat{\Delta}_{\text{LoRA}} = (X^\top X)^{-1/2}\widehat{G}_{\text{LoRA}}.$$

Now decompose $\widetilde{Y}$ into its projection onto $\text{col}(Z) = \text{col}(X)$ and its orthogonal complement:

$$\|\widetilde{Y} - ZG\|_F^2 = \|(I - P_X)\widetilde{Y}\|_F^2 + \|Z^\top \widetilde{Y} - G\|_F^2, \qquad P_X = ZZ^\top = X(X^\top X)^{-1}X^\top.$$

The first term does not depend on G. Hence minimizing over matrices G with $\operatorname{rank}(G) \le r$ is equivalent to finding the best rank-r Frobenius-norm approximation to $Z^\top \widetilde{Y}$.

By the Eckart–Young–Mirsky theorem [9], the minimizer is obtained by truncating the singular value decomposition of $Z^\top \widetilde{Y}$. Moreover,

$$Z^\top \widetilde{Y} = (X^\top X)^{-1/2} X^\top \widetilde{Y} = (X^\top X)^{1/2} (X^\top X)^{-1} X^\top \widetilde{Y} = (X^\top X)^{1/2} \widehat{\Delta}_{\text{full}},$$

where

$$\widehat{\Delta}_{\text{full}} = (X^\top X)^{-1} X^\top \widetilde{Y}.$$

Let

$$\widehat{Y}_{\text{OLS}} = X \widehat{\Delta}_{\text{full}} = U \Sigma V^\top$$

be the singular value decomposition of the unrestricted fitted response. Since

$$\widehat{Y}_{\text{OLS}} = Z\big((X^\top X)^{1/2} \widehat{\Delta}_{\text{full}}\big)$$

and $Z^\top Z = I_p$, left multiplication by Z preserves Frobenius norms, nonzero singular values, and right singular vectors. Therefore $(X^\top X)^{1/2} \widehat{\Delta}_{\text{full}}$ and $\widehat{Y}_{\text{OLS}}$ have the same nonzero singular values and the same right singular vectors.

Thus, if $V_r = V_{[:,1:r]}$ denotes any leading r-dimensional right singular subspace of $\widehat{Y}_{\text{OLS}}$, a best rank-r approximation to $(X^\top X)^{1/2} \widehat{\Delta}_{\text{full}}$ is

$$\widehat{G}_{\text{LoRA}} = (X^\top X)^{1/2} \widehat{\Delta}_{\text{full}} V_r V_r^\top.$$

Multiplying by $(X^\top X)^{-1/2}$ gives

$$\widehat{\Delta}_{\text{LoRA}} = \widehat{\Delta}_{\text{full}} V_r V_r^\top.$$

Consequently,

$$X \widehat{\Delta}_{\text{LoRA}} = \widehat{Y}_{\text{OLS}} V_r V_r^\top,$$

which is a best rank-r Frobenius-norm approximation to $\widehat{Y}_{\text{OLS}}$. This is precisely the classical reduced-rank regression estimator applied to the centered response $\widetilde{Y}$; see, for example, Reinsel and Velu [22, Chapter 2]. □

E Proof of Theorem 3.2

We prove each part of the theorem in turn. Throughout this section, assume

$$\operatorname{rank}(\Delta^*) = r < \min(p, q), \qquad Q_n := \frac{1}{n} X^\top X \to Q \succ 0.$$

Let

$$\Theta^* := Q^{1/2} \Delta^*.$$

Write the singular value decomposition of the whitened target matrix as

$$\Theta^* = U D V^\top, \qquad D = \operatorname{diag}(d_1, \dots, d_r), \qquad d_r > 0.$$

The condition $d_r > 0$, together with the assumption that the nonzero singular values are bounded away from zero, ensures that the rank-r manifold is locally smooth around Θ^* and that the rank-r projection is locally well defined to first order.

Proof of part (b): asymptotic law of full fine-tuning. By standard least-squares theory,

$$\widehat{\Delta}_{\text{full}} = \Delta^* + (X^\top X)^{-1} X^\top E.$$

Therefore,

$$\sqrt{n}\, \operatorname{vec}\Big(\widehat{\Delta}_{\text{full}} - \Delta^*\Big) = \operatorname{vec}\left(Q_n^{-1} \frac{1}{\sqrt{n}} X^\top E\right).$$

Conditionally on X,

$$\frac{1}{\sqrt{n}} X^\top E$$

is a $p \times q$ Gaussian matrix with column covariance $\sigma^2 Q_n$. Hence

$$\sqrt{n}\ \mathrm{vec}\Big(\widehat{\Delta}_{\mathrm{full}} - \Delta^*\Big) \Rightarrow \mathcal{N}\big(0, \sigma^2 (I_q \otimes Q^{-1})\big)\,,$$

where we use column-wise vectorization. This proves part (b).

In particular,

$$\widehat{\Delta}_{\mathrm{full}} - \Delta^* = O_p(n^{-1/2}),$$

so $\widehat{\Delta}_{\mathrm{full}} \to \Delta^*$ in probability. This proves the full fine-tuning part of consistency in part (a).

Whitened form of the reduced-rank estimator. By Theorem 3.1, the LoRA estimator is the reduced-rank regression estimator. Equivalently,

$$Q_n^{1/2} \widehat{\Delta}_{\mathrm{LoRA}}$$

is the rank-r truncated SVD of

$$Q_n^{1/2} \widehat{\Delta}_{\mathrm{full}}.$$

Indeed, multiplying by $Q_n^{1/2}$ differs from multiplying by $(X^\top X)^{1/2}$ only by the scalar $n^{-1/2}$, which does not change the singular vectors or the rank-r truncation.

Since $\widehat{\Delta}_{\mathrm{full}} \to \Delta^*$ and $Q_n^{1/2} \to Q^{1/2}$, we have

$$Q_n^{1/2} \widehat{\Delta}_{\mathrm{full}} \to Q^{1/2} \Delta^* = \Theta^*$$

in probability. Because Θ^* has rank r with $d_r > 0$, the rank-r truncation map is continuous at Θ^*. Therefore,

$$Q_n^{1/2} \widehat{\Delta}_{\mathrm{LoRA}} \to \Theta^* = Q^{1/2} \Delta^*$$

in probability. Since $Q_n^{-1/2} \to Q^{-1/2}$, this implies

$$\widehat{\Delta}_{\mathrm{LoRA}} \to \Delta^*$$

in probability. Together with the consistency of $\widehat{\Delta}_{\mathrm{full}}$, this proves part (a).

Proof of part (c): asymptotic law of LoRA. Let $\mathcal{T}$ be the tangent space of the rank-r matrix manifold at $\Theta^* = UDV^\top$. Explicitly,

$$\mathcal{T} = \{UA^\top + BV^\top : A \in \mathbb{R}^{q \times r},\ B \in \mathbb{R}^{p \times r}\}.$$

The orthogonal projection of a perturbation $H \in \mathbb{R}^{p \times q}$ onto $\mathcal{T}$ is

$$\mathcal{P}_{\mathcal{T}}(H) = UU^\top H + HVV^\top - UU^\top HVV^\top.$$

The rank-r truncation map is differentiable at Θ^* to first order, and its derivative is the orthogonal projector $\mathcal{P}_{\mathcal{T}}$. Hence,

$$Q_n^{1/2} \widehat{\Delta}_{\mathrm{LoRA}} - Q_n^{1/2} \Delta^* = \mathcal{P}_{\mathcal{T}} \left(Q_n^{1/2} \widehat{\Delta}_{\mathrm{full}} - Q_n^{1/2} \Delta^* \right) + o_p(n^{-1/2}).$$

Multiplying by $\sqrt{n}$, and using the whitened version of the OLS asymptotic normality,

$$\sqrt{n}\ \mathrm{vec}\Big(Q_n^{1/2} (\widehat{\Delta}_{\mathrm{full}} - \Delta^*) \Big) \Rightarrow \mathcal{N}(0, \sigma^2 I_{pq}),$$

we obtain

$$\sqrt{n}\ \mathrm{vec}\Big(Q_n^{1/2} (\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^*) \Big) \Rightarrow P_{\mathcal{T}} Z, \qquad Z \sim \mathcal{N}(0, \sigma^2 I_{pq}),$$

where $P_{\mathcal{T}}$ is the matrix representation of $\mathcal{P}_{\mathcal{T}}$ under column-wise vectorization.

Undoing the whitening gives

$$\sqrt{n}\ \mathrm{vec}\Big(\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^* \Big) = (I_q \otimes Q_n^{-1/2}) \sqrt{n}\ \mathrm{vec}\Big(Q_n^{1/2} (\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^*) \Big)\,.$$

Since $Q_n^{-1/2} \to Q^{-1/2}$, Slutsky's theorem yields

$$\sqrt{n}\ \mathrm{vec}\Big(\widehat{\Delta}_{\mathrm{LoRA}} - \Delta^* \Big) \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{LoRA}}),$$

with

$$\Sigma_{\mathrm{LoRA}} = \sigma^2 (I_q \otimes Q^{-1/2}) P_{\mathcal{T}} (I_q \otimes Q^{-1/2}).$$

This proves part (c).

Proof of part (d): first-order risk comparison. The asymptotic covariance of the unrestricted estimator from part (b) is

$$\Sigma_{\text{full}} = \sigma^2(I_q \otimes Q^{-1}).$$

The covariance from part (c) can be written as

$$\Sigma_{\text{LoRA}} = \sigma^2(I_q \otimes Q^{-1/2})P_{\mathcal{T}}(I_q \otimes Q^{-1/2}).$$

Since $P_{\mathcal{T}}$ is an orthogonal projector, $P_{\mathcal{T}} \preceq I_{pq}$. Therefore,

$$\Sigma_{\text{LoRA}} \preceq \sigma^2(I_q \otimes Q^{-1/2})I_{pq}(I_q \otimes Q^{-1/2}) = \sigma^2(I_q \otimes Q^{-1}) = \Sigma_{\text{full}}.$$

Taking traces gives

$$\operatorname{tr}(\Sigma_{\text{LoRA}}) \leq \operatorname{tr}(\Sigma_{\text{full}}) = \sigma^2 q \operatorname{tr}(Q^{-1}).$$

Moreover, because $r < \min(p, q)$, the tangent space $\mathcal{T}$ is a proper subspace of $\mathbb{R}^{p\times q}$. Hence $I_{pq} - P_{\mathcal{T}}$ is nonzero positive semidefinite. Since $I_q \otimes Q^{-1/2}$ is nonsingular,

$$\operatorname{tr}(\Sigma_{\text{full}}) - \operatorname{tr}(\Sigma_{\text{LoRA}}) = \sigma^2 \operatorname{tr}\Big[(I_q \otimes Q^{-1/2})(I_{pq} - P_{\mathcal{T}})(I_q \otimes Q^{-1/2})\Big] > 0.$$

Thus

$$\operatorname{tr}(\Sigma_{\text{LoRA}}) < \sigma^2 q \operatorname{tr}(Q^{-1}).$$

Finally, asymptotic normality and uniform integrability of the Gaussian least-squares limits imply the first-order risk expansions

$$\mathbb{E}\|\widehat{\Delta}_{\text{LoRA}} - \Delta^*\|_F^2 = \frac{1}{n}\operatorname{tr}(\Sigma_{\text{LoRA}}) + o(n^{-1}),$$

and

$$\mathbb{E}\|\widehat{\Delta}_{\text{full}} - \Delta^*\|_F^2 = \frac{1}{n}\operatorname{tr}(\Sigma_{\text{full}}) + o(n^{-1}).$$

Therefore the reduced-rank estimator has strictly smaller first-order asymptotic Frobenius risk than unrestricted least squares under correct rank specification. This proves part (d), and completes the proof of Theorem 3.2. □

F Proof of Theorem 3.3 (Oracle Inequality for Approximate Low Rank)

We condition on the fixed design matrix X. The centered model is

$$\widetilde{Y} = X\Delta^* + E, \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Let

$$\widehat{\Delta} = \widehat{\Delta}_{\text{LoRA}}$$

denote the rank-r estimator.

Let Δ_r be the best rank-r approximation of Δ^* in Frobenius norm:

$$\Delta_r = \arg\min_{\operatorname{rank}(\Delta)\leq r} \|\Delta - \Delta^*\|_F, \qquad \|\Delta_r - \Delta^*\|_F^2 = \sum_{j=r+1}^{\min(p,q)} d_j^2.$$

For brevity, define

$$\tau_r^2 := \|\Delta_r - \Delta^*\|_F^2 = \sum_{j>r} d_j^2.$$

Basic inequality. By optimality of $\widehat{\Delta}$ over the rank-r constraint set,

$$\|\widetilde{Y} - X\widehat{\Delta}\|_F^2 \leq \|\widetilde{Y} - X\Delta_r\|_F^2.$$

Substituting $\widetilde{Y} = X\Delta^* + E$, and writing

$$M := \widehat{\Delta} - \Delta_r, \qquad R := \Delta^* - \Delta_r,$$

we obtain

$$\|E + XR - XM\|_F^2 \leq \|E + XR\|_F^2.$$

Expanding and rearranging gives

$$\|XM\|_F^2 \leq 2\langle E, XM\rangle + 2\langle XR, XM\rangle. \tag{6}$$

Since both $\widehat{\Delta}$ and Δ_r have rank at most r,

$$\operatorname{rank}(M) \leq 2r.$$

Step 1: control of the approximation term. The approximation residual $R = \Delta^* - \Delta_r$ need not be low rank. This is why Assumption 3.1 includes an upper bound for all matrices. Using that upper bound,

$$\|XR\|_F^2 \le n\kappa_1 \|R\|_F^2 = n\kappa_1 \tau_r^2.$$

Hence

$$\|XR\|_F \le \sqrt{n\kappa_1}\, \tau_r.$$

Step 2: control of the stochastic term. We use the following standard Gaussian-width bound for low-rank matrices; see, for example, Bunea et al. [4], Vershynin [25], or Wainwright [26].

Lemma. There exists a universal constant $C_0 > 0$ such that, for every $\delta \in (0,1)$, with conditional probability at least $1-\delta$ given X,

$$\sup_{\operatorname{rank}(M)\le 2r,\ \|XM\|_F \le 1} \langle E, XM \rangle \le C_0 \sigma \sqrt{r(p+q) + \log(1/\delta)}.$$

Proof of the lemma. Since X has full column rank, define

$$Z = X(X^\top X)^{-1/2}, \qquad Z^\top Z = I_p.$$

For each M, define

$$G = (X^\top X)^{1/2} M.$$

Then

$$XM = ZG, \qquad \operatorname{rank}(G) = \operatorname{rank}(M), \qquad \|XM\|_F = \|G\|_F.$$

Therefore

$$\sup_{\operatorname{rank}(M)\le 2r,\ \|XM\|_F \le 1} \langle E, XM \rangle = \sup_{\operatorname{rank}(G)\le 2r,\ \|G\|_F \le 1} \langle Z^\top E, G \rangle.$$

Because Z has orthonormal columns and E has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, the matrix $Z^\top E \in \mathbb{R}^{p\times q}$ also has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries.

Let

$$\mathcal{G} = \{G \in \mathbb{R}^{p\times q} : \operatorname{rank}(G) \le 2r,\ \|G\|_F \le 1\}.$$

The Gaussian width of $\mathcal{G}$ is of order $\sqrt{r(p+q)}$. Thus,

$$\mathbb{E} \sup_{G\in\mathcal{G}} \langle Z^\top E, G \rangle \le C\sigma \sqrt{r(p+q)}$$

for a universal constant C. The supremum is a 1-Lipschitz function of $Z^\top E$ with respect to the Frobenius norm, so Gaussian concentration implies that, with probability at least $1-\delta$,

$$\sup_{G\in\mathcal{G}} \langle Z^\top E, G \rangle \le C\sigma \sqrt{r(p+q)} + C\sigma \sqrt{\log(1/\delta)}.$$

Absorbing constants gives

$$\sup_{\operatorname{rank}(M)\le 2r,\ \|XM\|_F \le 1} \langle E, XM \rangle \le C_0 \sigma \sqrt{r(p+q) + \log(1/\delta)}.$$

This proves the lemma. □

Step 3: bound $\|\widehat{\Delta} - \Delta_r\|_F^2$. On the event of the lemma,

$$\langle E, XM \rangle \le C_0 \sigma \sqrt{r(p+q) + \log(1/\delta)}\, \|XM\|_F.$$

Using the approximation bound from Step 1, equation (6) gives

$$\|XM\|_F^2 \le 2 \left(C_0 \sigma \sqrt{r(p+q) + \log(1/\delta)} + \sqrt{n\kappa_1}\, \tau_r \right) \|XM\|_F.$$

If $XM = 0$, the desired bound is immediate. Otherwise, divide both sides by $\|XM\|_F$ to obtain

$$\|XM\|_F \le 2C_0 \sigma \sqrt{r(p+q) + \log(1/\delta)} + 2\sqrt{n\kappa_1}\, \tau_r.$$

Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|XM\|_F^2 \le 8C_0^2\sigma^2\{r(p+q) + \log(1/\delta)\} + 8n\kappa_1\tau_r^2.$$

Because $\mathrm{rank}(M) \le 2r$, the lower restricted-isometry part of Assumption 3.1 gives

$$n\kappa_0\|M\|_F^2 \le \|XM\|_F^2.$$

Therefore

$$\|\widehat{\Delta} - \Delta_r\|_F^2 = \|M\|_F^2 \le \frac{8\kappa_1}{\kappa_0}\tau_r^2 + \frac{8C_0^2}{\kappa_0}\sigma^2\frac{r(p+q) + \log(1/\delta)}{n}.$$

This proves the first high-probability inequality in Theorem 3.3, with constants

$$c_1 = \frac{8\kappa_1}{\kappa_0}, \qquad c_2 = \frac{8C_0^2}{\kappa_0}.$$

Step 4: convert to error relative to Δ^*. Finally,

$$\widehat{\Delta} - \Delta^* = (\widehat{\Delta} - \Delta_r) + (\Delta_r - \Delta^*).$$

Hence

$$\|\widehat{\Delta} - \Delta^*\|_F^2 \le 2\|\widehat{\Delta} - \Delta_r\|_F^2 + 2\|\Delta_r - \Delta^*\|_F^2 = 2\|M\|_F^2 + 2\tau_r^2.$$

Substituting the bound for $\|M\|_F^2$ yields

$$\|\widehat{\Delta} - \Delta^*\|_F^2 \le \left(2 + \frac{16\kappa_1}{\kappa_0}\right)\tau_r^2 + \frac{16C_0^2}{\kappa_0}\sigma^2\frac{r(p+q) + \log(1/\delta)}{n}.$$

Since $\tau_r^2 = \sum_{j>r} d_j^2$, this proves the second high-probability inequality in Theorem 3.3, with

$$c_3 = 2 + \frac{16\kappa_1}{\kappa_0}, \qquad c_4 = \frac{16C_0^2}{\kappa_0}.$$

Expectation bound. Equivalently, setting $t = \log(1/\delta)$, the preceding high-probability bound says that, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\widehat{\Delta} - \Delta^*\|_F^2 \le c_3\tau_r^2 + c_4\sigma^2\frac{r(p+q) + t}{n}.$$

Integrating this tail inequality gives

$$\mathbb{E}\Big[\|\widehat{\Delta} - \Delta^*\|_F^2 \,\Big|\, X\Big] \le c_3\tau_r^2 + C\sigma^2\frac{r(p+q)}{n},$$

for a constant $C > 0$ depending only on κ_0, κ_1. This proves the expectation bound and completes the proof of Theorem 3.3. □