arXiv:2611.14397v1 [cs.LG] 10 Nov 2026

A Margin Perspective on LoRA: Robustness to Catastrophic Forgetting and Adapter Merging (MaLoRA)

Ziqing Xu[1,4*] **Hancheng Min**[7] **Lachlan Ewen MacDonald**[3,4]
Salma Tarmoun[2,4] **Enrique Mallada**[8] **Weijie Su**[1] **René Vidal**[2,3,4,5,6]
[1]Statistics and Data Science [2]AMCS [3]ESE [4]IDEAS [5]Radiology [6]CIS
University of Pennsylvania
[7]INS & SMS, Shanghai Jiao Tong University [8]ECE, Johns Hopkins University

Abstract

Low-rank adaptation (LoRA) is the de facto method for parameter-efficient fine-tuning of large neural networks, achieving strong task adaptation with minimal trainable parameters and low optimization cost. Beyond this efficiency, practitioners have observed two striking properties: *robustness to catastrophic forgetting* and the ability to *merge independently trained adapters* into a single model that performs competitively across multiple tasks. These properties are central to continual and multi-task learning, yet remain poorly understood. In this work, we provide a theoretical explanation through a unified *margin-based perspective*. We analyze LoRA in multiclass linear classification under a near-orthogonal task regime with ℓ_2-regularization, and characterize the optimal LoRA adapter across regularization regimes. Our analysis shows that, in an intermediate regime of regularization parameters, the optimal adapter aligns with the *max-margin solution* on the fine-tuning data. Building on this characterization, we derive two key consequences. First, we obtain closed-form expressions for the margins on pre-training and fine-tuning data, revealing a precise margin trade-off: the regularization parameter controls the balance between retention and adaptation. Second, we analyze adapter merging, proving that merged models achieve positive margin on each task and deriving optimal mixing coefficients that balance margins across tasks and maximize the margin over their union. These results lead to a simple, training-free merging rule, which we term *margin-based LoRA merging* (MaLoRA). Experiments on modern architectures and real datasets validate our theoretical predictions, showing that MaLoRA matches or outperforms several adapter-merging baselines across a range of vision and language classification tasks.

1 Introduction

Foundation models have transformed artificial intelligence across numerous domains. However, deploying these models in real-world applications typically requires fine-tuning on specialized datasets to meet accuracy, safety, and alignment requirements. Traditional full fine-tuning, which updates all parameters of a pre-trained model, incurs substantial computational cost, including large memory footprints, long training times, and significant storage demands. Low-Rank Adaptation (LoRA) [Hu et al., 2022] addresses this challenge by augmenting pre-trained weight matrices W with low-rank adapters B, A while keeping W frozen. Empirically, LoRA matches the performance of full fine-tuning while training only a small fraction of the parameters.

*Correspondence: ziqingxu@wharton.upenn.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

While LoRA was originally introduced for its computational efficiency, practitioners have since identified two striking empirical properties.

1. *Robustness to catastrophic forgetting.* In fine-tuning and continual learning, the degradation of performance on previously learned tasks after adapting to new ones (a.k.a. catastrophic forgetting) is a fundamental challenge [McCloskey and Cohen, 1989, Kirkpatrick et al., 2017, Ramasesh et al., 2021]. Surprisingly, recent studies show that LoRA exhibits strong robustness to catastrophic forgetting, even after substantial adaptation [Wistuba et al., 2023, Biderman et al., 2024, Qiao and Mahdavi, 2024]. This robustness makes LoRA particularly appealing for continual learning.
2. *Effective adapter merging.* In multi-task settings, maintaining separate fine-tuned models for each task is storage-intensive. Remarkably, LoRA adapters trained independently on different tasks can be merged (e.g., via simple weighted averaging) into a single adapter that performs well across all tasks [Huang et al., 2023, Yadav et al., 2023, Wang et al., 2024, Yu et al., 2024, Zhao et al., 2024]. This property enables flexible and efficient deployment.

Despite these empirical benefits, the theoretical mechanisms underlying LoRA's *robustness to catastrophic forgetting* and its *adapter merging* capabilities remain poorly understood. In this work, we provide a unified theoretical explanation for both phenomena through a margin-based perspective.

Paper contributions. We study LoRA in multiclass linear classification under a near-orthogonal-task regime, in which fine-tuning samples are nearly orthogonal to one another and to the pre-training samples. In this setting, we characterize the global minimizer of the LoRA objective and use this characterization to analyze forgetting and adapter merging. Compared with prior theoretical work on LoRA (see Appendix B), our work makes the following contributions:

- *Characterization of the optimal solution.* We characterize the global minimizer of LoRA across different regularization regimes (Theorem 3.1). In the high-penalty regime, the optimal adapter is zero. In the intermediate regime, the adapter aligns with the *max-margin* direction on the fine-tuning data. In the low-penalty regime, the adapter aligns with a non-max-margin classifier whose direction generally does not admit a simple closed form. Nevertheless, we show that this direction converges to an interpretable limit as the regularization parameter tends to zero.
- *Theory for reduced forgetting.* We define forgetting as the reduction in margin on the pre-training data after fine-tuning, and derive closed-form expressions for the normalized margins on the pre-training data, the fine-tuning data, and their union (Theorem 3.2). Our analysis shows that these margins are governed by the regularization parameter and satisfy a precise trade-off: decreasing the regularization parameter increases the fine-tuning margin but also increases the Frobenius-norm ratio of the adapter to the pre-trained weights, leading to greater forgetting. We further identify the regularization parameter that maximizes the margin on the union of pre-training and fine-tuning tasks, thereby optimally balancing retention and adaptation.
- *Theory for adapter merging.* Under the intermediate-regime assumption for each task, we derive closed-form expressions for the margin of the merged model on every task (Theorem 3.3). We show that the merged model attains strictly positive margin on each task, providing a theoretical explanation for the effectiveness of adapter merging. We also characterize, in closed form, the mixing coefficients that maximize the margin of the merged model on the union of all tasks, which naturally leads to *margin-based LoRA merging* (MaLoRA), a theoretically motivated adapter-merging method. This contrasts with many existing model-merging approaches, which rely primarily on heuristic prescriptions for choosing how to combine task-specific models.
- *Numerical validation.* Experiments on real datasets and modern architectures support the theory in three ways: *(i)* in the linear probing setting, the empirically optimal regularization parameter agrees with the value predicted by our analysis; *(ii)* in the same setting, the empirically optimal adapter mixing weights agree with our closed-form coefficients; and *(iii)* in LoRA adapter merging, MaLoRA achieves performance comparable to, and in several cases better than, strong heuristic adapter-merging baselines across vision and language classification tasks.

2 Preliminaries

In this section, we first introduce the multiclass linear classification problem and its corresponding hard-margin SVM formulation. We then describe the specific learning problem considered in this

work, which consists of a pre-training and fine-tuning phase. Finally, we state the key assumptions that underpin our analysis, and discuss their motivation and implications.

2.1 Background on multiclass linear classification

We first review the standard K-class linear classification problem. Given a dataset $\mathcal{D}$, the goal is to learn a weight matrix $W \in \mathbb{R}^{K\times d}$ that minimizes the mean empirical cross-entropy loss:

$$\min_{W\in\mathbb{R}^{K\times d}} L(W;\mathcal{D}) := \frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{y}, W\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input and $\boldsymbol{y} \in \mathbb{R}^K$ is a one-hot label vector. The cross-entropy loss is given by:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) := -\sum_{c=1}^{K} y_c \log\left(\frac{\exp(\hat{y}_c)}{\sum_{i=1}^{K}\exp(\hat{y}_i)}\right), \qquad \hat{\boldsymbol{y}} = W\boldsymbol{x}\,. \tag{2}$$

Recent studies have shown that when the data are linearly separable, gradient descent (GD) implicitly biases the solution of (1) toward the max-margin classifier [Soudry et al., 2018, Lyu and Li, 2019, Ravi et al., 2024]. Specifically, the iterates of GD converge in direction to the max-margin solution of the following hard-margin SVM problem[2]:

$$W^{\mathrm{SVM}}(\mathcal{D}) := \underset{W\in\mathbb{R}^{K\times d}}{\arg\min} \|W\|_F^2 \quad \text{subject to} \quad C(W;\mathcal{D}) \geq 1, \tag{3}$$

where $C(W;\mathcal{D})$ is the minimum margin over the dataset:

$$C(W;\mathcal{D}) := \min_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \min_{k\neq c, y_c=1} \left(\boldsymbol{w}_c^\top\boldsymbol{x} - \boldsymbol{w}_k^\top\boldsymbol{x}\right), \quad \text{with} \quad W = [\boldsymbol{w}_1;\ldots;\boldsymbol{w}_K]^\top. \tag{4}$$

The hard-margin SVM problem in (3) seeks a weight matrix of minimum norm that separates all examples with at least unit margin between the correct class and the nearest competing class.

Intuitively, classification decisions depend primarily on the direction of the weight matrix rather than its scale. To enable comparisons across different scales of W, we adopt the *normalized margin* used in [Crammer and Singer, 2001], i.e., $\gamma(W;\mathcal{D}) := \frac{C(W;\mathcal{D})}{\|W\|_F}$ for $W \neq 0$. This definition has been used in recent theoretical studies [Lyu and Li, 2019, Ravi et al., 2024]. We additionally set $\gamma(0;\mathcal{D}) := 0$ by convention.

2.2 Pre-training and fine-tuning setup

We now describe the specific pre-training and fine-tuning framework used in our theoretical analysis.

Pre-training stage. Let $\mathcal{D}_{\mathrm{pre}}$ be a labeled dataset with K classes, and let $W_{\mathrm{pre}} \in \mathbb{R}^{K\times d}$ denote a pre-trained linear classifier obtained by GD minimizing the cross-entropy loss $L(W;\mathcal{D}_{\mathrm{pre}})$. When $\mathcal{D}_{\mathrm{pre}}$ is linearly separable, this objective admits no finite minimizer. The GD iterates instead diverge in norm while converging in direction to the max-margin solution [Lyu and Li, 2019, Soudry et al., 2018]. We formalize this directional structure in Assumption 2.4.

Fine-tuning stage. We adapt the pre-trained model to a downstream dataset $\mathcal{D}_{\mathrm{ft}}$ consisting of n additional samples from each of $2 \leq \bar{K} \leq K$ classes selected from the original K pre-training classes. This restriction applies throughout the paper and appendix, including to every task in adapter merging. This setup captures practical scenarios such as *domain-incremental learning* [Esaki et al., 2024], *domain adaptation* [Zohrizadeh et al., 2019, Zhang et al., 2022], and *bias-rebalancing fine-tuning* [Li and Xu, 2021, ValizadehAslani et al., 2024].

When LoRA is applied to adapt the pre-trained model to the fine-tuning task, we introduce a low-rank update to the weights, parameterized by matrices $B \in \mathbb{R}^{K\times r}$ and $A \in \mathbb{R}^{r\times d}$:

$$\min_{B,A}\ L(W_{\mathrm{pre}} + BA;\mathcal{D}_{\mathrm{ft}}) + \tfrac{\lambda}{2}(\|B\|_F^2 + \|A\|_F^2). \tag{5}$$

Here, the Frobenius-norm penalties explicitly constrain the adapters in weight space, limiting the deviation of the fine-tuned model from the pretrained initialization.

[2]Whenever (3) is feasible, its solution is unique since the objective $\|W\|_F^2$ is strongly convex and the feasible set is convex.

Although some empirical studies of LoRA do not include this penalty explicitly in the loss, they train the adapters using the AdamW optimizer [Loshchilov and Hutter, 2017], whose decoupled weight decay induces an effect analogous to L_2 regularization on B and A. Thus, the objective in (5) captures the regularization that is, explicitly or implicitly, used in standard LoRA fine-tuning practice [Hu et al., 2022, Wistuba et al., 2023, Biderman et al., 2024].

Our goal is to understand the structure of the optimal solution to (5), and how it explains both LoRA's robustness to catastrophic forgetting and its effectiveness in enabling adapter merging.

2.3 Assumptions and their implications

Our theoretical analysis uses the following assumptions. Let $N_{\mathrm{ft}} = \bar{K}n$ be the number of fine-tuning samples, write $X_{\mathrm{ft}} = [\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_{N_{\mathrm{ft}}}]$, and let X_{pre} collect the pre-training feature vectors as columns.

Assumption 2.1 (Task geometry)**.** *The combined pre-training and fine-tuning datasets are linearly separable. There exist $\epsilon_1, \epsilon_2 \in [0, 1)$ such that, for all $i, j \in [N_{\mathrm{ft}}]$ and $\boldsymbol{x} \in \mathcal{D}_{\mathrm{pre}}$,*

$$(a)\ \|\bar{\boldsymbol{x}}_i\|_2 = \|\boldsymbol{x}\|_2 = 1, \quad (b)\ |\bar{\boldsymbol{x}}_i^\top \bar{\boldsymbol{x}}_j| \le \epsilon_1\ (i \neq j), \quad (c)\ |\bar{\boldsymbol{x}}_i^\top \boldsymbol{x}| \le \epsilon_2. \tag{6}$$

Assumption 2.2 (Dimension)**.** *In the strict-orthogonal case $\epsilon_1 = \epsilon_2 = 0$, the input dimension satisfies $d \ge N_{\mathrm{ft}} + \mathrm{rank}(X_{\mathrm{pre}})$. For near-orthogonal data, we assume $(N_{\mathrm{ft}} - 1)\epsilon_1 < 1$.*

Assumption 2.3 (LoRA rank)**.** *The LoRA rank parameter satisfies $r \ge \bar{K}$, where $\bar{K}$ is the number of fine-tuning classes.*

Assumption 2.4 (Pre-trained classifier)**.** *The pre-trained classifier W_{pre} is a scaled version of the hard-margin SVM solution on $\mathcal{D}_{\mathrm{pre}}$: $W_{\mathrm{pre}} = \rho_{\mathrm{pre}} \cdot W_{\mathrm{pre}}^{\mathrm{SVM}} / \|W_{\mathrm{pre}}^{\mathrm{SVM}}\|_F$, for some scalar $\rho_{\mathrm{pre}} > 0$, where $W_{\mathrm{pre}}^{\mathrm{SVM}} = W^{\mathrm{SVM}}(\mathcal{D}_{\mathrm{pre}})$.*

Assumption justification. The near-orthogonal task Assumption 2.1 is arguably a strong assumption in practice, but reflects the realistic geometric structure of fine-tuning data, and similar orthogonality-type assumptions are commonly used in theoretical analyses to enable closed-form characterizations [Bui Thi Mai and Lampert, 2021, Boursier et al., 2022, Frei et al., 2022, Kou et al., 2023]. The feature-correlation analyses on CIFAR-100 and six multi-domain benchmarks in Appendices H.1.1 and H.2.1 provide empirical motivation for studying this reference geometry. Assumption 2.2 makes the geometric feasibility and conditioning requirements explicit. The strict-case bound accounts for both the fine-tuning sample count and the pre-training feature span, specifying a high-dimensional, sample-limited setting. *Assumption 2.3* requires only $r \ge \bar{K}$, meaning the LoRA rank needs only to be at least the number of fine-tuning classes. Finally, *Assumption 2.4* is motivated by recent theoretical results on the implicit bias of GD in multiclass classification [Lyu et al., 2021, Ravi et al., 2024], as discussed in §2.1. ρ_{pre} captures the (unspecified) scale at which pre-training was stopped, which arises naturally in implicit-bias analyses where the GD iterate diverges along the SVM direction.

For clarity of exposition, all theorems in the main paper are stated under the strict-orthogonal special case ($\epsilon_1 = \epsilon_2 = 0$) of Assumption 2.1. The general versions of these results for $\epsilon_1, \epsilon_2 > 0$, together with their proofs, are deferred to the Appendix D–F.

Implications. The assumptions introduced above lead to a clean closed-form expression for the $\bar{K}$-class hard-margin SVM solution for the fine-tuning task, and also imply *near-orthogonality* between the pre-trained weights W_{pre} and the fine-tuning dataset $\mathcal{D}_{\mathrm{ft}}$, which underpins our subsequent theoretical analysis in §3. To formalize these implications, we begin by introducing the concept of a *simplex equiangular tight frame* (simplex ETF), which characterizes the geometry of the hard-margin SVM solution in our setting

Definition 2.1 (Simplex ETF)**.** *For $m \ge 2$, an m-simplex ETF is a collection of m vectors in $\mathbb{R}^p$ for some $p \ge m$, given by the columns of $M_m := \sqrt{\frac{m}{m-1}}\, P \left(\boldsymbol{I}_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m^\top\right)$, where $P \in \mathbb{R}^{p \times m}$ satisfies $P^\top P = \boldsymbol{I}_m$.*

Throughout, we use the canonical choice $p = m$ and $P = \boldsymbol{I}_m$, so $M_m \in \mathbb{R}^{m \times m}$. Using this definition, we now present a proposition that, under the strict-orthogonal specialization, characterizes the structure of the hard-margin SVM solution for the fine-tuning data ($\mathcal{D}_{\mathrm{ft}}$) and the orthogonality between the pre-trained weights (W_{pre}) and the fine-tuning data.

Proposition 2.1. *Let $\bar{K} \geq 2$. Under Assumption 2.1 with $\epsilon_1 = \epsilon_2 = 0$ and Assumption 2.4, the following hold:*

(i) The minimum-norm SVM solution $W_{\text{ft}}^{\text{SVM}} \in \mathbb{R}^{\bar{K} \times d}$ for $\mathcal{D}_{\text{ft}}$, defined by (3) with $\bar{K}$ output classes, admits the closed-form solution $W_{\text{ft}}^{\text{SVM}} = \sqrt{\frac{\bar{K}-1}{\bar{K}}} \left(M_{\bar{K}} \otimes \mathbf{1}_n^\top\right) X_{\text{ft}}^\top$, where $\boldsymbol{x}_{i,j}$ is the j-th sample from class i, and $X_{\text{ft}} = [\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,n}, \ldots, \boldsymbol{x}_{\bar{K},n}]$ is the fine-tuning data matrix.

(ii) The pre-trained weights are orthogonal to the fine-tuning data: $W_{\text{pre}} \bar{\boldsymbol{x}} = 0$ for all $\bar{\boldsymbol{x}} \in \mathcal{D}_{\text{ft}}$.

Remark on Proposition 2.1. Proposition 2.1 reveals that the hard-margin SVM solution for the fine-tuning task admits a compact closed-form expression: $W_{\text{ft}}^{\text{SVM}} = \sqrt{\frac{\bar{K}-1}{\bar{K}}} \left(M_{\bar{K}} \otimes \mathbf{1}_n^\top\right) X_{\text{ft}}^\top$. However, this form does not immediately reveal how the classifier aggregates information from the fine-tuning data. To provide insight, we explicitly derive the expression for the *first row* of $W_{\text{ft}}^{\text{SVM}}$ (the remaining rows follow symmetrically), yielding $\left(\frac{\bar{K}-1}{\bar{K}} \sum_{i=1}^n \bar{\boldsymbol{x}}_{1,i} - \frac{1}{\bar{K}} \sum_{k=2}^{\bar{K}} \sum_{j=1}^n \bar{\boldsymbol{x}}_{k,j}\right)^\top$.

Intuitively, each row of $W_{\text{ft}}^{\text{SVM}}$ encodes a direction that emphasizes its corresponding class mean while uniformly suppressing the influence of all other classes, leading to class separation in max-margin classification. The second result in Proposition 2.1 follows from the representer theorem for SVMs, which implies that the hard-margin SVM solution on pre-training data lies in the span of $\mathcal{D}_{\text{pre}}$. By Assumption 2.4, W_{pre} inherits this property; combined with condition (c) of Assumption 2.1 at $\epsilon_2 = 0$, each row of W_{pre} is therefore orthogonal to the fine-tuning data.

Building upon these assumptions and propositions, we develop a rigorous theoretical framework that provides a principled explanation for LoRA's empirical advantages in the subsequent section. Our analysis reveals properties that enable precise quantification of LoRA's benefits in terms of robustness to catastrophic forgetting and adapter merging performance.

3 Main Results

In this section, we first analyze the optimization landscape of the LoRA objective and characterize its global minimum in §3.1. We then derive closed-form expressions for the margins on the union of pre-training and fine-tuning datasets, explaining why LoRA mitigates *catastrophic forgetting* in §3.2. Finally, in §3.3, we extend our margin-based analysis to the multi-task setting, explaining the effectiveness of *adapter merging* and deriving optimal mixing coefficients that maximize the margin on the union of all tasks. Moreover, building on these closed-form coefficients, we translate them into a simple training-free merging rule, *margin-based LoRA merging* (MaLoRA, Algorithm 1).

3.1 Global minimum of the LoRA objective

In this subsection, we present our main theoretical result characterizing the global minimizer of the LoRA objective in (5). Interestingly, under suitable regularization, we find that part of the optimal LoRA adapter aligns exactly with the $\bar{K}$-class hard-margin SVM solution ($W_{\text{ft}}^{\text{SVM}}$) for the fine-tuning data (see Proposition 2.1 for its definition). While our analysis neither assumes nor relies on this alignment, its emergence reveals a geometric structure that partially explains LoRA's effectiveness.

Theorem 3.1. *Let K and $\bar{K} \leq K$ be the number of pre-training and fine-tuning classes, respectively. There exist scalar functions $a_\lambda,\ b_\lambda,\ c_\lambda, \Theta_\lambda$ of λ such that for any global minimizer $\left(B_\lambda^*, A_\lambda^*\right)$ of (5) with $\lambda > 0$, we have*

$$B_\lambda^* A_\lambda^* = \left(\begin{pmatrix} (a_\lambda + b_\lambda)\, \mathbf{I}_{\bar{K}} \ - \ b_\lambda\, \mathbf{1}_{\bar{K}} \mathbf{1}_{\bar{K}}^\top \\ -c_\lambda\, \mathbf{1}_{K-\bar{K}} \mathbf{1}_{\bar{K}}^\top \end{pmatrix} \otimes \mathbf{1}_n^\top \right) X_{\text{ft}}^\top \,, \quad \|B_\lambda^* A_\lambda^*\|_F = \Theta_\lambda \,. \tag{7}$$

When $\bar{K} \geq 2$, there exists a critical regularization weight $\lambda_{\text{crit}} \in [0, \frac{1}{\bar{K}\sqrt{n}})$ such that the scalar functions $a_\lambda,\ b_\lambda,\ c_\lambda, \Theta_\lambda$ are characterized as follows:

(i) High-penalty regime $\left(\lambda \geq \frac{1}{\bar{K}\sqrt{n}}\right)$: $a_\lambda = b_\lambda = c_\lambda = \Theta_\lambda = 0$, thus $B_\lambda^* A_\lambda^* = \mathbf{0}_{K \times d}$.

(ii) Intermediate regime $\left(\lambda_{\text{crit}} < \lambda < \frac{1}{K\sqrt{n}}\right)$*:* $a_\lambda = \frac{\Theta_\lambda\sqrt{\bar{K}-1}}{K\sqrt{n}}$, $b_\lambda = \frac{\Theta_\lambda}{\bar{K}\sqrt{n(\bar{K}-1)}}$, $c_\lambda = 0$, *and* Θ_λ *is the unique root of a nonlinear equation (see* (29) *in Appendix D), thus the minimizer is*

$$B_\lambda^* A_\lambda^* \;=\; \frac{\Theta_\lambda}{\sqrt{\bar{K}n}} \left(\begin{pmatrix} M_{\bar{K}} \\ \mathbf{0}_{(K-\bar{K})\times\bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top \right) X_{\text{ft}}^\top . \tag{8}$$

(iii) Low-penalty regime $\left(0 < \lambda \le \lambda_{\text{crit}}\right)$*:* $a_\lambda, b_\lambda, c_\lambda, \Theta_\lambda$ *are non-negative.*

Moreover, the small-regularization limit satisfies

$$\lim_{\lambda\to 0^+} \Theta_\lambda = \infty\,, \qquad \lim_{\lambda\to 0^+} \frac{B_\lambda^* A_\lambda^*}{\|B_\lambda^* A_\lambda^*\|_F} = \sqrt{\frac{1}{n\bar{K}}} \left(M_K^{(\bar{K})} \otimes \mathbf{1}_n^\top \right) X_{\text{ft}}^\top\,, \tag{9}$$

where $M_K^{(\bar{K})}$ *is the first* $\bar{K}$ *columns of the K-ETF matrix* M_K.

Remark 3.1. *For* $\bar{K} \ge 2$*, the critical weight satisfies* $\lambda_{\text{crit}} = 0$ *if and only if* $\bar{K} = K$*. In this case, the intermediate regime covers all* $0 < \lambda < 1/(K\sqrt{n})$*, and the low-penalty regime contains no positive regularization weights. The limits in* (9) *remain valid as* $\lambda \to 0^+$*, approached through the intermediate regime.*

The proof of Theorem 3.1 is provided in Appendix D. The theorem shows that the optimal LoRA update $B_\lambda^* A_\lambda^*$ has the unified form in (7), and identifies up to three regimes based on the regularization parameter λ. We make the following remarks:

Effect of regularization and connection with max-margin classifier. The theorem shows that the regularization parameter governs the structure of the optimal LoRA adapters. In the *high-penalty regime*, the regularization term dominates and drives the optimal adapters to zero. In the *intermediate regime*, the adapters balance cross-entropy and regularization, yielding the explicit form in (8). In particular, the first $\bar{K}$ rows of $B_\lambda^* A_\lambda^*$ align with the $\bar{K}$-class hard-margin SVM solution on the fine-tuning data (Proposition 2.1), which gives a clear max-margin interpretation of the learned solution. In the *low-penalty regime*, explicit formulas for a_λ and b_λ are no longer available. Nevertheless, as $\lambda \to 0^+$, the cross-entropy term dominates and $\Theta_\lambda \to \infty$, while our asymptotic analysis still characterizes the limiting direction of $B_\lambda^* A_\lambda^*$ in (9). This direction coincides with the hard-margin SVM solution only when $\bar{K} = K$.

Implication for the fine-tuned model. The structured form of the optimal LoRA adapters in (7) naturally enables $W_{\text{LoRA}}^\lambda := W_{\text{pre}} + B_\lambda^* A_\lambda^*$ to perform well on both pre-training and fine-tuning data. Under Assumptions 2.1 and 2.4, Proposition 2.1 and Theorem 3.1 imply that $W_{\text{pre}}\bar{\boldsymbol{x}} = 0$ for all $\bar{\boldsymbol{x}} \in \mathcal{D}_{\text{ft}}$ and $B_\lambda^* A_\lambda^* \boldsymbol{x} = 0$ for all $\boldsymbol{x} \in \mathcal{D}_{\text{pre}}$. That is, the two components operate independently: W_{pre} classifies the pre-training data, while $B_\lambda^* A_\lambda^*$ adapts to the fine-tuning data:

$$\forall \boldsymbol{x} \in \mathcal{D}_{\text{pre}} \cup \mathcal{D}_{\text{ft}}\,, \quad W_{\text{LoRA}}^\lambda \boldsymbol{x} = \begin{cases} W_{\text{pre}}\boldsymbol{x} & \text{if } \boldsymbol{x} \in \mathcal{D}_{\text{pre}}, \\ B_\lambda^* A_\lambda^* \boldsymbol{x} & \text{if } \boldsymbol{x} \in \mathcal{D}_{\text{ft}}. \end{cases} \quad \text{holds} \quad \forall \lambda > 0\,. \tag{10}$$

This clean separation lets us decompose the normalized margin of W_{LoRA}^λ on each component dataset into a contribution from the relevant component (either W_{pre} or $B_\lambda^* A_\lambda^*$) and a relative-norm factor:

$$\gamma(W_{\text{LoRA}}^\lambda; \mathcal{D}) = \gamma(W; \mathcal{D}) \frac{\|W\|_F}{\|W_{\text{LoRA}}^\lambda\|_F}, \qquad (W, \mathcal{D}) \in \{(W_{\text{pre}}, \mathcal{D}_{\text{pre}}), (B_\lambda^* A_\lambda^*, \mathcal{D}_{\text{ft}})\}\,. \tag{11}$$

The expression in (11) shows that the normalized margin on the combined dataset depends not only on the individual normalized margins of W_{pre} and $B_\lambda^* A_\lambda^*$ on the $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{ft}}$ respectively, but also on the relative Frobenius norms of these components. In the following sections, we use (11) to derive closed-form expressions for the normalized margin on the combined dataset and to compute the optimal mixing coefficients for *adapter merging*. For clarity of presentation, we focus on the case $\bar{K} = K$ and refer the reader to Appendix E and Appendix F for the full version of our results.

3.2 Margin trade-off between forgetting and fine-tuning under LoRA adaptation

We now analyze catastrophic forgetting through the normalized margin. Our goal is to characterize how the LoRA fine-tuned model trades off performance on the pre-training data and the fine-tuning

data as the regularization parameter varies, and to identify the regularization level that maximizes the margin on their union. For convenience, define $\gamma_{\mathrm{pre}} := \gamma(W_{\mathrm{pre}}; \mathcal{D}_{\mathrm{pre}})$, $\gamma_{\mathrm{ft},\lambda} := \gamma(B_\lambda^* A_\lambda^*; \mathcal{D}_{\mathrm{ft}})$. Moreover, we define the adaptation to the fine-tuning data and forgetting on the pre-training data by

$$\mathrm{Adapt}(\lambda) = \gamma\big(W_{\mathrm{LoRA}}^\lambda; \mathcal{D}_{\mathrm{ft}}\big) - \gamma(W_{\mathrm{pre}}, \mathcal{D}_{\mathrm{ft}}), \quad \mathrm{Forget}(\lambda) := \gamma_{\mathrm{pre}} - \gamma\big(W_{\mathrm{LoRA}}^\lambda; \mathcal{D}_{\mathrm{pre}}\big)\,. \tag{12}$$

Based on these definitions, we now present the theorem.

Theorem 3.2. *Adopt the setup of Theorem 3.1 under strict orthogonality, and assume $\bar{K} = K \geq 2$. Let γ_{ft}^* be the max normalized margin any linear classifier can obtain on the fine-tuning data $\mathcal{D}_{\mathrm{ft}}$, and recall the scalar ρ_{pre} from Assumption 2.4. Then, $\mathrm{Adapt}(\lambda)$ and $\mathrm{Forget}(\lambda)$ can be characterized uniformly over all λ as follows:*

$$\mathrm{Adapt}(\lambda) = \gamma_{\mathrm{ft},\lambda} \frac{\Theta_\lambda}{\sqrt{\Theta_\lambda^2 + \rho_{\mathrm{pre}}^2}}\,, \quad \mathrm{Forget}(\lambda) = \gamma_{\mathrm{pre}} \left(1 - \frac{\rho_{\mathrm{pre}}}{\sqrt{\Theta_\lambda^2 + \rho_{\mathrm{pre}}^2}}\right). \tag{13}$$

Moreover, the effect of the regularization parameter λ is characterized by two regimes:

(i) High-penalty regime: if $\lambda \geq \frac{1}{K\sqrt{n}}$, then $\Theta_\lambda = 0$, and hence $\mathrm{Forget}(\lambda) = \mathrm{Adapt}(\lambda) = 0$.

(ii) Intermediate and low-penalty regimes: if $\lambda < \frac{1}{K\sqrt{n}}$, then $\gamma_{\mathrm{ft},\lambda} = \gamma_{\mathrm{ft}}^*$, and Θ_λ is decreasing in λ. Therefore, $\mathrm{Adapt}(\lambda)$ and $\mathrm{Forget}(\lambda)$ are both decreasing in λ. In particular,

$$\lim_{\lambda \to 0^+} \big(\mathrm{Adapt}(\lambda), \mathrm{Forget}(\lambda)\big) = \big(\gamma_{\mathrm{ft}}^*, \gamma_{\mathrm{pre}}\big), \quad \lim_{\lambda \to \frac{1}{K\sqrt{n}}} \big(\mathrm{Adapt}(\lambda), \mathrm{Forget}(\lambda)\big) = (0, 0). \tag{14}$$

Remark 3.2. *Notably, the intermediate and low-penalty regimes coincide when $\bar{K} = K$. This merging is specific to the case $\bar{K} = K$; when $\bar{K} < K$, the two regimes remain distinct.*

Balancing adaptation and forgetting. Theorem 3.2 shows that the same quantity Θ_λ controls both adaptation and forgetting. As λ decreases, the adapter scale Θ_λ increases, improving $\mathrm{Adapt}(\lambda)$ on the fine-tuning data while also increasing $\mathrm{Forget}(\lambda)$ on the pre-training data. Thus, LoRA fine-tuning induces a direct margin trade-off: stronger adaptation comes at the cost of larger pre-training margin degradation. The natural question is how to choose λ to balance these effects. In the following Corollary 3.1, we quantify this trade-off by studying the normalized margin on the union of the pre-training and fine-tuning datasets, which leads to an explicit optimal regularization level.

Corollary 3.1 (Optimal balance between adaptation and forgetting)**.** *Under the conditions of Theorem 3.2, with $\bar{K} = K$, the normalized margin of $W_{\mathrm{LoRA}}^\lambda$ on the union $\mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}$ is*

$$\gamma(W_{\mathrm{LoRA}}^\lambda; \mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}) = \min\big\{\gamma_{\mathrm{pre}} - \mathrm{Forget}(\lambda), \mathrm{Adapt}(\lambda)\big\}\,. \tag{15}$$

This margin is uniquely maximized at the balancing point $\lambda^* \in (0, 1/(K\sqrt{n}))$, with

$$\max_{\lambda > 0} \gamma(W_{\mathrm{LoRA}}^\lambda; \mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}) = \left(\frac{1}{(\gamma_{\mathrm{ft}}^*)^2} + \frac{1}{\gamma_{\mathrm{pre}}^2}\right)^{-1/2}. \tag{16}$$

3.3 Margin characterization of merged LoRA models and optimal mixing coefficients

We now extend the margin-based analysis to the multi-task setting, where a single pre-trained model is independently adapted by T LoRA adapters that we wish to merge into a unified model. Through the same margin lens, we derive closed-form expressions for the merged model's margin on each task and identify mixing coefficients that maximize the margin on the union of all tasks (Theorem 3.3). Building on this characterization, we design a practical LoRA merging algorithm *margin-based LoRA merging* (MaLoRA, Algorithm 1).

Let $\big(B_i^*, A_i^*\big)$ denote the LoRA adapter trained on dataset $\mathcal{D}_i$, for $i \in [T]$. We merge the adapters with the pre-trained weight matrix as follows:

$$W_{\mathrm{LoRA}}(\boldsymbol{\alpha}) \;:=\; W_{\mathrm{pre}} \;+\; \textstyle\sum_{i=1}^T \alpha_i\, B_i^* A_i^*, \qquad \alpha_1, \ldots, \alpha_T \geq 0, \tag{17}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$ are mixing coefficients. Moreover, we generalize Assumption 2.1 to the setting of learning multiple tasks as follows.

Algorithm 1 Margin-based LoRA Merging (MaLoRA)

Require: $W_{\text{pre}}, \{(B_i^*, A_i^*)\}_{i=1}^T, \mathcal{D}_{\text{pre}}^{\text{val}}, \{\mathcal{D}_i^{\text{val}}\}_{i=1}^T, S \subset \mathbb{R}_{>0}$

$\bar{C}_{\text{pre}} \leftarrow \bar{C}(W_{\text{pre}}; \mathcal{D}_{\text{pre}}^{\text{val}})$

for $i = 1, \ldots, T$ **do**

$\quad \bar{C}_i \leftarrow \bar{C}(W_{\text{pre}} + B_i^* A_i^*; \mathcal{D}_i^{\text{val}}), \quad \beta_i \leftarrow \begin{cases} \bar{C}_{\text{pre}}/\bar{C}_i, & \bar{C}_i > 0, \\ 0, & \text{otherwise} \end{cases}$

end for

$s^* \in \arg\max_{s \in S} \text{ValAcc}\Big(W_{\text{pre}} + s \sum_{i=1}^T \beta_i B_i^* A_i^*\Big)$

return $W_{\text{pre}} + s^* \sum_{i=1}^T \beta_i B_i^* A_i^*$

Assumption 3.1. *There exist* $\epsilon_1, \epsilon_2 \in [0, 1)$ *such that*

(a) Every sample has unit norm: $\|\boldsymbol{x}\|_2 = 1$ *for every* $\boldsymbol{x} \in \bigcup_{i=1}^T \mathcal{D}_i \cup \mathcal{D}_{\text{pre}}$.

(b) Within-task: *for each* $i \in [T]$ *and each pair of distinct* $\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}' \in \mathcal{D}_i$, $|\bar{\boldsymbol{x}}^\top \bar{\boldsymbol{x}}'| \le \epsilon_1$.

(c) Cross-task and pre-training: $\forall \bar{\boldsymbol{x}} \in \mathcal{D}_i$, $\bar{\boldsymbol{x}}' \in \mathcal{D}_j, i \ne j$, *or* $x \in \mathcal{D}_{\text{pre}}$, *then* $|\bar{\boldsymbol{x}}^\top \bar{\boldsymbol{x}}'|$, $|\bar{\boldsymbol{x}}^\top \boldsymbol{x}| \le \epsilon_2$.

The following theorem characterizes the optimal mixing coefficients in the strict orthogonal case ($\epsilon_1 = \epsilon_2 = 0$). We refer the reader to Appendix F for the proof.

Theorem 3.3. *Under Assumption 3.1, suppose each fine–tuning task* $\mathcal{D}_i$ $(i = 1, \ldots, T)$ *is trained with a regularization parameter in the intermediate or low-penalty regime. Let* $\Theta_{\lambda,i} = \|B_i^* A_i^*\|_F$ *and* $\gamma_i = \gamma(B_i^* A_i^*; \mathcal{D}_i)$. *For any nonnegative coefficient vector* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T) \in \mathbb{R}_{\ge 0}^T$, *the merged model* (17) *achieves the normalized margins on each task as follows*

$$\gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\text{pre}}\big) = \frac{\gamma_{\text{pre}} \rho_{\text{pre}}}{\sqrt{\rho_{\text{pre}}^2 + \sum_{j=1}^T \alpha_j^2 \Theta_{\lambda,j}^2}}, \quad \gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_i\big) = \frac{\gamma_i \alpha_i \Theta_{\lambda,i}}{\sqrt{\rho_{\text{pre}}^2 + \sum_{j=1}^T \alpha_j^2 \Theta_{\lambda,j}^2}},$$
$$\gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\text{pre}} \cup \{D_i\}_{i=1}^T\big) = \min\big\{\gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\text{pre}}\big), \gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_i\big), i \in [T]\big\}. \tag{18}$$

Choosing the weights $\alpha_i = \frac{C(W_{\text{pre}}; \mathcal{D}_{\text{pre}})}{C(B_i A_i; \mathcal{D}_i)}$, $i \in [T]$, *maximizes the margin on the union of all tasks:*

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}_{\ge 0}^T} \gamma\big(W_{\text{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\text{pre}} \cup \{D_i\}_{i=1}^T\big) = \Big(\frac{1}{\gamma_{\text{pre}}^2} + \sum_{j=1}^T \frac{1}{(\gamma_j)^2}\Big)^{-1/2}.$$

Theorem 3.3 characterizes the margins on each task as well as on the union of tasks as functions of the mixing coefficients. In particular, it shows that the optimal mixing coefficients are given by the ratio between the unnormalized margin of the pre-trained model on $\mathcal{D}_{\text{pre}}$ and that of each fine-tuned adapter on its own task. The unnormalized margin measures how confidently a classifier separates classes on its dataset, with smaller values indicating more challenging classification problems. This formula admits a direct interpretation: an adapter that achieves a smaller margin on its task is assigned a larger mixing weight, increasing its contribution so as to preserve a positive margin for that task in the merged model.

Towards practical margin-based adapter mixing The preceding analysis suggests choosing adapter mixing coefficients according to how confidently each adapter separates its task data, as measured by the unnormalized margin. In practice, however, datasets are often not perfectly separable, so the minimum margin $C(\cdot;\cdot)$ can be negative and overly sensitive to hard examples. We therefore use the average margin $\bar{C}(\cdot; \mathcal{D})$ as a more stable proxy for adapter performance. Under the strict-orthogonal assumptions of Theorem 3.3, the per-sample margins of each optimal adapter are equal on its own fine-tuning dataset. Consequently, minimum and average margins coincide on these fine-tuning datasets. In our practical algorithm, we therefore use average margins as empirical proxies for the minimum margins appearing in the theoretical mixing rule.

Algorithm 1: Margin-based LoRA merging (MaLoRA). Motivated by Theorem 3.3, we propose a practical LoRA merging algorithm using the average-margin proxies $\bar{C}(W; \mathcal{D})$. When a task has non-positive average margin $\bar{C}_i$, the corresponding coefficient would be negative and no longer

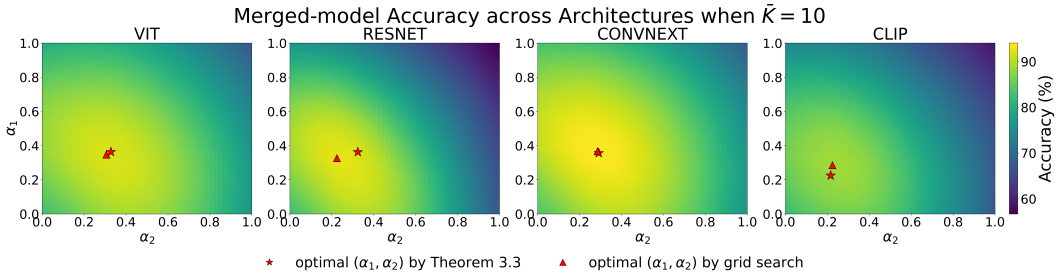


Figure 1: Merged-model accuracy across architectures. Each panel shows the accuracy of the merged model evaluated on the combined dataset, across a 50×50 grid of mixing coefficients $(\alpha_1, \alpha_2) \in (0, 1)^2$. For each architecture, the red star indicates the theoretically predicted optimal coefficients from Theorem 3.3, while the red triangle marks the empirically optimal coefficients.

consistent with the regime motivated by the theorem. Algorithm 1 therefore drops such tasks by setting their coefficients to zero. We also tune a global scaling factor s^* on a held-out validation set, following Ilharco et al. [2022], and find that this further improves performance. In §4.2, we compare the resulting method with other popular adapter-merging baselines.

4 Experiments

Our experiments aim to validate the theoretical predictions in Section 3 and assess how well they transfer beyond the linear setting, rather than to establish state-of-the-art performance. Since our analysis gives closed-form predictions for optimal regularization and adapter mixing coefficients, we test whether these quantities match empirical behavior in modern architectures. Specifically, we focus on two questions: *(i)* whether the mixing coefficients derived from our margin analysis agree with empirically optimal coefficients, and *(ii)* whether the resulting theory-driven merging rule, MaLoRA, remains competitive with adapter-merging baselines in nonlinear models. Experiments on the effect of regularization are deferred to Appendix H.1.2.

Setup and metric. We evaluate our theory and method in two complementary settings. First, in linear probing on CIFAR-100 [Krizhevsky et al., 2009] (§4.1), we use four frozen pre-trained backbones (`ResNet-50`, `ViT-B/16`, `ConvNeXt`, and `CLIP`) as feature extractors [He et al., 2016, Dosovitskiy et al., 2021, Radford et al., 2021, Liu et al., 2022], apply LoRA only to the final linear classifier [Kornblith et al., 2019], and construct tasks from the CIFAR-100 superclass hierarchy. Second, in multi-domain adapter merging (§4.2), we evaluate ViT-B/16 on PACS, VLCS, and TerraIncognita [Fang et al., 2013, Li et al., 2017, Beery et al., 2018], and XLM-RoBERTa Base on MNLI, Amazon Reviews, and CivilComments [Zhang et al., 2015, Williams et al., 2018, Borkan et al., 2019, Conneau et al., 2020]. Although our theory is stated in terms of normalized margin, we report classification accuracy in experiments, since real-world classification tasks often do not achieve perfect separation and can have negative normalized margins. Additional details are given in Appendix H.

4.1 Optimal mixing coefficients for adapter merging: theory vs. grid search

We fine-tune LoRA adapters on two disjoint tasks $(\mathcal{D}_1, \mathcal{D}_2)$, then merge them with the pre-trained weights using mixing coefficients (α_1, α_2). We evaluate the merged model's performance on the combined dataset $\mathcal{D}_{\text{pre}} \cup \mathcal{D}_1 \cup \mathcal{D}_2$. As shown in Figure 1, our predicted mixing coefficients closely match the grid-searched optima, confirming the practical accuracy of our merging theory.

4.2 Comparison between MaLoRA and other training-free baselines

Following Ilharco et al. [2022], we report *average normalized accuracy* (per-task merged accuracy divided by per-task finetuned accuracy, averaged across tasks) and its complement, *relative accuracy drop*, in Table 1. The aggregate performance of the methods is comparable, with average normalized

accuracy ranging from 92.87% to 93.32%. MaLoRA lies at the top of this range, achieving an average normalized accuracy of 93.32% and the smallest relative accuracy drop of 6.68%. We emphasize that although MaLoRA is competitive with baselines, our purpose here is not to achieve state-of-the-art results but to demonstrate the efficacy of our theoretically-grounded method.

Table 1: Average normalized accuracy and relative accuracy drop, following Ilharco et al. [2022]. Results are averaged over 10 seeds. **Bold** indicates the best result in each column.

	Per-benchmark normalized accuracy						Aggregate metrics	
Method	PACS	VLCS	Terra Inc.	MNLI	Amazon	CivilComm.	Avg. norm. acc	Rel. drop
Soup	96.38	96.89	64.31	101.16	100.64	**98.27**	92.94	7.06
TA	97.52	97.19	64.46	**101.27**	100.68	97.80	93.15	6.85
TIES	97.52	96.95	63.78	100.87	100.64	98.00	92.96	7.04
DARE-TIES	**97.57**	97.23	64.23	100.90	100.45	97.92	93.05	6.95
DARE-Linear	96.48	96.77	64.73	101.12	100.50	97.61	92.87	7.13
MaLoRA	96.69	**97.54**	**65.75**	101.23	**100.69**	98.00	**93.32**	**6.68**

Computational overhead. MaLoRA computes task margins once and reuses them throughout the scale search. Under a separate matched 49-configuration timing protocol, the mean runtime is 972.6 s versus 950.3 s for TA, approximately 2.35% overhead across the six benchmarks. Detailed timings and the measurement protocol are in Appendix H.2.2.

5 Conclusion, Limitations, and Future Work

In this work, we study LoRA fine-tuning through the normalized margin under a near-orthogonal task regime. We characterize the global minimizer of the regularized LoRA objective across three regularization regimes (Theorem 3.1), derive closed-form expressions for the margins on the pre-training and fine-tuning data, and identify the regularization level that maximizes the margin on their union (Theorem 3.2). We further extend the analysis to multi-task settings by deriving closed-form optimal mixing coefficients for adapter merging (Theorem 3.3). Based on this theory, we propose *margin-based LoRA merging* (MaLoRA, Algorithm 1), which matches or outperforms several adapter-merging baselines on a range of vision and language benchmarks. Across architectures and tasks, the empirical results closely align with the theoretical predictions.

Limitations and future work. First, our analysis relies on simplifying assumptions, most notably near-orthogonal task structure (Assumption 2.1) and balanced class sizes in the fine-tuning data. These assumptions enable a sharp closed-form characterization, but they also limit the scope of the theory. An important direction for future work is to relax them and study LoRA under more general data distributions and pre-training conditions. Second, experiments applying *MaLoRA* to nonlinear models provide empirical support for the practical usefulness of our merging rule beyond multiclass linear classification. These findings motivate extending the margin-based theoretical analysis to nonlinear adaptation. Third, our theory characterizes margins on the training datasets; establishing corresponding generalization or test-risk guarantees requires further analysis.

Acknowledgments and Disclosure of Funding

Ziqing Xu, Lachlan Ewen MacDonald, Salma Tarmoun, and René Vidal acknowledge support from NSF grant 2031985, Simons Foundation grant 814201 (Theorinet), ONR MURI 503405-78051, and University of Pennsylvania startup funds. Enrique Mallada acknowledges support from NSF Global Centers grant 2330450 and DOE Office of Science (ASCR) award 826565. Hancheng Min acknowledges support from the Science and Technology Commission of Shanghai Municipality through the Shanghai Pujiang Programme (grant 25PJA070) and Shanghai Jiao Tong University startup funds.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference*

on Computer Vision (ECCV), pages 139–154, 2018.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL `http://arxiv.org/abs/1903.04561`.

Etienne Boursier, Loucas Pullaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022.

Phuong Bui Thi Mai and Christoph Lampert. The inductive bias of relu networks on orthogonally separable data. In *9th International Conference on Learning Representations*, 2021.

Christopher J Burges and David Crisp. Uniqueness of the svm solution. *Advances in neural information processing systems*, 12, 1999.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451, 2020.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2021.

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. *arXiv preprint arXiv:2010.11929*, 2021.

Yasushi Esaki, Satoshi Koide, and Takuro Kutsuna. One-shot domain incremental learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079. PMLR, 2022.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, pages 1657–1664, 2013.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *arXiv preprint arXiv:2309.09725*, 2023.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Uijeong Jang, Jason D Lee, and Ernest K Ryu. Lora training in the ntk regime has no spurious local minima. *arXiv preprint arXiv:2402.11867*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data. *Advances in Neural Information Processing Systems*, 36:30167–30221, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization, 2017. URL `https://arxiv.org/abs/1710.03077`.

Xuhong Li, Yves Grandvalet, and François Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning (ICML)*, pages 2825–2834, 2018.

Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, pages 614–629, 2016.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pages 23610–23641. PMLR, 2023.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Fuli Qiao and Mehrdad Mahdavi. Learn more, but bother less: parameter efficient continual learning. *Advances in Neural Information Processing Systems*, 37:97476–97498, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International conference on learning representations*, 2021.

Hrithik Ravi, Clay Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in Neural Information Processing Systems*, 37:81324–81359, 2024.

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

Kurt S Riedel. A sherman–morrison–woodbury identity for rank augmenting matrices with application to centering. *SIAM Journal on Matrix Analysis and Applications*, 13(2):659–662, 1992.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2018.

Jonathan Schwarz, Wojciech M Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning (ICML)*, pages 4528–4537, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Rangarajan K Sundaram. *A first course in optimization theory*. Cambridge university press, 1996.

Taha ValizadehAslani, Yiwen Shi, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. Two-stage fine-tuning with chatgpt data augmentation for learning class-imbalanced data. *Neurocomputing*, 592:127801, 2024.

Hanqing Wang, Bowen Ping, Shuo Wang, Xu Han, Yun Chen, Zhiyuan Liu, and Maosong Sun. Lora-flow: Dynamic lora fusion for large language models in generative tasks. *arXiv preprint arXiv:2402.11455*, 2024.

G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl*, 170(1):33–45, 1992.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.

Martin Wistuba, Prabhu Teja Sivaprasad, Lukas Balles, and Giovanni Zappella. Continual learning with low rank adaptation. *arXiv preprint arXiv:2311.17601*, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. *arXiv preprint arXiv:2503.06982*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 1, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

Wenyu Zhang, Li Shen, Wanyue Zhang, and Chuan-Sheng Foo. Few-shot adaptation of pre-trained networks for domain shift. *arXiv preprint arXiv:2205.15234*, 2022.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *arXiv preprint arXiv:2402.09997*, 2024.

Fariba Zohrizadeh, Mohsen Kheirandishfard, and Farhad Kamangar. Class subset selection for partial domain adaptation. In *CVPR Workshops*, 2019.

Contents

1 Introduction **1**

2 Preliminaries **2**

2.1 Background on multiclass linear classification 3

2.2 Pre-training and fine-tuning setup 3

2.3 Assumptions and their implications 4

3 Main Results **5**

3.1 Global minimum of the LoRA objective 5

3.2 Margin trade-off between forgetting and fine-tuning under LoRA adaptation 6

3.3 Margin characterization of merged LoRA models and optimal mixing coefficients . 7

4 Experiments **9**

4.1 Optimal mixing coefficients for adapter merging: theory vs. grid search 9

4.2 Comparison between MaLoRA and other training-free baselines 9

5 Conclusion, Limitations, and Future Work **10**

A LLM Usage **17**

B Related work **17**

C Characterization of Hard-margin SVM Solution for Orthogonal data **18**

D Proof of Theorem 3.1 **19**

D.1 Proof of Theorem 3.1 19

D.2 Proof of General Version 25

E Proof of Theorem 3.2 and Corollary 3.1 **29**

E.1 Proof of Theorem 3.2 30

E.2 Proof of Corollary 3.1 32

E.3 Proof of General Version 32

F Proof of Theorem 3.3 **36**

F.1 Proof of Theorem 3.3 36

F.2 Proof of General Version 37

G Lemmas **40**

G.1 Preliminary Lemmas 40

G.2 Proof of Lemma D.1 41

G.3 Proof of Lemma D.2 42

G.4 Proof of Lemma G.6 45

H Experiment **46**
H.1 Details for experiments in §H.1.2 and §4.1 46
H.1.1 Empirical Motivation for the Task Geometry Assumption 47
H.1.2 Effect of regularization 49
H.1.3 Details for experiments in §4.1 49
H.2 Details for experiments in §4.2 51
H.2.1 Feature correlations across domains 53
H.2.2 Computational overhead 54

A LLM Usage

In this work, we use LLM to revise the writing, and help build standard training baselines of the experiments, such as dataset downloading, data processing, and pre-trained model loading.

B Related work

Theory of LoRA. There are many works theoretically studying the expressiveness, characterizing the loss landscape, and understanding the learning dynamics of LoRA. Zeng and Lee [2023] prove that, under mild assumptions, LoRA can approximate any deep linear, feed-forward, or transformer network. Within the NTK regime, Malladi et al. [2023] characterize the conditions under which one can study LoRA in the NTK regime, while Jang et al. [2024] show that when the LoRA rank is $r_i \gtrsim \sqrt{N}$, where N is the number of samples, the optimization landscape of LoRA has no spurious local minima, and GD can find $\mathcal{O}(\sqrt{N})$-rank solutions that generalize well. Moreover, Xu et al. [2025] study the learning dynamics of LoRA in the context of matrix factorization, and show that smaller initialization leads to longer training time and lower training error. However, none of these studies theoretically characterize why LoRA reduces *catastrophic forgetting* and support effective *adapter merging*.

LoRA merging. A variety of techniques have been proposed to *merge* task-specific LoRA adapters by forming a weighted average of their parameters. For example, *LoRA-Hub* [Huang et al., 2023] learns per-task mixing coefficients first, and then applies a weighted average of LoRA weight matrices B and A separately. *LoRA-Flow* [Wang et al., 2024] introduces token-level gates that dynamically assign merging coefficients to each LoRA adapter before taking weighted averaging. *LoRA-Retriever* [Zhao et al., 2024] first retrieves the most relevant adapters for each input, and then averages the selected LoRA adapters. Beyond weighted averaging, another line of work attempts to address redundancy and conflicting updates directly. *TIES* [Yadav et al., 2023] proceeds in three steps: trimming redundant parameters, resolving sign conflicts into an aggregate vector, and averaging only the parameters consistent with the aggregate sign, thereby mitigating degradation from redundant or conflicting updates. *DARE* [Yu et al., 2024] can be used as a preprocessing step for other merging methods, where parameters are randomly dropped according to a specified rate and the remaining ones are rescaled, reducing redundancy and potential interference among merged adapters. However, all these methods remain largely heuristic and lack theoretical guarantees.

Regularization and task orthogonality help reduce forgetting. A large body of work in fine-tuning and continual learning shows that *controlling deviation of the fine-tuned model from the pretrained model* reduces catastrophic forgetting. Concretely, these approaches impose *weight-space* penalties that restrict the magnitude and direction of parameter updates away from the pretrained solution. Weight-anchoring and importance-aware penalties exemplify this idea: Kirkpatrick et al. [2017] penalize movement along directions identified by the Fisher information matrix, while Zenke et al. [2017] accumulate path-wise importance during training. Schwarz et al. [2018] extend this idea to long sequences of tasks. Related approaches estimate parameter sensitivity [Aljundi et al., 2018], apply output-level distillation to preserve prior behavior [Li and Hoiem, 2016], or explicitly regularize parameters toward the pretrained anchor [Li et al., 2018]. Although these approaches differ in how importance or anchoring is computed, the common mechanism is the same: *weight-space regularizers constrain updates during fine-tuning relative to the pretrained model to mitigate catastrophic forgetting*. Another complementary perspective highlights the role of task relations. When task representations are sufficiently decorrelated, interference between tasks is naturally reduced and forgetting is alleviated. In linear and kernelized models, it has been shown that when tasks are orthogonal or nearly orthogonal, cross-task gradient interference vanishes in expectation, implying negligible forgetting [Doan et al., 2021, Evron et al., 2022]. These insights resonate with our analysis of LoRA, where we explicitly study a regularized objective on low-rank adapters and make the simplifying assumption that the fine-tuning data are orthogonal to the pretraining data. This assumption directly connects to prior findings on task orthogonality and provides a tractable setting in which we can isolate and analyze the role of regularization in mitigating forgetting.

C Characterization of Hard-margin SVM Solution for Orthogonal data

In this section, we characterize the hard-margin SVM solution for $\bar{K}$-class classification problem for orthogonal data. We show the solution is closely related to the $\bar{K}$-ETF simplex.

Problem Formulation. Consider a $\bar{K}$-class classification problem with data $\{(x_{i,j}, y_i)\}_{i=1,j=1}^{\bar{K},n}$ where:

- $\bar{K} \geq 2$ is the number of classes in the dataset
- n is the number of samples per class
- $y_i \in [\bar{K}]$ denotes the class label for samples $x_{i,j}$
- All data vectors $x_{i,j}$ have unit norm: $\|x_{i,j}\| = 1$
- All data vectors are mutually orthogonal: $x_{i,j}^T x_{k,l} = 0$ for $(i,j) \neq (k,l)$

Let $W \in \mathbb{R}^{\bar{K} \times d}$ be a weight matrix for a linear classifier, where the predicted class for input x is $\arg\max_k (W_k x)$. For simplicity, we use the following notation to denote all the data points: $X = [x_{1,1}, \cdots, x_{\bar{K},n}]$. We use the canonical ETF $M_{\bar{K}} := \sqrt{\bar{K}/(\bar{K}-1)}\big(\boldsymbol{I}_{\bar{K}} - \mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top/\bar{K}\big)$, corresponding to $P = \boldsymbol{I}_{\bar{K}}$ in Definition 2.1.

Now, we are ready to present the main result.

Theorem C.1 (Maximum Margin with Frobenius Norm)**.** *Under the above data assumptions, the largest normalized margin any nonzero linear classifier can achieve is:*

$$\gamma^* = \frac{1}{\sqrt{n(\bar{K}-1)}} \tag{19}$$

The unique unit-margin minimum-Frobenius-norm SVM solution is

$$W^{\mathrm{SVM}} = \sqrt{\frac{\bar{K}-1}{\bar{K}}}\left(M_{\bar{K}} \otimes \mathbf{1}_n^\top\right) X^\top .$$

The normalized margin is maximized precisely by $W = cW^{\mathrm{SVM}}$ with $c > 0$.

Proof. We first solve the unit-margin minimum-norm SVM problem (3) with $\bar{K}$ output classes. By Proposition G.1, its solution has the form $W^{\mathrm{SVM}} = A^* X^\top$ for some $A^* \in \mathbb{R}^{\bar{K} \times \bar{K}n}$. Since $X^\top X = \boldsymbol{I}_{\bar{K}n}$, the logits satisfy $W^{\mathrm{SVM}} X = A^*$ and $\|W^{\mathrm{SVM}}\|_F = \|A^*\|_F$. Thus, A^* solves

$$\begin{aligned} A^* = \underset{A \in \mathbb{R}^{\bar{K} \times \bar{K}n}}{\arg\min} \frac{1}{2}\|A\|_F^2 \quad & \text{subject to} \\ A_{i,(i-1)n+j} - A_{k,(i-1)n+j} \geq 1, \qquad & i,k \in [\bar{K}],\ i \neq k,\ j \in [n]. \end{aligned} \tag{20}$$

This feasible problem has a strictly convex objective and a convex constraint set, so its solution is unique [Burges and Crisp, 1999]. We use the permutation invariances of this fixed-scale problem to characterize A^*. We first decompose A^* into block matrices, i.e., $A^* = \begin{bmatrix} A_1^* & \cdots & A_{\bar{K}}^* \end{bmatrix}$ where $A_i^* \in \mathbb{R}^{\bar{K} \times n}, i \in [\bar{K}]$.

Invariance under column permutation within block matrices. Permuting columns within a class block preserves both the constraints and the objective in (20). Thus, for any permutation matrices $\Pi_i \in \mathbb{R}^{n \times n}$, the matrix $\begin{bmatrix} A_1^*\Pi_1 & \cdots & A_{\bar{K}}^*\Pi_{\bar{K}} \end{bmatrix}$ is also optimal. Uniqueness gives $A_i^*\Pi_i = A_i^*$ for every Π_i, so all columns of each block are equal: $A_i^* = \boldsymbol{v}_i \mathbf{1}_n^\top$ for some $\boldsymbol{v}_i \in \mathbb{R}^{\bar{K}}$.

Invariance under permutation across block matrices. Let $V := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\bar{K}}]$. Simultaneously permuting output rows and the corresponding class blocks preserves the constraints and objective in (20). Uniqueness therefore gives $\Pi V \Pi^\top = V$ for every permutation matrix $\Pi \in \mathbb{R}^{\bar{K} \times \bar{K}}$. By Lemma G.5, $V = a\boldsymbol{I}_{\bar{K}} - b\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top$ for some $a, b \in \mathbb{R}$.

Scalar optimization. The unit-margin constraints reduce to $a \geq 1$, and

$$\begin{aligned}\|A^*\|_F^2 &= n\bar{K}(a-b)^2 + n\bar{K}(\bar{K}-1)b^2 \\ &= n\left[(\bar{K}-1)a^2 + \bar{K}^2\left(b - \frac{a}{\bar{K}}\right)^2\right].\end{aligned}$$

The unique minimum is attained at $a=1$ and $b=1/\bar{K}$. Hence,

$$W^{\mathrm{SVM}} = \left[\left(\boldsymbol{I}_{\bar{K}} - \frac{1}{\bar{K}}\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top\right) \otimes \mathbf{1}_n^\top\right] X^\top, \qquad \|W^{\mathrm{SVM}}\|_F = \sqrt{n(\bar{K}-1)},$$

which gives the stated ETF formula and has $C(W^{\mathrm{SVM}}; \mathcal{D}) = 1$.

Finally, any classifier W with $C(W;\mathcal{D}) > 0$ can be rescaled to the feasible unit-margin classifier $W/C(W;\mathcal{D})$. Minimum-norm optimality yields

$$\gamma(W;\mathcal{D}) = \frac{1}{\|W/C(W;\mathcal{D})\|_F} \leq \frac{1}{\|W^{\mathrm{SVM}}\|_F} = \frac{1}{\sqrt{n(\bar{K}-1)}}.$$

Equality holds precisely when $W/C(W;\mathcal{D}) = W^{\mathrm{SVM}}$, by uniqueness of the unit-margin minimum-norm solution. Thus all normalized-margin maximizers are exactly the positive multiples of W^{SVM}. Classifiers with nonpositive margin cannot attain this positive optimum. □

D Proof of Theorem 3.1

In this section, we first present the proof of Theorem 3.1. Then, under additional assumptions on ϵ_1, ϵ_2, we show that the same conclusion still holds, with additional error terms that depend on ϵ_1, ϵ_2.

D.1 Proof of Theorem 3.1

For the case $2 \leq \bar{K} < K$, we first define the following functions used in the proof of Theorem 3.1. The balanced case $\bar{K} = K$ is treated separately below.

$$\begin{aligned}
g_N(x) &:= \sqrt{\frac{\rho(K-\bar{K})}{\bar{K}(K-\bar{K})x^2 + \bar{K}K(1-x)^2}} \\
g_a(x) &:= \exp\left(\frac{g_N(x)\sqrt{\frac{\rho}{\bar{K}n}}}{\sqrt{x^2 + \frac{K}{K-\bar{K}}(1-x)^2}}\left[x^2 + \sqrt{\frac{K}{K-\bar{K}}}(1-x)^2\right]\right) \\
g_b(x) &:= \exp\left(\frac{g_N(x)\sqrt{\frac{\rho}{\bar{K}n}}}{\sqrt{x^2 + \frac{K}{K-\bar{K}}(1-x)^2}}\left[-\frac{x^2}{\bar{K}-1} + \sqrt{\frac{K}{K-\bar{K}}}(1-x)^2\right]\right) \\
g_c(x) &:= \exp\left(\frac{g_N(x)\sqrt{\frac{\rho}{\bar{K}n}}}{\sqrt{x^2 + \frac{K}{K-\bar{K}}(1-x)^2}}\left[-\frac{\bar{K}}{K-\bar{K}} \cdot \sqrt{\frac{K}{K-\bar{K}}}(1-x)^2\right]\right).
\end{aligned} \tag{21}$$

Moreover, let $x_0 \in [0,1]$ be the root of the equation

$$\frac{g_b(x)}{g_c(x)} = \frac{1}{\sqrt{\frac{K}{K-\bar{K}}} + 1}.$$

Remark D.1. *Our proof technique follows Fang et al. [2021], and the readers can see the same definition of $g_a(x), g_b(x), g_c(x)$ and x_0 in Fang et al. [2021]. Moreover, the authors prove the existence of x_0 under certain conditions.*

Now, we begin the proof of Theorem 3.1.

Theorem D.1 (Restatement of Theorem 3.1)**.** *Let* $\left(B_\lambda^*, A_\lambda^*\right)$ *be any minimizer of* (5)*. There exist scalar functions* $a_\lambda,\, b_\lambda,\, c_\lambda, \Theta_\lambda$ *of* λ *such that for any global minimizer* $\left(B_\lambda^*, A_\lambda^*\right)$ *of* (5) *with* $\lambda > 0$*, we have*

$$B_\lambda^* A_\lambda^* = \left(\begin{pmatrix} (a_\lambda + b_\lambda)\, \mathbf{I}_{\bar{K}} \;-\; b_\lambda\, \mathbf{1}_{\bar{K}} \mathbf{1}_{\bar{K}}^\top \\ -c_\lambda\, \mathbf{1}_{K-\bar{K}} \mathbf{1}_{\bar{K}}^\top \end{pmatrix} \otimes \mathbf{1}_n^\top \right) X_{\mathrm{ft}}^\top\,, \quad \|B_\lambda^* A_\lambda^*\|_F = \Theta_\lambda\,. \tag{22}$$

When $\bar{K} \geq 2$*, there exists a critical regularization weight* $\lambda_{\mathrm{crit}} \in [0, \frac{1}{\bar{K}\sqrt{n}})$*, with* $\lambda_{\mathrm{crit}} = 0$ *if and only if* $\bar{K} = K$*, such that the scalar functions* $a_\lambda,\, b_\lambda,\, c_\lambda, \Theta_\lambda$ *are characterized as follows:*

(i) High-penalty regime $\left(\lambda \geq \frac{1}{\bar{K}\sqrt{n}}\right)$*:* $a_\lambda = b_\lambda = c_\lambda = \Theta_\lambda = 0$*, thus* $B_\lambda^* A_\lambda^* = \mathbf{0}_{K \times d}$*.*

(ii) Intermediate regime $\left(\lambda_{\mathrm{crit}} < \lambda < \frac{1}{\bar{K}\sqrt{n}}\right)$*:* $a_\lambda = \frac{\Theta_\lambda \sqrt{\bar{K}-1}}{\bar{K}\sqrt{n}}$, $b_\lambda = \frac{\Theta_\lambda}{\bar{K}\sqrt{n(\bar{K}-1)}}$, $c_\lambda = 0$*, and* Θ_λ *is the unique root of a nonlinear equation (see* (29) *in Appendix D), thus the minimizer is*

$$B_\lambda^* A_\lambda^* \;=\; \frac{\Theta_\lambda}{\sqrt{\bar{K}n}} \left(\begin{pmatrix} M_{\bar{K}} \\ \mathbf{0}_{(K-\bar{K}) \times \bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top \right) X_{\mathrm{ft}}^\top. \tag{23}$$

(iii) Low-penalty regime $\left(0 < \lambda \leq \lambda_{\mathrm{crit}}\right)$*:* $a_\lambda, b_\lambda, c_\lambda, \Theta_\lambda$ *are non-negative.*

Moreover, the small-regularization limit satisfies

$$\lim_{\lambda \to 0^+} \Theta_\lambda = \infty\,, \qquad \lim_{\lambda \to 0^+} \frac{B_\lambda^* A_\lambda^*}{\|B_\lambda^* A_\lambda^*\|_F} = \sqrt{\frac{1}{n\bar{K}}} \left(M_K^{(\bar{K})} \otimes \mathbf{1}_n^\top \right) X_{\mathrm{ft}}^\top\,, \tag{24}$$

where $M_K^{(\bar{K})}$ *is the first* $\bar{K}$ *columns of the K-ETF matrix* M_K*.*

Proof. We first show that one can simplify the objective in (5) based on the following Lemma.

Lemma D.1. *Under Assumptions 2.1 and 2.4 with* $\epsilon_1 = \epsilon_2 = 0$*, the original fine-tuning problem in* (5) *is equivalent to the following symmetric formulation:*

$$\min_{B, \tilde{A}} \; \mathcal{L}(B, \tilde{A}) := \frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \mathcal{L}_{\mathrm{CE}}\big(y_c,\, B\tilde{A} e_{n(c-1)+j}\big) + \frac{\lambda}{2}\big(\|B\|_F^2 + \|\tilde{A}\|_F^2\big)\,, \tag{25}$$

where $X_{\mathrm{ft}} = [\bar{x}_{1,1}, \ldots, \bar{x}_{1,n}, \cdots, \bar{x}_{\bar{K},1}, \cdots \bar{x}_{\bar{K},n}], B \in \mathbb{R}^{K \times r}, \tilde{A} = A X_{\mathrm{ft}} \in \mathbb{R}^{r \times \bar{K}n}$*.*
Let $X_{\mathrm{ft},\perp}$ *be an orthonormal complement of* X_{ft}*. Then:*

1. *If* (B^*, A^*) *is any global minimizer of* (5)*, then*

$$(B^*,\, \tilde{A}^*) \quad \textit{with} \quad \tilde{A}^* = A^* X_{\mathrm{ft}}$$

   is a global minimizer of (25)*.*

2. *Conversely, if* $(B^*, \tilde{A}^*)$ *is any global minimizer of* (25)*, then* $(B^*, \tilde{A}^* X_{\mathrm{ft}}^\top)$ *yields a global minimizer of the original problem in* (5)*.*

We refer the readers to Appendix G.2 for the proof of Lemma D.1. Lemma D.1 indicates that to characterize the global minimizer of the objective in (5), one can instead characterize the global minimizer of the objective in (25).

We start with the following equations that characterize the global minimizer of (25)

$$\frac{\partial \mathcal{L}(B, \tilde{A})}{\partial B} = \lambda B + \frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \frac{\partial}{\partial B\tilde{A}} \left(\mathcal{L}_{\mathrm{CE}} \left(y_c, B\tilde{A} e_{(c-1)n+j} \right) \right) e_{(c-1)n+j}^\top \tilde{A}^\top = 0$$

$$\frac{\partial \mathcal{L}(B, \tilde{A})}{\partial \tilde{A}} = \lambda \tilde{A} + B^\top \frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \frac{\partial}{\partial B\tilde{A}} \left(\mathcal{L}_{\mathrm{CE}} \left(y_c, B\tilde{A} e_{(c-1)n+j} \right) \right) e_{(c-1)n+j}^\top = 0\,.$$

Based on the above equation, we can conclude that any minimizer $(B, \tilde{A})$ of (25) must satisfy

$$\lambda B^\top B = \lambda \tilde{A}\tilde{A}^\top \Rightarrow \|B\|_F^2 = \|\tilde{A}\|_F^2 . \tag{26}$$

Our second step is to converge the problem into a constraint optimization problem. The following lemma characterizes this.

Lemma D.2. *Let $B \in \mathbb{R}^{K\times r}$, $\tilde{A} \in \mathbb{R}^{r\times \bar{K}n}$, and assume $r \geq \bar{K}$ as in Assumption 2.3. For each $\lambda > 0$, there exists a unique non-negative value ρ_λ such that the solution set for the following optimization problems are equivalent*

$$\min_{B,\tilde{A}} \ \frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \mathcal{L}_{\mathrm{CE}}\big(y_c,\, B\tilde{A}e_{(c-1)n+j}\big) + \frac{\lambda}{2}\big(\|B\|_F^2 + \|\tilde{A}\|_F^2\big) , \tag{Problem one}$$

$$\min_{B,\tilde{A}} \ \frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \mathcal{L}_{\mathrm{CE}}\big(y_c,\, B\tilde{A}e_{(c-1)n+j}\big) \quad \textit{s.t.} \|B\|_F^2 \leq \rho_\lambda, \|\tilde{A}\|_F^2 \leq \rho_\lambda . \tag{Problem two}$$

Moreover, the map $\lambda \mapsto \rho_\lambda$ enjoys the following properties:

(i) ***Monotonicity and continuity.*** *ρ_λ is* continuous *and* non-increasing *on* $(0, \infty)$.

(ii) ***Asymptotic behaviour.***

$$\lim_{\lambda\to 0^+} \rho_\lambda = \infty, \qquad \lim_{\lambda\to\infty} \rho_\lambda = 0.$$

(iii) ***Flat regions.*** *For any $0 < \lambda_1 < \lambda_2$, $\rho(\lambda_1) = \rho(\lambda_2) \iff \rho(\lambda_1) = \rho(\lambda_2) = 0$.*

We refer the readers to Appendix G.3 for the proof.

Balanced case $\bar{K} = K \geq 2$. By Lemmas D.1 and D.2, it suffices to solve the convex problem (Problem three). Let $\ell(Z)$ denote its mean cross-entropy term and set $C := (\boldsymbol{I}_K - \mathbf{1}_K\mathbf{1}_K^\top/K) \otimes \mathbf{1}_n^\top$. The $K-1$ nonzero singular values of C equal $\sqrt{n}$, so for $t > 0$,

$$\nabla\ell(tC) = -\frac{C}{n(e^t + K - 1)} , \qquad \frac{C}{\sqrt{n}} \in \partial\|tC\|_* .$$

Thus the global optimality condition $0 \in \nabla\ell(Z) + \lambda\partial\|Z\|_*$ holds at $Z = tC$ exactly when $e^t + K - 1 = 1/(\lambda\sqrt{n})$. At zero, $\|\nabla\ell(0)\|_{\mathrm{op}} = 1/(K\sqrt{n})$, so zero is optimal exactly when $\lambda \geq 1/(K\sqrt{n})$. Uniqueness (Lemma G.6) therefore gives

$$Z_\lambda^* = t_\lambda C , \qquad t_\lambda = \begin{cases} \log\big(\frac{1}{\lambda\sqrt{n}} - (K-1)\big), & 0 < \lambda < \frac{1}{K\sqrt{n}}, \\ 0, & \lambda \geq \frac{1}{K\sqrt{n}}. \end{cases}$$

Since $\mathrm{rank}(Z_\lambda^*) \leq K-1 \leq r$, this solution is feasible for the factorized problem. Writing $\bar{A}_\lambda^* := A_\lambda^* X_{\mathrm{ft}}$, every global minimizer satisfies $B_\lambda^*\bar{A}_\lambda^* = Z_\lambda^*$ and $B_\lambda^* A_\lambda^* = t_\lambda C X_{\mathrm{ft}}^\top$. Consequently, $\rho_\lambda = \sqrt{n}(K-1)t_\lambda$ and $\Theta_\lambda = \sqrt{n(K-1)}t_\lambda$. These formulas prove the intermediate ETF form and both limits in (9). We set $\lambda_{\mathrm{crit}} = 0$ and $\rho_{\mathrm{crit}} = +\infty$ as limiting boundaries; no minimizer at $\lambda = 0$ is asserted.

For the remainder of this strict-orthogonal proof, assume $2 \leq \bar{K} < K$. For simplicity, we will use ρ to denote ρ_λ for convenience.

Characterization to solution of (Problem two)**.** (Problem two) can be viewed as a special case of the neural collapse phenomenon in the extreme setting of an imbalanced dataset: for the first $\bar{K}$ classes we observe n data points per class, while all remaining classes have zero samples. In this context, we directly invoke the following result from Fang et al. Fang et al. [2021] (see p. 26 in . Fang et al. [2021]).

Lemma D.3 (Neural collapse under extreme imbalance; [Fang et al., 2021, Page 26]). *Assume $2 \le \bar{K} < K$. For any constants $c_1, c_2, c_3, \rho > 0$, define*

$$
\begin{aligned}
c_1' &= \frac{c_1}{c_1 + (\bar{K}-1)c_2 + (K-\bar{K})c_3} \\
c_2' &= \frac{c_2}{c_1 + (\bar{K}-1)c_2 + (K-\bar{K})c_3} \\
c_3' &= \frac{c_3}{c_1 + (\bar{K}-1)c_2 + (K-\bar{K})c_3} \\
c_4 &= -c_1' \log c_1' - c_2'(\bar{K}-1)\log c_2' - (K-\bar{K})c_3' \log(c_3') \\
c_5 &= \frac{\bar{K}c_2}{\bar{K}c_2 + (K-\bar{K})c_3} \\
c_6 &= \frac{(K-\bar{K})c_3}{\bar{K}c_2 + (K-\bar{K})c_3} \\
c_7 &= \frac{\bar{K}c_2 + (K-\bar{K})c_3}{c_1 + (\bar{K}-1)c_2 + (K-\bar{K})c_3}
\end{aligned}
$$

For any feasible solution $(B, \tilde{A})$ of (Problem two)*, the objective value of* (Problem two) *can be simplified to*

$$
\frac{1}{\bar{K}n} \sum_{c=1}^{\bar{K}} \sum_{j=1}^{n} \mathcal{L}_{\mathrm{CE}}\big(y_c,\, B\tilde{A}e_{(c-1)n+j}\big) \overset{a}{\ge} -\frac{c_7}{\bar{K}} \sqrt{\frac{\rho}{n}} \sqrt{\sum_{i=1}^{\bar{K}} \|c_5 w_1 + c_6 w_2 - b_i\|^2} + c_4
$$

where $w_1 = \frac{1}{\bar{K}} \sum_{i=1}^{\bar{K}} b_i, w_2 = \frac{1}{K-\bar{K}} \sum_{i=\bar{K}+1}^{K} b_i$, and $\overset{a}{\ge}$ becomes equality under certain choices of c_1, c_2, c_3.

Readers are referred to Fang et al. [2021] for the full proof. Note the following distinction between Lemma D.3 and its analogue in Fang et al. [2021]: the inequality in Fang et al. [2021] becomes tight only when all minority-class features are zero and certain choices of c_1, c_2, c_3, whereas in our lemma, we do not require that all minority-class features are zero. This difference is resolved by our change of variables $\tilde{A} = AX_{\mathrm{ft}}$, which effectively confines $\tilde{A}$ to the subspace of observed (majority-class) features and thus guarantees equality in Lemma D.3.

Now, we carefull pick the $B, \tilde{A}$ and c_1, c_2, c_3 to achieve the global minimum of the objective in (Problem two). The following lemma characterizes this exactly.

Lemma D.4 (Lemma 5 in Fang et al. [2021]). *Under the assumptions of Theorem 3.1, with $2 \le \bar{K} < K$, define*

$$
\rho_{\mathrm{crit}} = \sqrt{n}\bar{K}(\bar{K}-1)\log\left(\sqrt{\frac{K}{K-\bar{K}}}+1\right)
$$

such that the optimal value to (Problem two) *is as follows*

- *when $\bar{K} \ge 2$ and $\rho_\lambda < \rho_{\mathrm{crit}}$, we choose $c_1 = \exp\left(\frac{\rho_\lambda}{K\sqrt{n}}\right), c_2 = \exp\left(\frac{-\rho_\lambda}{\sqrt{n}\bar{K}(K-1)}\right), c_3 = 1$, and* (Problem two) *attain its minimum*

$$
-\frac{\rho_\lambda}{\sqrt{n}\bar{K}} + \log\left(K - \bar{K} + (\bar{K}-1)\exp\big(-\frac{\rho_\lambda}{\sqrt{n}\bar{K}(\bar{K}-1)}\big) + \exp\big(\frac{\rho_\lambda}{\sqrt{n}\bar{K}}\big)\right),
$$

where $B^*, \tilde{A}^*$ takes the following form

$$
\left[b_1^*, b_2^*, \ldots, b_{\bar{K}}^*\right] = \sqrt{n}\left[\bar{a}_1^*, \bar{a}_2^*, \ldots, \bar{a}_{\bar{K}}^*\right] = \sqrt{\frac{\rho}{\bar{K}}}\, PM_{\bar{K}},
$$

$$
\begin{aligned}
a_i^* &= \bar{a}_{\lceil i/n \rceil}^*, \quad i = 1, \ldots, \bar{K}n, \\
b_j^* &= 0, \qquad j > \bar{K}.
\end{aligned}
$$

Here $P \in \mathbb{R}^{r\times \bar{K}}$ satisfies $P^\top P = \mathbf{I}_{\bar{K}}$, a_i^* is the i-th column of $\tilde{A}^*$, and $\bar{a}_c^*$ is the common column for samples in class c. Writing $\bar{A}^* := \tilde{A}^* = A^* X_{\mathrm{ft}}$ for the data-coordinate factor, the product takes the following form

$$B^*\bar{A}^* \;=\; \frac{\rho_\lambda}{\sqrt{\bar{K}n(\bar{K}-1)}} \begin{pmatrix} M_{\bar{K}} \\ \mathbf{0}_{(K-\bar{K})\times\bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top . \tag{27}$$

- *when $\bar{K} \geq 2$ and $\rho \geq \rho_{\mathrm{crit}}$, one choose*

$$c_1 = g_a(x_0), \quad c_2 = g_b(x_0), \quad c_3 = g_c(x_0)\,,$$

where $g_a(x), g_b(x), g_c(x)$ and x_0 are defined in (21)*. Then,* (Problem two) *attain its minimum*

$$\log\!\left(\frac{g_a(x_0) + (\bar{K}-1)g_b(x_0) + (K-\bar{K})g_c(x_0)}{g_a(x_0)}\right),$$

where

$$b_i^* \;=\; \begin{cases} g_N(x_0)\,P\!\left[\dfrac{\bar{K}\,x_0}{\sqrt{\bar{K}(\bar{K}-1)}}\,\bar{y}_i + \Big(\dfrac{1-x_0}{\sqrt{\bar{K}}} - \dfrac{x_0}{\sqrt{\bar{K}(\bar{K}-1)}}\Big)\mathbf{1}_{\bar{K}}\right], & i \leq \bar{K}, \\ -\dfrac{g_N(x_0)\,\sqrt{\bar{K}}\,(1-x_0)}{K-\bar{K}}\;P\mathbf{1}_{\bar{K}}, & i > \bar{K}. \end{cases}$$

$$\bar{a}_c^* \;=\; \frac{g_N(x_0)}{\sqrt{n}} P\!\left[x_0\sqrt{\frac{\bar{K}}{\bar{K}-1}}\bar{y}_c + \Big(\frac{(1-x_0)\sqrt{\frac{K}{K-\bar{K}}}}{\sqrt{\bar{K}}} - \frac{x_0}{\sqrt{\bar{K}(\bar{K}-1)}}\Big)\mathbf{1}_{\bar{K}}\right],$$

where $\bar{y}_c \in \mathbb{R}^{\bar{K}}$ is a one-hot vector with the c-th entry equal to one (and $\bar{y}_i$ is defined analogously with the i-th entry equal to one), and $P \in \mathbb{R}^{r\times\bar{K}}$ is a partial orthogonal matrix such that $P^\top P = \mathbf{I}_{\bar{K}}$. The sample columns satisfy $a^*_{(c-1)n+j} = \bar{a}_c^*$ for $c \in [\bar{K}]$ and $j \in [n]$. Moreover, $B^*\tilde{A}^*$ takes the following form

$$B^*\tilde{A}^* = \begin{pmatrix} (\log g_a(x_0) - \log g_b(x_0))\mathbf{I}_{\bar{K}} + \log g_b(x_0)\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top \\ \log g_c(x_0)\mathbf{1}_{K-\bar{K}}\mathbf{1}_{\bar{K}}^\top \end{pmatrix} \otimes \mathbf{1}_n^\top .$$

Lemma D.4 is Lemma 5 in Fang et al. [2021]. We refer the readers to Fang et al. [2021] for the proof. Lemma D.4 exactly characterize the optimal solution and value of the objective in (Problem two).

Computation of the product of LoRA adapters. Based on Lemma D.4, we can characterize the solution to the LoRA objective in (5), then it suffices to compute the product of the minimizer to prove Theorem 3.1.

In the **Intermediate regime**, restricting B^* to its first $\bar{K}$ rows gives

$$\begin{aligned} \big(B^*\big)_{1:\bar{K},:}\,[\bar{a}_1^*, \bar{a}_2^*, \ldots, \bar{a}_{\bar{K}}^*] &= \frac{\rho_\lambda}{\bar{K}\sqrt{n}} M_{\bar{K}}^\top P^\top P M_{\bar{K}} \\ &= \frac{\rho_\lambda}{\bar{K}\sqrt{n}}\,\frac{\bar{K}}{\bar{K}-1}\big(\boldsymbol{I}_{\bar{K}} - \frac{1}{\bar{K}}\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top\big) \\ &= \frac{\rho_\lambda}{\sqrt{\bar{K}n(\bar{K}-1)}} M_{\bar{K}}\,, \end{aligned}$$

and it leads to the product of final solution

$$B^*\tilde{A}^* \;=\; \frac{\rho_\lambda}{\sqrt{\bar{K}n(\bar{K}-1)}} \begin{pmatrix} M_{\bar{K}} \\ \mathbf{0}_{(K-\bar{K})\times\bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top . \tag{28}$$

In the **Low penalty regime**, in Fang et al. [2021], the authors further show that

$$(b_i^*)^\top \bar{a}_j^* = \begin{cases} \log g_a(x_0) & i = j \leq \bar{K} \\ \log g_b(x_0) & i \neq j, i \leq \bar{K} \\ \log g_c(x_0) & \bar{K} < i \leq K \end{cases} \qquad i \in [K],\ j \in [\bar{K}].$$

Thus, one can show that

$$B^*\tilde{A}^* = \begin{pmatrix} (\log g_a(x_0) - \log g_b(x_0))\mathbf{I}_{\bar{K}} + \log g_b(x_0)\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top \\ \log g_c(x_0)\mathbf{1}_{K-\bar{K}}\mathbf{1}_{\bar{K}}^\top \end{pmatrix} \otimes \mathbf{1}_n^\top .$$

Characterization of λ_{crit}. In Lemma D.4, when $\bar{K} \geq 2$, we see that the solution takes different form based on whether $\rho_\lambda < \rho_{\text{crit}}$ or $\rho_\lambda \geq \rho_{\text{crit}}$. Moreover, based on Lemma D.2, there must exist a unique λ_{crit} such that $\rho_{\text{crit}} = \rho(\lambda_{\text{crit}})$. Now, we characterize the range of λ_{crit} here.

Due to Lemma D.4, we know when $\lambda > \lambda_{\text{crit}}$, the solution takes the form

$$B^*\bar{A}^* = \frac{\rho}{\sqrt{n}(\bar{K}-1)} \begin{pmatrix} \mathbf{I}_{\bar{K}} - \frac{1}{\bar{K}}\mathbf{1}_{\bar{K}}\mathbf{1}_{\bar{K}}^\top \\ \mathbf{0}_{(K-\bar{K})\times\bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top ,$$

and adding the penalty $\lambda\rho$ to this constrained minimum cross entropy gives the scalar penalized objective

$$\psi(\rho) := \lambda\rho - \frac{\rho}{\sqrt{n\bar{K}}} + \log\left(K - \bar{K} + (\bar{K}-1)\exp\big(-\frac{\rho}{\sqrt{n\bar{K}(\bar{K}-1)}}\big) + \exp\big(\frac{\rho}{\sqrt{n\bar{K}}}\big) \right),$$

For this ETF-form update, $\rho = \sqrt{\bar{K}-1}\Theta$. Differentiating $\psi(\rho)$ and substituting this relation into the stationarity condition gives

$$\begin{aligned} &\frac{d\psi(\rho)}{d\rho} = 0 \\ \iff &(\lambda - \frac{1}{\sqrt{n\bar{K}}}) + \frac{1}{\sqrt{n\bar{K}}} \cdot \frac{\exp\big(\frac{\rho}{\sqrt{n\bar{K}}}\big) - \exp\big(-\frac{\rho}{\sqrt{n\bar{K}(\bar{K}-1)}}\big)}{K - \bar{K} + (\bar{K}-1)\exp\big(-\frac{\rho}{\sqrt{n\bar{K}(\bar{K}-1)}}\big) + \exp\big(\frac{\rho}{\sqrt{n\bar{K}}}\big)} = 0 \\ \iff &(\lambda - \frac{1}{\sqrt{n\bar{K}}}) + \frac{1}{\sqrt{n\bar{K}}} \cdot \frac{\exp\big(\frac{\Theta}{\sqrt{n(\bar{K}-1)}}\big) - 1}{(K-\bar{K})\exp\big(\frac{\Theta}{\bar{K}\sqrt{n(\bar{K}-1)}}\big) + (\bar{K}-1) + \exp\big(\frac{\Theta}{\sqrt{n(\bar{K}-1)}}\big)} = 0 \end{aligned} \quad (29)$$

We argue by contradiction. Suppose $\lambda_{\text{crit}} \geq 1/(\sqrt{n}\bar{K})$. For every $\lambda > \lambda_{\text{crit}}$, the preceding ETF form (including the zero product) applies and (29) gives $d\psi/d\rho > 0$ for $\rho > 0$, so $\rho_\lambda = 0$. Letting $\lambda \downarrow \lambda_{\text{crit}}$ and using continuity from Lemma D.2 yields $\rho(\lambda_{\text{crit}}) = 0$, contradicting its defining value $\rho_{\text{crit}} > 0$. This covers both equality and strict inequality in the contradiction hypothesis. Therefore

$$\lambda_{\text{crit}} < \frac{1}{\sqrt{n\bar{K}}} .$$

Characterization of ρ_λ. In this part, we characterize the ρ_λ when $\lambda > \lambda_{\text{crit}}$. By Lemma D.4, in this ETF regime $\psi(\rho)$ equals the constrained minimum cross entropy at budget ρ, plus the penalty $\lambda\rho$. Thus, ρ_λ must attain the minimum of $\psi(\rho)$. Based on (29), one can first see that if $\lambda \geq \frac{1}{\sqrt{n\bar{K}}}$, $\frac{d\psi(\rho)}{d\rho} > 0$, and the minimum is attained when $\rho_\lambda = 0$. When $\lambda_{\text{crit}} < \lambda < \frac{1}{\sqrt{n\bar{K}}}$, solve (29) for Θ_λ and recover $\rho_\lambda = \sqrt{\bar{K}-1}\Theta_\lambda$.

Asymptotic behaviour of ρ_λ. On one hand, as $\lambda \to 0+$, one obviously has $\lim_{\lambda\to 0+} \|B^*\tilde{A}^*\|_F = \infty$, since by Lemma D.2 we have $Z_\lambda^* = B^*\tilde{A}^*$ where Z_λ^* is the unique minimizer of (Problem three) and $\|Z_\lambda^*\|_* = \rho_\lambda \to \infty$. Then, we characterize its asymptotic direction, which is equivalent to study the limit of $\frac{\log g_a(x_0)}{\log g_b(x_0)}$ and $\frac{\log g_a(x_0)}{\log g_c(x_0)}$ as $\lambda \to 0+$. We will use $x_{0,\lambda}$ denote the choice of x_0 for fix regularization effect λ.

First, the equation to derive for x_0 is as follows

$$\begin{aligned} &\frac{g_b(x)}{g_c(x)} = \frac{1}{1 + \sqrt{\frac{K}{K-\bar{K}}}} \\ \iff &\frac{\rho_\lambda}{\bar{K}\sqrt{n}} \cdot \frac{-\frac{x^2}{\bar{K}-1} + (\frac{K}{K-\bar{K}})^{3/2}(1-x)^2}{x^2 + \frac{K}{K-\bar{K}}(1-x)^2} = -\log\big(1 + \sqrt{\frac{K}{K-\bar{K}}}\big) \end{aligned}$$

Since $\lim_{\lambda\to 0+}\rho_\lambda=\infty$, one must have $x_{0,\lambda}$ approaches the solution of the following quadratic equation

$$\frac{x_{0,\lambda}^2}{\bar{K}-1}=\big(\frac{K}{K-\bar{K}}\big)^{3/2}(1-x_{0,\lambda})^2\,.$$

Moreover, one can see that the expression for $\frac{\log g_a(x_{0,\lambda})}{\log g_b(x_{0,\lambda})}$ and $\frac{\log g_a(x_{0,\lambda})}{\log g_c(x_{0,\lambda})}$ takes the following form

$$\frac{\log g_a(x_{0,\lambda})}{\log g_b(x_{0,\lambda})}=\frac{x_{0,\lambda}^2+\sqrt{\frac{K}{K-\bar{K}}}(1-x_{0,\lambda})^2}{-\frac{x_{0,\lambda}^2}{\bar{K}-1}+\sqrt{\frac{K}{K-\bar{K}}}(1-x_{0,\lambda})^2}$$

$$\frac{\log g_a(x_{0,\lambda})}{\log g_c(x_{0,\lambda})}=\frac{x_{0,\lambda}^2+\sqrt{\frac{K}{K-\bar{K}}}(1-x_{0,\lambda})^2}{-\frac{\bar{K}}{K-\bar{K}}\cdot\sqrt{\frac{K}{K-\bar{K}}}(1-x_{0,\lambda})^2}\,.$$

Thus, combine the above two equations, we can compute that

$$\lim_{\lambda\to 0+}\frac{\log g_a(x_{0,\lambda})}{\log g_b(x_{0,\lambda})}=\lim_{\lambda\to 0+}\frac{\log g_a(x_{0,\lambda})}{\log g_c(x_{0,\lambda})}=-(K-1)\,.$$

Connection between ρ_λ and Θ_λ. In previous analysis, we have shown that

$$\rho_\lambda=\|B_\lambda^*A_\lambda^*\|_*\,,\qquad \Theta_\lambda=\|B_\lambda^*A_\lambda^*\|_F \tag{30}$$

It remains to relate ρ_λ and Θ_λ across the regularization regimes. Write $\bar{A}_\lambda^*:=A_\lambda^*X_{\mathrm{ft}}$ for the data-coordinate factor denoted by $\bar{A}^*$ above. By Lemma D.1, $B_\lambda^*A_\lambda^*=B_\lambda^*\bar{A}_\lambda^*X_{\mathrm{ft}}^\top$. Since $X_{\mathrm{ft}}^\top X_{\mathrm{ft}}=\boldsymbol{I}_{\bar{K}n}$, the weight-space and data-coordinate products have the same nonzero singular values, and hence the same nuclear and Frobenius norms.

First, it is obvious in the high-penalty regime, $\rho_\lambda=\Theta_\lambda=0$.

Second, in the *intermediate regime*, the data-coordinate product is

$$B_\lambda^*\bar{A}_\lambda^*\;=\;\frac{\rho_\lambda}{\sqrt{\bar{K}n(\bar{K}-1)}}\begin{pmatrix}M_{\bar{K}}\\ \mathbf{0}_{(K-\bar{K})\times\bar{K}}\end{pmatrix}\otimes\mathbf{1}_n^\top\,.$$

Consequently,

$$\rho_\lambda=\|B_\lambda^*\bar{A}_\lambda^*\|_*\,,\qquad \Theta_\lambda=\|B_\lambda^*\bar{A}_\lambda^*\|_F=\frac{\rho_\lambda}{\sqrt{\bar{K}-1}}\,.$$

Equivalently, the data-coordinate product in the *intermediate regime* is

$$B_\lambda^*\bar{A}_\lambda^*\;=\;\frac{\Theta_\lambda}{\sqrt{n\bar{K}}}\begin{pmatrix}M_{\bar{K}}\\ \mathbf{0}_{(K-\bar{K})\times\bar{K}}\end{pmatrix}\otimes\mathbf{1}_n^\top\,.$$

Multiplying by $X_{\mathrm{ft}}^\top$ recovers the weight-space formula (8). Third, in the *low-penalty regime*, the preceding asymptotic analysis gives $\lim_{\lambda\to 0+}\|B_\lambda^*\bar{A}_\lambda^*\|_F=\infty$ and characterizes the limiting direction.

The solution of (5). Finally, one can use Lemma D.1 to characterize the optimal solution of (5) based on the above results.

□

D.2 Proof of General Version

In this subsection we present the general version of Theorem 3.1: under the near-orthogonal regime in Assumption 2.1 (i.e., ϵ_1,ϵ_2 small but possibly nonzero), the optimal LoRA update $B_\lambda^*A_\lambda^*$ remains close, in Frobenius norm, to its closed-form orthogonal-case description, with the deviation controlled linearly in ϵ_1 and ϵ_2.

Notation. Write $G:=X_{\mathrm{ft}}^\top X_{\mathrm{ft}}\in\mathbb{R}^{\bar{K}n\times\bar{K}n}$, $R:=G^{1/2}$, $Q:=X_{\mathrm{ft}}R^{-1}\in\mathbb{R}^{d\times\bar{K}n}$ (so that $Q^\top Q=I_{\bar{K}n}$ and $X_{\mathrm{ft}}=QR$), $S:=W_{\mathrm{pre}}X_{\mathrm{ft}}\in\mathbb{R}^{K\times\bar{K}n}$, and let $N_{\mathrm{pre}}:=|\mathcal{D}_{\mathrm{pre}}|$ be the number

of pre-training samples. Set $\delta := (\bar{K}n - 1)\,\epsilon_1$. For any $\bar{S} \in \mathbb{R}^{K\times \bar{K}n}$ and any positive-definite $\bar{R} \in \mathbb{R}^{\bar{K}n\times \bar{K}n}$, define

$$\Phi(Z;\bar{S},\bar{R}) := \frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\mathrm{CE}}\big(y_c,(Z\bar{R}+\bar{S})\,e_{(c-1)n+j}\big) + \lambda\|Z\|_*\,, \tag{31}$$

$$Z_\lambda^*(\bar{S},\bar{R}) := \arg\min_{Z\in\mathbb{R}^{K\times\bar{K}n}} \Phi(Z;\bar{S},\bar{R})\,. \tag{32}$$

The minimizer Z_λ^* is unique by Lemma G.6 (for $\bar{S}=0, \bar{R}=I$; the same argument applies for general $\bar{R}\succ 0$).

Theorem D.2 (General version of Theorem 3.1)**.** *Fix $\lambda>0$ and suppose Assumptions 2.1–2.4 hold with $\epsilon_1,\epsilon_2\in[0,1)$. There exist constants $\delta_0\in(0,1/2]$, $s_0\in(0,1]$, and $\mu_0, C_0>0$, depending on $\lambda, K, \bar{K}, n, \|Z_\lambda^*(0,I)\|_*$ but not on ϵ_1,ϵ_2, such that whenever $\delta := (\bar{K}n-1)\,\epsilon_1 \le \delta_0$ and $\|S\|_F \le s_0$,*

$$\big\|B_\lambda^*A_\lambda^* - Z_\lambda^*(0,I)\,X_{\mathrm{ft}}^\top\big\|_F \le \frac{C_0}{\mu_0}\Big(\|S\|_F + \delta\,\big(1+\|Z_\lambda^*(0,I)\|_F\big)\Big)\,. \tag{33}$$

In particular, as $\epsilon_1,\epsilon_2\to 0$, $B_\lambda^*A_\lambda^*$ converges in Frobenius norm to its orthogonal-case form $Z_\lambda^*(0,I)X_{\mathrm{ft}}^\top$.

Remark D.2 (Bounding $\|S\|_F$ in ϵ_2)**.** *The bound* (33) *is stated directly in terms of $\|S\|_F = \|W_{\mathrm{pre}}X_{\mathrm{ft}}\|_F$. To translate this into a bound in ϵ_2, one needs a lower bound on $\sigma_{\min}(X_{\mathrm{pre}})$. By Proposition G.1 and Assumption 2.4, each row w_k of W_{pre} admits the representation $w_k = X_{\mathrm{pre}}\beta_k$ with $\|\beta_k\|_2 \le \|w_k\|_2/\sigma_{\min}(X_{\mathrm{pre}})$. Combined with Assumption 2.1(c),*

$$|w_k^\top \bar{x}_{c,j}| = |\beta_k^\top X_{\mathrm{pre}}^\top \bar{x}_{c,j}| \le \|\beta_k\|_2\,\sqrt{N_{\mathrm{pre}}}\,\epsilon_2 \le \frac{\|w_k\|_2\,\sqrt{N_{\mathrm{pre}}}\,\epsilon_2}{\sigma_{\min}(X_{\mathrm{pre}})}\,, \tag{34}$$

hence $\|S\|_F \le \sqrt{\bar{K}n\,N_{\mathrm{pre}}}\,\|W_{\mathrm{pre}}\|_F\,\epsilon_2/\sigma_{\min}(X_{\mathrm{pre}})$. In particular, $\|S\|_F\to 0$ as $\epsilon_2\to 0$ whenever $\sigma_{\min}(X_{\mathrm{pre}})$ is bounded below by a positive constant. The cleanest specialization is the orthonormal case $X_{\mathrm{pre}}^\top X_{\mathrm{pre}} = I_{N_{\mathrm{pre}}}$ (so $\sigma_{\min}(X_{\mathrm{pre}})=1$), which yields

$$\|S\|_F \le \sqrt{\bar{K}n\,N_{\mathrm{pre}}}\,\|W_{\mathrm{pre}}\|_F\,\epsilon_2\,. \tag{35}$$

The same conclusion holds, up to a multiplicative conditioning constant, whenever $\sigma_{\min}(X_{\mathrm{pre}})$ can be lower-bounded (e.g., near-orthogonal pre-training features, or unit-norm pre-training features with bounded condition number). We highlight that we do not claim (35) *is tight. The goal of Remark D.2 is to show that $\|S\|_F$ can be controlled by ϵ_2 under mild assumptions, so that the right-hand side of* (33) *goes to zero as $\epsilon_2\to 0$.*

Proof of Theorem D.2. We first reformulate the optimization problem in Z as it was done in the previous proof. By the same change of variable used in the proof of Lemma D.1, decompose $A = \tilde{A}Q^\top + A_\perp$ where $\tilde{A} := AQ \in \mathbb{R}^{r\times\bar{K}n}$ and $A_\perp := A(I_d - QQ^\top)$. Then $Ax_{c,j} = \tilde{A}Re_{(c-1)n+j}$ and (5) can be rewritten as

$$\mathcal{L}(B,A) = \frac{1}{\bar{K}n}\sum_{c,j}\mathcal{L}_{\mathrm{CE}}\big(y_c,(B\tilde{A}R+S)e_{(c-1)n+j}\big) + \frac{\lambda}{2}\big(\|B\|_F^2 + \|\tilde{A}\|_F^2 + \|A_\perp\|_F^2\big)\,. \tag{36}$$

The component $A_\perp$ contributes only to the regularizer and is therefore zero at any minimizer. Setting $Z = B\tilde{A}$ and applying the variational form of the nuclear norm (Lemma G.2), the LoRA problem becomes the rank-constrained minimization $\min_{Z\in\mathbb{R}^{K\times\bar{K}n},\,\mathrm{rank}(Z)\le r}\Phi(Z;S,R)$, where $\Phi(Z;\bar{S},\bar{R}) := L(Z;\bar{S},\bar{R}) + \lambda\|Z\|_*$. Let $Z_\lambda^*(S,R)$ denote the unique unconstrained minimizer of $\Phi(\cdot;S,R)$ over $\mathbb{R}^{K\times\bar{K}n}$ (Lemma G.6). We first establish a perturbation bound for this unconstrained problem, then verify that its minimizer satisfies the rank constraint.

Then, we provide a spectral bound on R. By Assumption 2.1, $G_{ii} = \|\bar{x}_i\|^2 = 1$ and $|G_{ij}| = |\bar{x}_i^\top\bar{x}_j| \le \epsilon_1$ for $i\neq j$. By Gershgorin's circle theorem,

$$\|G-I\|_{\mathrm{op}} \le \max_i\sum_{j\neq i}|G_{ij}| \le (\bar{K}n-1)\,\epsilon_1 = \delta\,. \tag{37}$$

Hence every eigenvalue of G lies in $[1-\delta, 1+\delta]$, so $\sigma_i(R) = \sqrt{\sigma_i(G)} \in [\sqrt{1-\delta}, \sqrt{1+\delta}]$, which gives $\sigma_{\min}(R) \geq \sqrt{1-\delta}$. Using $|\sqrt{x}-1| \leq |x-1|$ for $x \geq 0$,

$$\|R - I\|_{\mathrm{op}} = \max_i |\sqrt{\sigma_i(G)} - 1| \leq \max_i |\sigma_i(G) - 1| = \|G - I\|_{\mathrm{op}} \leq \delta\,. \tag{38}$$

Based on the previous spectral bound, we can now bound the deviation of $Z^*_\lambda(S,R)$ from $Z^*_\lambda(0,I)$, which will lead to the desired bound on $\|B^*_\lambda A^*_\lambda - Z^*_\lambda(0,I) X_{\mathrm{ft}}^\top\|_F$. Let $L(Z;\bar{S},\bar{R}) := \Phi(Z;\bar{S},\bar{R}) - \lambda\|Z\|_*$ be the smooth (cross-entropy) part of the objective, and write $Z_1 := Z^*_\lambda(S,R)$, $Z_0 := Z^*_\lambda(0,I)$. The first-order optimality conditions read

$$-\nabla_Z L(Z_1; S, R) \in \lambda\, \partial\|Z_1\|_*\,, \qquad -\nabla_Z L(Z_0; 0, I) \in \lambda\, \partial\|Z_0\|_*\,. \tag{39}$$

Subdifferentials of convex functions are monotone, so $\langle G_1 - G_0,\, Z_1 - Z_0\rangle \geq 0$ for any $G_i \in \partial\|Z_i\|_*$. Substituting,

$$\langle \nabla L(Z_1; S, R) - \nabla L(Z_0; 0, I),\, Z_1 - Z_0\rangle \leq 0\,. \tag{40}$$

We work under the preliminary bound $\|S\|_F \leq 1$ and will choose $s_0 \leq 1$ below. By Lemma G.6, both Z_0 and Z_1 lie in the subspace $\{Z : \mathbf{1}_K^\top Z = 0\}$. We next establish an a priori F-norm bound to enable application of local strong convexity. The unperturbed bound $\|Z_0\|_* \leq \log K/\lambda$ is already given in the proof of Lemma D.1 (using $\Phi(0;0,I) = \log K$). The same argument extends to the perturbed problem: by optimality of Z_1 and feasibility of $Z = 0$,

$$\lambda\,\|Z_1\|_F \leq \lambda\,\|Z_1\|_* \leq \Phi(Z_1; S, R) \leq \Phi(0; S, R) = L(0; S, R) \leq \log K + 2\|S\|_F\,, \tag{41}$$

where the last inequality uses $\mathcal{L}_{\mathrm{CE}}(y_c, z) = -z_c + \log\sum_i e^{z_i} \leq \log K + 2\|z\|_\infty$ together with $\|S e_{(c-1)n+j}\|_\infty \leq \|S\|_F$. Define

$$R_0 := \frac{\log K + 2}{\lambda}\,, \qquad \mathcal{B} := \left\{Z \in \mathbb{R}^{K \times \bar{K}n} : \|Z\|_F \leq R_0,\, \mathbf{1}_K^\top Z = 0\right\},$$

so that $\{Z_0, Z_1\} \subset \mathcal{B}$. By Lemma G.7 applied with radius R_0, $L(\cdot; 0, I)$ is μ_0-strongly convex on $\mathcal{B}$ in directions $\Delta_Z \in \{\Delta_Z : \mathbf{1}_K^\top \Delta_Z = 0\}$, where $\mu_0 = \mu_0(R_0) > 0$. Consequently

$$\langle \nabla L(Z_1; 0, I) - \nabla L(Z_0; 0, I),\, Z_1 - Z_0\rangle \geq \mu_0 \|Z_1 - Z_0\|_F^2\,. \tag{42}$$

Subtracting from (40),

$$\begin{aligned} \mu_0\|Z_1 - Z_0\|_F^2 &\leq \langle \nabla L(Z_1; 0, I) - \nabla L(Z_1; S, R),\, Z_1 - Z_0\rangle & (43)\\ &\leq \|\nabla L(Z_1; 0, I) - \nabla L(Z_1; S, R)\|_F\, \|Z_1 - Z_0\|_F\,, & (44)\end{aligned}$$

hence

$$\mu_0\,\|Z_1 - Z_0\|_F \leq \|\nabla L(Z_1; 0, I) - \nabla L(Z_1; S, R)\|_F\,. \tag{45}$$

It remains to bound the right-hand side of (45). Direct calculation gives

$$\nabla L(Z; \bar{S}, \bar{R}) = \frac{1}{\bar{K}n} \sum_{c,j} \sigma_{c,j}(Z; \bar{S}, \bar{R})\, (\bar{R}\, e_{(c-1)n+j})^\top\,, \tag{46}$$

where $\sigma_{c,j}(Z; \bar{S}, \bar{R}) := \mathrm{softmax}\big((Z\bar{R} + \bar{S}) e_{(c-1)n+j}\big) - y_c \in \mathbb{R}^K$. Splitting,

$$\begin{aligned} \nabla L(Z; S, R) - \nabla L(Z; 0, I) = \frac{1}{\bar{K}n} \sum_{c,j} \Big[&\big(\sigma_{c,j}(Z; S, R) - \sigma_{c,j}(Z; 0, I)\big)\, (R\, e_{(c-1)n+j})^\top & (47)\\ &+ \sigma_{c,j}(Z; 0, I)\, ((R - I)\, e_{(c-1)n+j})^\top \Big]\,. & (48)\end{aligned}$$

The softmax map is 1-Lipschitz, so

$$\begin{aligned} \|\sigma_{c,j}(Z; S, R) - \sigma_{c,j}(Z; 0, I)\|_2 &\leq \|((ZR + S) - Z) e_{(c-1)n+j}\|_2 \\ &\leq \|Z\|_F\, \|R - I\|_{\mathrm{op}} + \|S\, e_{(c-1)n+j}\|_2\,, \end{aligned} \tag{49}$$

and $\|\sigma_{c,j}(Z;0,I)\|_2 \le \sqrt{2}$ since $\sigma_{c,j}$ is the difference of a probability vector and a one-hot label, both in $[0,1]^K$. Combining via the triangle inequality and Cauchy–Schwarz (using $\sum_{c,j}\|Se_{(c-1)n+j}\|_2 \le \sqrt{\bar{K}n}\,\|S\|_F$ and $\sum_{c,j}\|(R-I)e_{(c-1)n+j}\|_2 \le \sqrt{\bar{K}n}\,\|R-I\|_F \le \bar{K}n\,\delta$),

$$\|\nabla L(Z;S,R) - \nabla L(Z;0,I)\|_F \le (1+\delta)\Big(\delta\|Z\|_F + \frac{\|S\|_F}{\sqrt{\bar{K}n}}\Big) + \sqrt{2}\,\delta\,. \tag{50}$$

Evaluating at $Z = Z_1$, using $\|Z_1\|_F \le \|Z_0\|_F + \|Z_1 - Z_0\|_F$, and substituting into (45):

$$\big(\mu_0 - (1+\delta)\delta\big)\,\|Z_1 - Z_0\|_F \le (1+\delta)\Big(\delta\,\|Z_0\|_F + \frac{\|S\|_F}{\sqrt{\bar{K}n}}\Big) + \sqrt{2}\,\delta\,. \tag{51}$$

Set $\eta_0 := \min(1/2, \mu_0/3) \in (0, 1/2]$. For any $\delta \le \eta_0$ we have $1+\delta \le 3/2$ and

$$(1+\delta)\delta \le \tfrac{3}{2}\delta \le \tfrac{3}{2}\cdot\tfrac{\mu_0}{3} = \tfrac{\mu_0}{2}\,,$$

so $\mu_0 - (1+\delta)\delta \ge \mu_0/2 > 0$. Dividing through and using $1+\delta \le 3/2$, $\sqrt{2}\delta \le \sqrt{2}\delta\,(1+\|Z_0\|_F)$,

$$\|Z_1 - Z_0\|_F \le \frac{C_1}{\mu_0}\left(\|S\|_F + \delta\,\big(1 + \|Z_0\|_F\big)\right), \tag{52}$$

where C_1 is an absolute constant absorbing the factors 2, $3/\sqrt{\bar{K}n}$, $\sqrt{2}$ from the previous display.

By Lemma D.5 below, there exist thresholds $\delta_0' \le \eta_0$ and $s_0 \le 1$ such that $\operatorname{rank}(Z_1) \le \bar{K} \le r$ whenever $\delta \le \delta_0'$ and $\|S\|_F \le s_0$. Choose the theorem's threshold $\delta_0 := \delta_0'$. Under the joint conditions in the statement, the rank constraint is therefore non-binding, so $B^*\bar{A}^* = Z_1$ and $B^*A^* = Z_1Q^\top$.

Finally, we bound $\|B_\lambda^* A_\lambda^* - Z_0\,X_{\mathrm{ft}}^\top\|_F$. From the reformulation above, $B_\lambda^* A_\lambda^* = Z_1\,Q^\top$. Using $X_{\mathrm{ft}} = QR$, hence $X_{\mathrm{ft}}^\top = RQ^\top$, we obtain $Z_0\,X_{\mathrm{ft}}^\top = Z_0\,R\,Q^\top$, so

$$\begin{aligned}
\big\|B_\lambda^* A_\lambda^* - Z_0\,X_{\mathrm{ft}}^\top\big\|_F &= \big\|(Z_1 - Z_0R)\,Q^\top\big\|_F \\
&= \big\|Z_1 - Z_0R\big\|_F \qquad\qquad (\text{since } Q^\top Q = I) \\
&\le \|Z_1 - Z_0\|_F + \big\|Z_0(I-R)\big\|_F \\
&\le \|Z_1 - Z_0\|_F + \delta\,\|Z_0\|_F\,.
\end{aligned} \tag{53}$$

Combining with (52) and using $\delta\,\|Z_0\|_F \le \delta\,(1+\|Z_0\|_F)$,

$$\big\|B_\lambda^* A_\lambda^* - Z_0\,X_{\mathrm{ft}}^\top\big\|_F \le \frac{C_1}{\mu_0}\big(\|S\|_F + \delta\,(1+\|Z_0\|_F)\big) + \delta\,(1+\|Z_0\|_F) \tag{54}$$

$$\le \frac{C_1+\mu_0}{\mu_0}\big(\|S\|_F + \delta\,(1+\|Z_0\|_F)\big)\,. \tag{55}$$

Setting $C_0 := C_1 + \mu_0$ (which depends only on $\lambda, K, \bar{K}, n, \|Z_0\|_*$, as required) gives (33). □

Lemma D.5 (Rank stability of the unconstrained minimizer)**.** *Adopt the notation from the proof of Theorem D.2. Let $Z_\lambda^*(\bar{S},\bar{R})$ denote the unique unconstrained minimizer of $\Phi(\cdot;\bar{S},\bar{R}) = L(\cdot;\bar{S},\bar{R}) + \lambda\|\cdot\|_*$ over $\mathbb{R}^{K\times\bar{K}n}$ (Lemma G.6), and write $Z_0 := Z_\lambda^*(0,I)$. There exist constants $\delta_0' \in (0,\eta_0]$ and $s_0 \in (0,1]$, depending on $\lambda, K, \bar{K}, n, \mu_0, C_1, \|Z_0\|_*$ but not on ϵ_1, ϵ_2, such that whenever $\delta \le \delta_0'$ and $\|S\|_F \le s_0$,*

$$\operatorname{rank}\big(Z_\lambda^*(S,R)\big) \le \bar{K}\,.$$

Proof. Write $Z_1 := Z_\lambda^*(S,R)$ and $G_i := -\nabla L(Z_i;\bar{S}_i,\bar{R}_i)/\lambda$ for $i \in \{0,1\}$, with $(\bar{S}_0,\bar{R}_0) = (0,I)$ and $(\bar{S}_1,\bar{R}_1) = (S,R)$.

By the within-class permutation argument in the proof of Lemma D.2, $Z_0 = V \otimes \mathbf{1}_n^\top$ for some $V \in \mathbb{R}^{K\times\bar{K}}$.

Consequently, $\sigma(Z_0\,e_{(c-1)n+j}) = \sigma(V_{:,c}) =: p_c$ depends only on c. Letting $D := \sum_{c=1}^{\bar{K}}(p_c - y_c)\,e_c^\top \in \mathbb{R}^{K\times\bar{K}}$,

$$\nabla L(Z_0;0,I) = \frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}(p_c - y_c)\sum_{j=1}^{n} e_{(c-1)n+j}^\top = \frac{1}{\bar{K}n}\,D\otimes\mathbf{1}_n^\top\,, \tag{56}$$

so $\operatorname{rank}\big(\nabla L(Z_0;0,I)\big) \le \operatorname{rank}(D) \le \bar{K}$, and

$$\sigma_i(G_0) = 0 \quad \text{for all } i > \bar{K}\,. \tag{57}$$

Then, we study the first-order optimality condition for the unconstrained convex problem reads $G_1 \in \partial\|Z_1\|_*$. Letting $Z_1 = U_1\Sigma_1 V_1^\top$ be a compact SVD of rank $r_1 := \operatorname{rank}(Z_1)$, the standard subdifferential characterization of the nuclear norm [Watson, 1992] gives $G_1 = U_1V_1^\top + W$ with $U_1^\top W = 0$, $WV_1 = 0$, and $\|W\|_{\mathrm{op}} \le 1$. Since $U_1V_1^\top$ and W have orthogonal column and row spaces, their singular values combine: $\sigma_i(G_1) = 1$ for $i \le r_1$, and $\sigma_i(G_1) = \sigma_{i-r_1}(W) \le 1$ for $i > r_1$. Applying Weyl's inequality for singular values to G_1 and G_0, together with (57),

$$\sigma_{\bar{K}+1}(G_1) \le \sigma_{\bar{K}+1}(G_0) + \|G_1 - G_0\|_{\mathrm{op}} = \|G_1 - G_0\|_{\mathrm{op}} \le \|G_1 - G_0\|_F\,. \tag{58}$$

Next, based on the previous analysis, we bound the difference between the gradients. Decompose

$$\nabla L(Z_1;S,R) - \nabla L(Z_0;0,I) = \underbrace{\nabla L(Z_1;S,R) - \nabla L(Z_1;0,I)}_{=:(\mathrm{I})} + \underbrace{\nabla L(Z_1;0,I) - \nabla L(Z_0;0,I)}_{=:(\mathrm{II})}\,.$$

The bound (59) on (I) follows from the gradient calculation in the proof of Theorem D.2 (the display preceding (52)): for $\delta \le \eta_0$,

$$\|(\mathrm{I})\|_F \le (1+\delta)\Big(\delta\|Z_1\|_F + \frac{\|S\|_F}{\sqrt{\bar{K}n}}\Big) + \sqrt{2}\,\delta\,. \tag{59}$$

For (II), the softmax map $\sigma : \mathbb{R}^K \to \Delta^{K-1}$ has Jacobian $\operatorname{diag}(p) - pp^\top$ with operator norm at most 1, and is therefore 1-Lipschitz in ℓ_2. Since $\{e_{(c-1)n+j}\}_{c,j}$ is an orthonormal basis of $\mathbb{R}^{\bar{K}n}$,

$$\|(\mathrm{II})\|_F^2 = \frac{1}{(\bar{K}n)^2}\sum_{c,j}\big\|\sigma(Z_1e_{(c-1)n+j}) - \sigma(Z_0e_{(c-1)n+j})\big\|_2^2 \le \frac{\|Z_1 - Z_0\|_F^2}{(\bar{K}n)^2}\,. \tag{60}$$

By (52), $\|Z_1 - Z_0\|_F \le \frac{C_1}{\mu_0}\big(\|S\|_F + \delta(1+\|Z_0\|_F)\big)$. Combined with $\|Z_1\|_F \le \|Z_0\|_F + \|Z_1 - Z_0\|_F$ and restricting to $\delta \le \eta_0$ and $\|S\|_F \le 1$, there exists a constant

$$C_2 := (1+\eta_0)\Big(\|Z_0\|_F + \tfrac{C_1}{\mu_0}(1+\|Z_0\|_F) + \tfrac{1}{\sqrt{\bar{K}n}}\Big) + \sqrt{2} + \tfrac{C_1(1+\|Z_0\|_F)}{\mu_0\bar{K}n}\,,$$

depending only on $\lambda, K, \bar{K}, n, \mu_0, C_1, \|Z_0\|_*$, such that

$$\|\nabla L(Z_1;S,R) - \nabla L(Z_0;0,I)\|_F \le C_2\,(\delta + \|S\|_F)\,. \tag{61}$$

Finally, we use the above bounds to establish a condition under which the rank of Z_1 is bounded. Set

$$\delta_0' := \min\big(\eta_0,\ \tfrac{\lambda}{4C_2}\big)\,, \qquad s_0 := \min\big(1,\ \tfrac{\lambda}{4C_2}\big)\,.$$

For $\delta \le \delta_0'$ and $\|S\|_F \le s_0$,

$$\|G_1 - G_0\|_F = \frac{1}{\lambda}\,\|\nabla L(Z_1;S,R) - \nabla L(Z_0;0,I)\|_F \le \frac{C_2(\delta_0' + s_0)}{\lambda} \le \frac{1}{2} < 1\,.$$

Combined with (58), $\sigma_{\bar{K}+1}(G_1) < 1$. Since at most $\bar{K}$ singular values of G_1 can equal 1 by Step 2, we conclude $\operatorname{rank}(Z_1) \le \bar{K}$. □

E Proof of Theorem 3.2 and Corollary 3.1

In this section, we first present the proofs of Theorem 3.2 and Corollary 3.1. Then, under additional smallness assumptions on ϵ_1, ϵ_2, we show that the same conclusions still hold, with additional error terms that depend on ϵ_1, ϵ_2.

E.1 Proof of Theorem 3.2

For convenience, we will first define the following notations. Let $W_{\text{LoRA}}^{\lambda} = W_{\text{pre}} + B_{\lambda}^{*}A_{\lambda}^{*}$, and define the following shorthand for margins as: $\gamma_{\text{pre}} = \gamma(W_{\text{pre}}; \mathcal{D}_{\text{pre}})$, $\gamma_{\text{ft},\lambda} = \gamma(B_{\lambda}^{*}A_{\lambda}^{*}; \mathcal{D}_{\text{ft}})$, and $\gamma_{\lambda} = \gamma(W_{\text{LoRA}}^{\lambda}; \mathcal{D}_{\text{pre}} \cup \mathcal{D}_{\text{ft}})$. Additionally, given an $\bar{K}$-class max-margin classifier $W_{\text{ft}}^{*} \in \mathbb{R}^{\bar{K}\times d}$ on the fine-tuning data, let $\gamma_{\text{ft}}^{\text{pad}} = \gamma(([W_{\text{ft}}^{*}; \mathbf{0}_{(K-\bar{K})\times d}]); \mathcal{D}_{\text{ft}})$. When $\bar{K} < K$ this is in general a lower bound on $\max_{W\in\mathbb{R}^{K\times d}} \gamma(W; \mathcal{D}_{\text{ft}})$, with equality iff $\bar{K} = K$. In the balanced case $\bar{K} = K$, $\gamma_{\text{ft}}^{\text{pad}} = \gamma_{\text{ft}}^{*}$ as defined in the main theorem.

With these definitions in place, we restate Theorem 3.2 in expanded form covering the general $\bar{K} \le K$ setting. The expanded version specifies the closed-form values of Θ_{λ} and $\gamma_{\text{ft},\lambda}$ in three regimes when $\bar{K} < K$. In the balanced case $\bar{K} = K$, the intermediate and low-penalty regimes coincide in their ETF geometry and margin formulas, recovering the two-regime statement of Theorem 3.2. More precisely, $\lambda_{\text{crit}} = 0$, so the intermediate formula holds for every $0 < \lambda < 1/(K\sqrt{n})$, and the limit $\lambda \to 0^{+}$ retains the same ETF geometry.

Theorem E.1 (Detailed version of Theorem 3.2)**.** *Adopt the setup of Theorem 3.1 and let* $\Theta_{\lambda} := \|B_{\lambda}^{*}A_{\lambda}^{*}\|_{F}$*. The normalized margins of* $W_{\text{LoRA}}^{\lambda}$ *on the pre-training and fine-tuning data are*

$$\gamma\big(W_{\text{LoRA}}^{\lambda}; \mathcal{D}_{\text{pre}}\big) = \gamma_{\text{pre}} \frac{\rho_{\text{pre}}}{\sqrt{\Theta_{\lambda}^{2} + \rho_{\text{pre}}^{2}}}, \quad \gamma\big(W_{\text{LoRA}}^{\lambda}; \mathcal{D}_{\text{ft}}\big) = \gamma_{\text{ft},\lambda} \frac{\Theta_{\lambda}}{\sqrt{\Theta_{\lambda}^{2} + \rho_{\text{pre}}^{2}}}, \tag{62}$$

which yield the closed-form $\text{Adapt}(\lambda)$ *and* $\text{Forget}(\lambda)$ *in* (13) *via* $\text{Adapt}(\lambda) = \gamma(W_{\text{LoRA}}^{\lambda}; \mathcal{D}_{\text{ft}})$ *(since* $\gamma(W_{\text{pre}}; \mathcal{D}_{\text{ft}}) = 0$ *in the strict-orthogonal regime) and* $\text{Forget}(\lambda) = \gamma_{\text{pre}} - \gamma(W_{\text{LoRA}}^{\lambda}; \mathcal{D}_{\text{pre}})$*. Moreover,* Θ_{λ} *and* $\gamma_{\text{ft},\lambda}$ *take the following values:*

(i) ***High–penalty regime*** $\lambda \ge \frac{1}{\bar{K}\sqrt{n}}$*:* $\Theta_{\lambda} = \gamma_{\text{ft},\lambda} = 0$*, and hence* $\text{Adapt}(\lambda) = \text{Forget}(\lambda) = 0$*.*

(ii) ***Intermediate regime*** $\lambda_{\text{crit}} < \lambda < \frac{1}{K\sqrt{n}}$*:* $\Theta_{\lambda} = \frac{\rho_{\lambda}}{\sqrt{\bar{K}-1}}$*,* $\gamma_{\text{ft},\lambda} = \gamma_{\text{ft}}^{\text{pad}}$*, and* Θ_{λ} *is a strictly decreasing function of* λ*. Both* $\text{Adapt}(\lambda)$ *and* $\text{Forget}(\lambda)$ *are therefore strictly decreasing in* λ *on this regime.*

(iii) ***Low–penalty regime*** $0 < \lambda \le \lambda_{\text{crit}}$*:*

$$\Theta_{\lambda} = \sqrt{n\bar{K}a_{\lambda}^{2} + n\bar{K}(\bar{K}-1)b_{\lambda}^{2} + n\bar{K}(K-\bar{K})c_{\lambda}^{2}}, \quad \gamma_{\text{ft},\lambda} = \frac{a_{\lambda} + c_{\lambda}}{\Theta_{\lambda}}.$$

Remark E.1. ***Full fine-tuning as an alternative method.*** *While our work focuses on LoRA, it is instructive to briefly contrast it with the more conventional strategy of full fine-tuning. Full fine-tuning updates all parameters of the pretrained model, which is considerably less efficient in both computation and storage, whereas LoRA achieves adaptation through a compact low-rank parameterization. In the linear classification setting, the two approaches correspond to the following regularized objectives:*

$$\min_{B,A} \; L(W_{\text{pre}} + BA; \mathcal{D}_{\text{ft}}) + \tfrac{\lambda}{2}\big(\|B\|_{F}^{2} + \|A\|_{F}^{2}\big), \qquad \textit{LoRA objective}, \tag{63}$$

$$\min_{W} \; L(W; \mathcal{D}_{\text{ft}}) + \tfrac{\lambda}{2}\|W - W_{\text{pre}}\|_{F}^{2}, \qquad \textit{Full fine-tuning objective}. \tag{64}$$

Beyond efficiency, the two formulations induce fundamentally different (implicit) biases. By exploiting the variational form of the nuclear norm [Recht et al., 2010], one can show that for any solution (B^{*}, A^{*}) *of* (63)*, the induced update* $W_{\text{LoRA}}^{\lambda}$ *is equivalently the solution of*

$$\min_{W, \text{rank}(W - W_{\text{pre}}) \le r} \; L(W; \mathcal{D}_{\text{ft}}) + \lambda\|W - W_{\text{pre}}\|_{*}, \qquad \textit{Implicit bias of LoRA fine-tuning}, \tag{65}$$

which highlights LoRA's connection to nuclear-norm regularization under rank constraints. We refer the readers to Jang et al. [2024] for the proof of the argument. Understanding the consequences of this distinction between nuclear norm regularization and Frobenius norm regularization remains an open direction for future work. Importantly, our goal is not to compare LoRA and full fine-tuning, but to rigorously analyze why LoRA exhibits reduced forgetting and to provide principled guidance for selecting regularization parameters in adapter-based fine-tuning.

Proof. Our key observation is that when one computes the margin of $W_{\mathrm{LoRA}}^{\lambda}$ on the pre-training and fine-tuning datasets, only W_{pre} or $B_\lambda^* A_\lambda^*$ is activated. This is because by Theorem 3.1, the product of the optimal LoRA adapters lies in the span of fine-tuning data, and by Proposition G.1 together with Assumption 2.1, W_{pre} lies in the span of pre-training data. Thus, under Assumption 2.1, $B_\lambda^* A_\lambda^* x = 0$ if $x \in \mathcal{D}_{\mathrm{pre}}$, and $W_{\mathrm{pre}} x = 0$ if $x \in \mathcal{D}_{\mathrm{ft}}$. In particular, $\gamma(W_{\mathrm{pre}}; \mathcal{D}_{\mathrm{ft}}) = 0$, so $\mathrm{Adapt}(\lambda) = \gamma(W_{\mathrm{LoRA}}^{\lambda}; \mathcal{D}_{\mathrm{ft}})$. Before proving that the margins take the structures shown in Theorem E.1, we introduce the following notation. Let $W_{\mathrm{pre}} = [w_1, \ldots, w_K]^\top$ and $B_\lambda^* A_\lambda^* = [\delta w_{\lambda,1}, \cdots, \delta w_{\lambda,K}]^\top$.

Now, when we compute the margin of $W_{\mathrm{LoRA}}^{\lambda}$ on pre-training data, one has

$$\begin{aligned}
\gamma(W_{\mathrm{LoRA}}^{\lambda}; \mathcal{D}_{\mathrm{pre}}) &= \frac{\min_{(x_{i,c}, y_c) \in \mathcal{D}_{\mathrm{pre}}} (w_c + \delta_{\lambda,c})^\top x - \max_{i \neq y} (w_i + \delta_{\lambda,i})^\top x}{\|W_{\mathrm{LoRA}}^{\lambda}\|_F} \\
&= \frac{\min_{(x_{i,c}, y_c) \in \mathcal{D}_{\mathrm{pre}}} w_c^\top x - \max_{i \neq y} w_i^\top x}{\|W_{\mathrm{pre}}\|_F} \cdot \frac{\|W_{\mathrm{pre}}\|_F}{\|W_{\mathrm{LoRA}}^{\lambda}\|_F} \\
&= \gamma_{\mathrm{pre}} \cdot \frac{\rho_{\mathrm{pre}}}{\sqrt{\rho_{\mathrm{pre}}^2 + \|B_\lambda^* A_\lambda^*\|_F^2}} \,.
\end{aligned}$$

Then, for the fine-tuning data, one has

$$\begin{aligned}
\gamma(W_{\mathrm{LoRA}}^{\lambda}; \mathcal{D}_{\mathrm{ft}}) &= \frac{\min_{(x_{i,c}, y_c) \in \mathcal{D}_{\mathrm{ft}}} (w_c + \delta_{\lambda,c})^\top x - \max_{i \neq y} (w_i + \delta_{\lambda,i})^\top x}{\|W_{\mathrm{LoRA}}^{\lambda}\|_F} \\
&= \frac{C(B_\lambda^* A_\lambda^*; \mathcal{D}_{\mathrm{ft}})}{\sqrt{\rho_{\mathrm{pre}}^2 + \|B_\lambda^* A_\lambda^*\|_F^2}} \,, \\
&= \gamma_{\mathrm{ft},\lambda} \cdot \frac{\|B_\lambda^* A_\lambda^*\|_F}{\sqrt{\rho_{\mathrm{pre}}^2 + \|B_\lambda^* A_\lambda^*\|_F^2}} \,.
\end{aligned}$$

The last equality also holds when $B_\lambda^* A_\lambda^* = 0$ by the convention $\gamma(0; \mathcal{D}) = 0$. These are exactly the formulas in (62) with $\Theta_\lambda = \|B_\lambda^* A_\lambda^*\|_F$, and substituting them into the definitions of $\mathrm{Adapt}(\lambda)$ and $\mathrm{Forget}(\lambda)$ recovers (13). It remains to compute $\|B_\lambda^* A_\lambda^*\|_F$ in each regime.

High-penalty regime. In this regime, $B_\lambda^* A_\lambda^* = 0$, so $\|B_\lambda^* A_\lambda^*\|_F = 0$.

Intermediate regime. In the intermediate regime, one has

$$B_\lambda^* A_\lambda^* \;=\; \frac{\rho_\lambda}{\sqrt{\bar{K} n (\bar{K} - 1)}} \left(\begin{pmatrix} M_{\bar{K}} \\ \mathbf{0}_{(K-\bar{K}) \times \bar{K}} \end{pmatrix} \otimes \mathbf{1}_n^\top \right) X_{\mathrm{ft}}^\top \,, \tag{66}$$

and one can compute its norm as follows

$$\begin{aligned}
\|B_\lambda^* A_\lambda^*\|_F^2 &= \frac{\rho_\lambda^2}{\bar{K} n (\bar{K} - 1)} \cdot \frac{\bar{K}}{\bar{K} - 1} \left\| \left((\boldsymbol{I}_{\bar{K}} - \frac{1}{\bar{K}} \mathbf{1}_{\bar{K}} 1_{\bar{K}}^\top) \right) \otimes \mathbf{1}_n^\top X_{\mathrm{ft}}^\top \right\|_F^2 \\
&= \frac{\rho_\lambda^2}{\bar{K} n (\bar{K} - 1)} \cdot \frac{\bar{K}}{\bar{K} - 1} \cdot \bar{K} \left(n \cdot (\frac{\bar{K} - 1}{\bar{K}})^2 + \frac{(\bar{K} - 1) n}{\bar{K}^2} \right) \\
&= \frac{\rho_\lambda^2}{\bar{K} - 1} \,.
\end{aligned}$$

Thus, $\Theta_\lambda = \|B_\lambda^* A_\lambda^*\|_F = \frac{\rho_\lambda}{\sqrt{\bar{K} - 1}}$.

Low-penalty regime. In this regime, $B_\lambda^* A_\lambda^*$ takes the following form

$$B_\lambda^* A_\lambda^* = \left(\begin{pmatrix} (a_\lambda + b_\lambda) \mathbf{I}_{\bar{K}} \;-\; b_\lambda \mathbf{1}_{\bar{K}} \mathbf{1}_{\bar{K}}^\top \\ -c_\lambda \mathbf{1}_{K - \bar{K}} \mathbf{1}_{\bar{K}}^\top \end{pmatrix} \otimes 1_n^\top \right) X_{\mathrm{ft}}^\top \,, \tag{67}$$

and its Frobenius norm is

$$\begin{aligned}
\|B_\lambda^* A_\lambda^*\|_F &= \sqrt{n} \left\| \begin{pmatrix} (a_\lambda + b_\lambda) \mathbf{I}_{\bar{K}} \;-\; b_\lambda \mathbf{1}_{\bar{K}} \mathbf{1}_{\bar{K}}^\top \\ -c_\lambda \mathbf{1}_{K - \bar{K}} \mathbf{1}_{\bar{K}}^\top \end{pmatrix} \right\|_F \\
&= \sqrt{n} \cdot \sqrt{\bar{K} a_\lambda^2 + \bar{K} (\bar{K} - 1) b_\lambda^2 + (K - \bar{K}) \bar{K} c_\lambda^2} \,.
\end{aligned}$$

For any sample $\bar{\boldsymbol{x}}_{c,j} \in \mathcal{D}_{\text{ft}}$ in class $c \le \bar{K}$, the score on the correct class is a_λ, the score on any other class $k \le \bar{K}, k \ne c$ is $-b_\lambda$, and the score on any class $k > \bar{K}$ is $-c_\lambda$. Hence the per-sample margin is $a_\lambda - \max(-b_\lambda, -c_\lambda) = a_\lambda + \min(b_\lambda, c_\lambda)$. By the defining equation of x_0 in Lemma D.4, $g_b(x_0)/g_c(x_0) = 1/(\sqrt{K/(K-\bar{K})} + 1) < 1$, so $\log g_b(x_0) < \log g_c(x_0)$. Combined with $b_\lambda = -\log g_b(x_0)$ and $c_\lambda = -\log g_c(x_0)$, this gives $b_\lambda > c_\lambda$, so $\min(b_\lambda, c_\lambda) = c_\lambda$ and the unnormalized fine-tuning margin equals $a_\lambda + c_\lambda$. Dividing by $\Theta_\lambda = \|B_\lambda^* A_\lambda^*\|_F$ yields $\gamma_{\text{ft},\lambda} = (a_\lambda + c_\lambda)/\Theta_\lambda$. □

E.2 Proof of Corollary 3.1

Proof. The first identity in Corollary 3.1,

$$\gamma(W_{\text{LoRA}}^\lambda; \mathcal{D}_{\text{pre}} \cup \mathcal{D}_{\text{ft}}) = \min\{\gamma_{\text{pre}} - \text{Forget}(\lambda),\ \text{Adapt}(\lambda)\},$$

follows from the definition of the union margin together with $\gamma(W_{\text{LoRA}}^\lambda; \mathcal{D}_{\text{pre}}) = \gamma_{\text{pre}} - \text{Forget}(\lambda)$ and $\gamma(W_{\text{LoRA}}^\lambda; \mathcal{D}_{\text{ft}}) = \text{Adapt}(\lambda)$ from Theorem 3.2. It remains to maximize this minimum over λ.

We work in the balanced setting $\bar{K} = K$ of Corollary 3.1. In the high-penalty regime ($\lambda \ge 1/(K\sqrt{n})$), $\Theta_\lambda = 0$ and so $\text{Adapt}(\lambda) = 0$, which gives a union margin of 0. Every $0 < \lambda < 1/(K\sqrt{n})$ gives a positive union margin, so the optimum lies in this interval.

On this interval, $\gamma_{\text{ft},\lambda} = \gamma_{\text{ft}}^*$ is constant. By the balanced-case solution in Appendix D, Θ_λ decreases continuously and strictly from $+\infty$ to 0 as λ increases from 0 to $1/(K\sqrt{n})$. From (62), the pre-training margin is strictly decreasing in Θ_λ and the fine-tuning margin is strictly increasing in Θ_λ. Their minimum is therefore uniquely maximized where they are equal:

$$\gamma_{\text{pre}} \frac{\rho_{\text{pre}}}{\sqrt{\Theta_\lambda^2 + \rho_{\text{pre}}^2}} = \gamma_{\text{ft}}^* \frac{\Theta_\lambda}{\sqrt{\Theta_\lambda^2 + \rho_{\text{pre}}^2}} \iff \Theta_\lambda = \Theta^* := \frac{\rho_{\text{pre}}\, \gamma_{\text{pre}}}{\gamma_{\text{ft}}^*}.$$

Since $0 < \Theta^* < \infty$, there is a unique $\lambda^* \in (0, 1/(K\sqrt{n}))$ with $\Theta_{\lambda^*} = \Theta^*$. No additional balancing condition is needed. Substituting Θ^* into (62) gives

$$\gamma(W_{\text{LoRA}}^{\lambda^*}; \mathcal{D}_{\text{pre}} \cup \mathcal{D}_{\text{ft}}) = \frac{\gamma_{\text{pre}}\, \gamma_{\text{ft}}^*}{\sqrt{\gamma_{\text{pre}}^2 + (\gamma_{\text{ft}}^*)^2}} = \left(\frac{1}{(\gamma_{\text{ft}}^*)^2} + \frac{1}{\gamma_{\text{pre}}^2}\right)^{-1/2}.$$

□

E.3 Proof of General Version

In this subsection we present the general version of Theorem 3.2: under the near-orthogonal regime in Assumption 2.1, the normalized margins of W_{LoRA}^λ on pre-training, fine-tuning, and combined data remain close to their strict-orthogonal closed-form values from Theorem E.1, with deviations controlled by the perturbation error $E_\lambda + \epsilon_2$ defined below.

Notation. We adopt the notation of Theorem D.2: $G, R, Q, S, \delta, \delta_0, s_0, \mu_0, C_0$, and $Z_\lambda^*(\bar{S}, \bar{R})$, with the key bound (33). Additionally, write

$$\Theta_\lambda^{(0)} := \|Z_\lambda^*(0, I)\|_F, \qquad \gamma_{\text{ft},\lambda}^{(0)} := \gamma_{\text{ft},\lambda}\big|_{\epsilon_1 = \epsilon_2 = 0}, \tag{68}$$

for the strict-orthogonal adapter Frobenius norm and the strict-orthogonal fine-tuning margin from Theorem E.1. (When $X_{\text{ft}}^\top X_{\text{ft}} = I$, $\Theta_\lambda^{(0)} = \|B_\lambda^* A_\lambda^*\|_F$ and $\gamma_{\text{ft},\lambda}^{(0)} = \gamma_{\text{ft},\lambda}$.) Set

$$E_\lambda := \frac{C_0}{\mu_0}\big(\|S\|_F + \delta\,(1 + \|Z_\lambda^*(0, I)\|_F)\big) \quad \text{(the bound in (33))}. \tag{69}$$

The construction $C_0 = C_1 + \mu_0$ in the proof of Theorem D.2 gives $C_0/\mu_0 > 1$, so $\|S\|_F \le E_\lambda$ and $\delta \le E_\lambda$. Therefore, $\|S\|_F = O(\epsilon_2)$, and $E_\lambda \to 0$ as $\epsilon_1, \epsilon_2 \to 0$.

Theorem E.2 (General version of Theorem 3.2)**.** *Adopt the setup of Theorem D.2 (so $\delta \le \delta_0$ and $\|S\|_F \le s_0$). Assume additionally that pre-training samples are unit-norm. There exists a constant*

$C>0$, *depending on* $\lambda, K, \bar{K}, n, \rho_{\mathrm{pre}}, \mu_0, C_0, \|Z^*_\lambda(0,I)\|_*$ *but not on* ϵ_1, ϵ_2, *such that*

$$\left|\gamma\big(W^\lambda_{\mathrm{LoRA}}; \mathcal{D}_{\mathrm{pre}}\big) - \gamma_{\mathrm{pre}}\,\frac{\rho_{\mathrm{pre}}}{\sqrt{(\Theta^{(0)}_\lambda)^2+\rho^2_{\mathrm{pre}}}}\right| \le C\,\big(E_\lambda+\epsilon_2\big)\,, \tag{70}$$

$$\left|\gamma\big(W^\lambda_{\mathrm{LoRA}}; \mathcal{D}_{\mathrm{ft}}\big) - \gamma^{(0)}_{\mathrm{ft},\lambda}\,\frac{\Theta^{(0)}_\lambda}{\sqrt{(\Theta^{(0)}_\lambda)^2+\rho^2_{\mathrm{pre}}}}\right| \le C\,\big(E_\lambda+\epsilon_2\big)\,. \tag{71}$$

The corresponding bound on $\gamma(W^\lambda_{\mathrm{LoRA}}; \mathcal{D}_{\mathrm{pre}}\cup\mathcal{D}_{\mathrm{ft}}) = \min\{\cdot,\cdot\}$ *inherits the same right-hand side. In particular, as* $E_\lambda+\epsilon_2\to 0$, *the three margins converge to their strict-orthogonal closed-form values from Theorem E.1.*

Proof. Throughout the proof, C denotes a positive constant whose value may change line to line, but always depends only on the quantities listed in the theorem statement.

We first establish leakage bounds quantifying how much pre-training and fine-tuning components mix in the near-orthogonal regime. Since $S=W_{\mathrm{pre}}X_{\mathrm{ft}}$, for any fine-tuning sample $\bar{x}_j$ and any row w_k of W_{pre},

$$|w_k^\top \bar{x}_j| = |S_{kj}| \le \|S\|_F \le E_\lambda\,. \tag{72}$$

For the symmetric direction, write $\delta_{\lambda,k} := (B^*_\lambda A^*_\lambda)_k^\top$ for the k-th row of $B^*_\lambda A^*_\lambda$. By the reformulation in the proof of Theorem D.2, $B^*_\lambda A^*_\lambda = Z^*_\lambda(S,R)\,Q^\top$, so for any unit-norm $x\in\mathcal{D}_{\mathrm{pre}}$,

$$|\delta_{\lambda,k}^\top x| \le \|(Z^*_\lambda(S,R))_k\|_2\,\|Q^\top x\|_2\,. \tag{73}$$

Since $\|Q^\top x\|_2 = \|R^{-1}X_{\mathrm{ft}}^\top x\|_2 \le \|R^{-1}\|_{\mathrm{op}}\,\|X_{\mathrm{ft}}^\top x\|_2$, $\|R^{-1}\|_{\mathrm{op}}\le 1/\sqrt{1-\delta}$, and $\|X_{\mathrm{ft}}^\top x\|_2^2 = \sum_l |\bar{x}_l^\top x|^2 \le \bar{K}n\,\epsilon_2^2$ by Assumption 2.1(c), we have $|\delta_{\lambda,k}^\top x| \le \frac{\sqrt{\bar{K}n}\,\epsilon_2}{\sqrt{1-\delta}}\,\|(Z^*_\lambda(S,R))_k\|_2$. Using $\|Z^*_\lambda(S,R)\|_F \le \|Z^*_\lambda(0,I)\|_F + E_\lambda$ from (52),

$$|\delta_{\lambda,k}^\top x| \le C\,(\|Z^*_\lambda(0,I)\|_F + E_\lambda)\,\epsilon_2 \le C\,\epsilon_2\,. \tag{74}$$

Then, we bound the perturbation of the Frobenius norm $\|W^\lambda_{\mathrm{LoRA}}\|_F$ from $\sqrt{\rho^2_{\mathrm{pre}}+(\Theta^{(0)}_\lambda)^2}$. Expanding,

$$\|W^\lambda_{\mathrm{LoRA}}\|_F^2 = \|W_{\mathrm{pre}}\|_F^2 + 2\,\langle W_{\mathrm{pre}}, B^*_\lambda A^*_\lambda\rangle + \|B^*_\lambda A^*_\lambda\|_F^2\,. \tag{75}$$

Each row of $B^*_\lambda A^*_\lambda$ lies in $\mathrm{span}(\mathcal{D}_{\mathrm{ft}})$ (the column space of X_{ft}), so

$$\big|\langle W_{\mathrm{pre}}, B^*_\lambda A^*_\lambda\rangle\big| = \Big|\sum_k w_k^\top \delta_{\lambda,k}\Big| \le \sum_k \|P_{X_{\mathrm{ft}}} w_k\|_2\,\|(B^*_\lambda A^*_\lambda)_k\|_2\,, \tag{76}$$

where $P_{X_{\mathrm{ft}}}$ is the orthogonal projection onto $\mathrm{span}(\mathcal{D}_{\mathrm{ft}})$. The identity $W_{\mathrm{pre}}Q = SR^{-1}$ gives $\|W_{\mathrm{pre}}Q\|_F \le \|S\|_F/\sqrt{1-\delta}$. Thus, Cauchy–Schwarz and $B^*_\lambda A^*_\lambda = Z^*_\lambda(S,R)Q^\top$ yield

$$\big|\langle W_{\mathrm{pre}}, B^*_\lambda A^*_\lambda\rangle\big| \le \frac{\|S\|_F\,\|B^*_\lambda A^*_\lambda\|_F}{\sqrt{1-\delta}} \le C\,(\Theta^{(0)}_\lambda + E_\lambda)\,\|S\|_F\,. \tag{77}$$

Write $\Delta_\lambda := B^*_\lambda A^*_\lambda$. Equation (52) and $Q^\top Q = I$ give $\big|\|\Delta_\lambda\|_F - \Theta^{(0)}_\lambda\big| \le E_\lambda$. Comparing first with $\sqrt{\rho^2_{\mathrm{pre}}+\|\Delta_\lambda\|_F^2}$ and then using this norm bound yields

$$\begin{aligned}\Big|\,\|W^\lambda_{\mathrm{LoRA}}\|_F - \sqrt{\rho^2_{\mathrm{pre}}+(\Theta^{(0)}_\lambda)^2}\,\Big| &\le \frac{2\big|\langle W_{\mathrm{pre}}, \Delta_\lambda\rangle\big|}{\|W^\lambda_{\mathrm{LoRA}}\|_F + \sqrt{\rho^2_{\mathrm{pre}}+\|\Delta_\lambda\|_F^2}} + E_\lambda \\ &\le \frac{2\|S\|_F}{\sqrt{1-\delta}} + E_\lambda \le C\,(E_\lambda+\epsilon_2)\,.\end{aligned} \tag{78}$$

Next, we bound the perturbation of the unnormalized margins $C(W^\lambda_{\mathrm{LoRA}};\cdot)$ on each dataset. For pre-training data $(x,y_c)\in\mathcal{D}_{\mathrm{pre}}$ and $k\neq c$,

$$[(w_c+\delta_{\lambda,c}) - (w_k+\delta_{\lambda,k})]^\top x = (w_c-w_k)^\top x + (\delta_{\lambda,c}-\delta_{\lambda,k})^\top x\,. \tag{79}$$

Taking the minimum over (x, c, k) and combining $\min(w_c - w_k)^\top x = C(W_{\rm pre}; \mathcal{D}_{\rm pre}) = \rho_{\rm pre}\gamma_{\rm pre}$ (definition of $\gamma_{\rm pre}$) with $|\delta_{\lambda,c}^\top x - \delta_{\lambda,k}^\top x| \le 2\,C\epsilon_2$ from (74),

$$\left|C(W_{\rm LoRA}^\lambda; \mathcal{D}_{\rm pre}) - \rho_{\rm pre}\gamma_{\rm pre}\right| \le C\,\epsilon_2\,. \tag{80}$$

For fine-tuning data $(\bar{x}, y_c) \in \mathcal{D}_{\rm ft}$ and $k \ne c$,

$$[(w_c + \delta_{\lambda,c}) - (w_k + \delta_{\lambda,k})]^\top \bar{x} = (\delta_{\lambda,c} - \delta_{\lambda,k})^\top \bar{x} + (w_c - w_k)^\top \bar{x}\,. \tag{81}$$

The unnormalized fine-tuning margin is 2-Lipschitz in the F-norm of the classifier matrix (since for any $W, W' \in \mathbb{R}^{K\times d}$ and any unit-norm $\bar{x}$, $|((W - W')_c - (W - W')_k)^\top \bar{x}| \le 2\|W - W'\|_F$), so by Theorem D.2,

$$\left|C(B_\lambda^* A_\lambda^*; \mathcal{D}_{\rm ft}) - C(Z_\lambda^*(0, I)X_{\rm ft}^\top; \mathcal{D}_{\rm ft})\right| \le 2\,E_\lambda\,. \tag{82}$$

We further compare $C(Z_\lambda^*(0, I)X_{\rm ft}^\top; \mathcal{D}_{\rm ft})$ — the unnormalized margin of the strict-orthogonal classifier evaluated on the actual (near-orthogonal) data — with its strict-orthogonal value $\Theta_\lambda^{(0)}\gamma_{{\rm ft},\lambda}^{(0)}$. For any sample $\bar{x}_{c',j'} \in \mathcal{D}_{\rm ft}$,

$$X_{\rm ft}^\top \bar{x}_{c',j'} = e_{(c'-1)n+j'} + r_{c',j'}\,, \qquad r_{c',j'} := X_{\rm ft}^\top \bar{x}_{c',j'} - e_{(c'-1)n+j'}\,, \tag{83}$$

where the $((c''-1)n + j'')$-th entry of $r_{c',j'}$ is $\bar{x}_{c'',j''}^\top \bar{x}_{c',j'}$ for $(c'', j'') \ne (c', j')$ and 0 otherwise. By Assumption 2.1(b),

$$\|r_{c',j'}\|_2^2 \le \sum_{(c'',j'')\ne(c',j')} \epsilon_1^2 = (\bar{K}n - 1)\,\epsilon_1^2\,, \tag{84}$$

hence $\|r_{c',j'}\|_2 \le \sqrt{\bar{K}n - 1}\,\epsilon_1 = \delta/\sqrt{\bar{K}n - 1}$. For any class $k \in [K]$,

$$\left|(Z_\lambda^*(0, I)X_{\rm ft}^\top \bar{x}_{c',j'})_k - (Z_\lambda^*(0, I))_{k,(c'-1)n+j'}\right| = \left|(Z_\lambda^*(0, I))_k r_{c',j'}\right| \le \|(Z_\lambda^*(0, I))_k\|_2\,\|r_{c',j'}\|_2\,, \tag{85}$$

where $(Z_\lambda^*(0, I))_k$ denotes the k-th row. Since $\max_k \|(Z_\lambda^*(0, I))_k\|_2 \le \|Z_\lambda^*(0, I)\|_F = \Theta_\lambda^{(0)}$, the difference of two-class scores satisfies

$$\begin{aligned}&\left|[(Z_\lambda^*(0, I)X_{\rm ft}^\top \bar{x}_{c',j'})_c - (Z_\lambda^*(0, I)X_{\rm ft}^\top \bar{x}_{c',j'})_k] - [(Z_\lambda^*(0, I))_{c,(c'-1)n+j'} - (Z_\lambda^*(0, I))_{k,(c'-1)n+j'}]\right| \\ \le &\frac{2\,\Theta_\lambda^{(0)}\,\delta}{\sqrt{\bar{K}n - 1}}\end{aligned} \tag{86}$$

for every $\bar{x} \in \mathcal{D}_{\rm ft}$ and every pair $c \ne k$. Setting $c = c'$ and taking the minimum over $c' \in [\bar{K}]$, $j' \in [n]$, and $k \in [K] \setminus \{c'\}$, the 1-Lipschitz property of $\min$ gives

$$\left|C(Z_\lambda^*(0, I)X_{\rm ft}^\top; \mathcal{D}_{\rm ft}) - \min_{c',j',k\ne c'}\left[(Z_\lambda^*(0, I))_{c',(c'-1)n+j'} - (Z_\lambda^*(0, I))_{k,(c'-1)n+j'}\right]\right| \le \frac{2\,\Theta_\lambda^{(0)}\,\delta}{\sqrt{\bar{K}n - 1}}\,. \tag{87}$$

The displayed minimum is precisely the strict-orthogonal unnormalized margin $\Theta_\lambda^{(0)}\,\gamma_{{\rm ft},\lambda}^{(0)}$ (it is the minimum-margin definition applied to $Z_\lambda^*(0, I)$ on a one-hot $\bar{K}n$-dimensional input set, which is what Theorem E.1 characterizes). Hence

$$\left|C(Z_\lambda^*(0, I)X_{\rm ft}^\top; \mathcal{D}_{\rm ft}) - \Theta_\lambda^{(0)}\gamma_{{\rm ft},\lambda}^{(0)}\right| \le \frac{2\,\Theta_\lambda^{(0)}}{\sqrt{\bar{K}n - 1}}\,\delta \le C\,\delta\,. \tag{88}$$

The leakage from the pre-training rows on $\mathcal{D}_{\rm ft}$ contributes at most $2\|S\|_F$ to the per-pair score difference via (72), so taking the min,

$$\left|C(W_{\rm LoRA}^\lambda; \mathcal{D}_{\rm ft}) - C(B_\lambda^* A_\lambda^*; \mathcal{D}_{\rm ft})\right| \le 2\|S\|_F \le 2E_\lambda\,. \tag{89}$$

Combining with (82) and (88) via the triangle inequality (and noting $\delta \le (\mu_0/C_0)\,E_\lambda$ by definition of E_λ, so the δ from (88) is absorbed into $C\,E_\lambda$),

$$\left|C(W_{\rm LoRA}^\lambda; \mathcal{D}_{\rm ft}) - \Theta_\lambda^{(0)}\gamma_{{\rm ft},\lambda}^{(0)}\right| \le C\,(E_\lambda + \epsilon_2)\,. \tag{90}$$

Finally, we combine the bounds. For a dataset of unit-norm inputs, $|\gamma(W;\mathcal{D})| \le \sqrt{2}$ whenever $W \neq 0$. Let $a := \|W^{\lambda}_{\mathrm{LoRA}}\|_F$, $b := \sqrt{\rho^2_{\mathrm{pre}} + (\Theta^{(0)}_{\lambda})^2}$, $u := C(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D})$, and let u_0 be the corresponding strict-orthogonal unnormalized margin. Then

$$\left|\frac{u}{a} - \frac{u_0}{b}\right| = \left|\frac{u-u_0}{b} + \frac{u}{a}\frac{b-a}{b}\right| \le \frac{|u-u_0| + \sqrt{2}\,|a-b|}{b}\,. \tag{91}$$

Since $b \ge \rho_{\mathrm{pre}} > 0$, (78) combined with (80) gives (70) and combined with (90) gives (71). Taking the minimum of the pre-training and fine-tuning margins gives the combined-data bound by the 1-Lipschitz property of $\min$. □

We now translate Theorem E.2 into perturbation bounds on $\mathrm{Adapt}(\lambda)$ and $\mathrm{Forget}(\lambda)$, and on the optimal-trade-off value of Corollary 3.1.

Corollary E.1 (Near-orthogonal bounds on Adapt and Forget)**.** *Under the hypotheses of Theorem E.2, with C as there,*

$$\left|\mathrm{Adapt}(\lambda) - \gamma^{(0)}_{\mathrm{ft},\lambda}\,\frac{\Theta^{(0)}_{\lambda}}{\sqrt{(\Theta^{(0)}_{\lambda})^2 + \rho^2_{\mathrm{pre}}}}\right| \le C\,(E_\lambda + \epsilon_2)\,, \tag{92}$$

$$\left|\mathrm{Forget}(\lambda) - \gamma_{\mathrm{pre}}\Big(1 - \frac{\rho_{\mathrm{pre}}}{\sqrt{(\Theta^{(0)}_{\lambda})^2 + \rho^2_{\mathrm{pre}}}}\Big)\right| \le C\,(E_\lambda + \epsilon_2)\,. \tag{93}$$

In particular, as $E_\lambda + \epsilon_2 \to 0$, both $\mathrm{Adapt}(\lambda)$ and $\mathrm{Forget}(\lambda)$ converge to their strict-orthogonal closed-form values from Theorem E.1.

Proof. For the Forget bound, note that $\gamma_{\mathrm{pre}} = \gamma(W_{\mathrm{pre}};\mathcal{D}_{\mathrm{pre}})$ does not depend on ϵ_1, ϵ_2 and that $\mathrm{Forget}(\lambda) = \gamma_{\mathrm{pre}} - \gamma(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D}_{\mathrm{pre}})$, so (93) is just (70) with the γ_{pre} term absorbed. For the Adapt bound, use $\mathrm{Adapt}(\lambda) = \gamma(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D}_{\mathrm{ft}}) - \gamma(W_{\mathrm{pre}};\mathcal{D}_{\mathrm{ft}})$. The leakage bound (72) controls the second term: for any $\bar{x} \in \mathcal{D}_{\mathrm{ft}}$ and any pair $c \neq k$, $|(w_c - w_k)^\top \bar{x}| \le 2\|S\|_F$, so taking the min over $(\bar{x}, c, k)$ and dividing by $\|W_{\mathrm{pre}}\|_F = \rho_{\mathrm{pre}}$ gives $|\gamma(W_{\mathrm{pre}};\mathcal{D}_{\mathrm{ft}})| \le 2\|S\|_F/\rho_{\mathrm{pre}} \le 2E_\lambda/\rho_{\mathrm{pre}}$. Combining with (71) via the triangle inequality yields (92). □

Corollary E.2 (Near-orthogonal version of Corollary 3.1)**.** *Let $\gamma^{\mathrm{pad},(0)}_{\mathrm{ft}}$ denote the strict-orthogonal reference value of $\gamma^{\mathrm{pad}}_{\mathrm{ft}}$. Adopt the hypotheses of Theorem E.2, and set $\Theta^* := \rho_{\mathrm{pre}}\gamma_{\mathrm{pre}}/\gamma^{\mathrm{pad},(0)}_{\mathrm{ft}}$. If $\bar{K} < K$, assume $\Theta^* < \Theta^{(0)}_{\lambda_{\mathrm{crit}}}$; when $\bar{K} = K$, no additional balancing condition is needed. Let λ^* be the unique solution of $\Theta^{(0)}_{\lambda^*} = \Theta^*$ in $(\lambda_{\mathrm{crit}}, 1/(\bar{K}\sqrt{n}))$, and choose a compact subinterval Λ of this open interval containing λ^*. Writing $\delta_0(\lambda), s_0(\lambda)$ for the local thresholds in Theorem D.2, require*

$$\delta \le \inf_{\lambda\in\Lambda} \delta_0(\lambda)\,, \qquad \|S\|_F \le \inf_{\lambda\in\Lambda} s_0(\lambda)\,.$$

Suppose $E_\lambda \le \bar{E}$ for all $\lambda \in \Lambda$, and the constant C in Theorem E.2 can be taken uniform on Λ with value $\bar{C}$. Then

$$\left|\max_{\lambda\in\Lambda} \gamma(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}) - \Big(\frac{1}{(\gamma^{\mathrm{pad},(0)}_{\mathrm{ft}})^2} + \frac{1}{\gamma^2_{\mathrm{pre}}}\Big)^{-1/2}\right| \le \bar{C}\,(\bar{E} + \epsilon_2)\,. \tag{94}$$

In particular, as $\bar{E} + \epsilon_2 \to 0$ (with Λ fixed), the maximum converges to the corresponding strict-orthogonal intermediate-regime balancing value, which reduces to the value in Corollary 3.1 when $\bar{K} = K$.

Proof. By Theorem E.2 applied uniformly on Λ, together with the 1-Lipschitz property of $\min$,

$$\left|\gamma(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}) - \gamma^{(0)}(\lambda)\right| \le \bar{C}\,(\bar{E} + \epsilon_2) \quad \text{for every } \lambda \in \Lambda\,,$$

where $\gamma^{(0)}(\lambda) := \min\big\{\gamma_{\mathrm{pre}}\rho_{\mathrm{pre}}/\sqrt{(\Theta^{(0)}_{\lambda})^2 + \rho^2_{\mathrm{pre}}},\ \gamma^{(0)}_{\mathrm{ft},\lambda}\Theta^{(0)}_{\lambda}/\sqrt{(\Theta^{(0)}_{\lambda})^2 + \rho^2_{\mathrm{pre}}}\big\}$ is the strict-orthogonal union margin. Taking $\sup_{\lambda\in\Lambda}$ on both sides and using $|\sup_\Lambda f - \sup_\Lambda g| \le \sup_\Lambda |f - g|$,

$$\left|\max_{\lambda\in\Lambda} \gamma(W^{\lambda}_{\mathrm{LoRA}};\mathcal{D}_{\mathrm{pre}} \cup \mathcal{D}_{\mathrm{ft}}) - \max_{\lambda\in\Lambda} \gamma^{(0)}(\lambda)\right| \le \bar{C}\,(\bar{E} + \epsilon_2)\,.$$

On Λ, $\gamma^{(0)}_{\mathrm{ft},\lambda} = \gamma^{\mathrm{pad},(0)}_{\mathrm{ft}}$ and $\Theta^{(0)}_{\lambda}$ is strictly decreasing in λ. The pre-training term in $\gamma^{(0)}$ is strictly decreasing in $\Theta^{(0)}_{\lambda}$, while the fine-tuning term is strictly increasing. Their minimum is therefore uniquely maximized when $\Theta^{(0)}_{\lambda} = \Theta^*$, which occurs at $\lambda^* \in \Lambda$. Hence $\max_{\lambda\in\Lambda} \gamma^{(0)}(\lambda) = \gamma^{(0)}(\lambda^*) = (1/(\gamma^{\mathrm{pad},(0)}_{\mathrm{ft}})^2 + 1/\gamma^2_{\mathrm{pre}})^{-1/2}$, proving (94) for both $\bar{K} = K$ and $\bar{K} < K$. □

Remark E.2 (Explicit bounds on $\bar{E}$ and $\bar{C}$)**.** *The constants in Corollary E.2 can be chosen with the following explicit upper bounds. Set $\bar{E} := \sup_{\lambda\in\Lambda} E_\lambda$ and $\Theta^{(0)}_{\max} := \max_{\lambda\in\Lambda} \Theta^{(0)}_{\lambda}$. Since $a := \min\Lambda > 0$, the proof of Theorem D.2 uses radii bounded by $(\log K + 2)/a$. Thus $\mu_0(\lambda)$ can be chosen uniformly bounded below by a positive constant and $C_0(\lambda)$ uniformly bounded above. Write $\kappa_\Lambda := \sup_{\lambda\in\Lambda} C_0(\lambda)/\mu_0(\lambda) < \infty$.*

***Bound on** $\bar{E}$. For the conversion to an epsilon-only bound, assume the uniform positive singular-value lower bound discussed in Remark D.2. Combining the definition of E_λ in* (69) *with $\delta = (\bar{K}n - 1)\,\epsilon_1$ from Theorem D.2 and the explicit bound $\|S\|_F \le \sqrt{\bar{K}n\,N_{\mathrm{pre}}}\,\rho_{\mathrm{pre}}\,\epsilon_2/\sigma_{\min}(X_{\mathrm{pre}})$ from that remark,*

$$\bar{E} \le \kappa_\Lambda \left(\frac{\sqrt{\bar{K}n\,N_{\mathrm{pre}}}\,\rho_{\mathrm{pre}}}{\sigma_{\min}(X_{\mathrm{pre}})}\,\epsilon_2 + (\bar{K}n - 1)\big(1 + \Theta^{(0)}_{\max}\big)\,\epsilon_1 \right), \tag{95}$$

which is linear in ϵ_1 and ϵ_2.

***Bound on** $\bar{C}$. The constant C in the proof of Theorem E.2 is built from three sources:*

- *(i) the direct leakage bound* (72) *has coefficient* 1 *in $\|S\|_F$, the projection bound* (77) *contributes $L_{\mathrm{proj}} := \sqrt{2}$, and* (74) *contributes $L_{\mathrm{ft}} := \sqrt{2\bar{K}n}$, uniformly since $\delta \le 1/2$;*
- *(ii) Cauchy–Schwarz over K classes contributes a factor $\sqrt{K}$;*
- *(iii) the unit-norm input bound $|\gamma(W;\mathcal{D})| \le \sqrt{2}$ and* (91) *contribute a factor at most $(1 + \sqrt{2})/\rho_{\mathrm{pre}}$ when converting numerator and norm errors to normalized-margin errors.*

Using $E_\lambda\epsilon_2 \le E_\lambda$ in (74) *and absorbing universal numerical factors into $C_* > 0$,*

$$\bar{C} \le \frac{C_*\,\sqrt{K}\,(1 + L_{\mathrm{proj}} + L_{\mathrm{ft}})}{\rho_{\mathrm{pre}}}\,\big(1 + \rho_{\mathrm{pre}} + \Theta^{(0)}_{\max}\big)\,. \tag{96}$$

Under the stated conditioning requirement, substituting (95) *and* (96) *into the right-hand side of* (94) *yields a fully explicit linear-in-(ϵ_1, ϵ_2) bound on the gap between the near-orthogonal optimum and the strict-orthogonal value.*

F Proof of Theorem 3.3

In this section, we first present the proof of Theorem 3.3. Then, under additional smallness assumptions on ϵ_1, ϵ_2, we show that the same conclusions still hold, with additional error terms that depend on ϵ_1, ϵ_2.

F.1 Proof of Theorem 3.3

Proof. For $\boldsymbol{\alpha} \in \mathbb{R}^T_{>0}$, following the same argument as in Appendix E, one can show that the margin of linear classifier W_{LoRA} for each task is

$$\gamma\big(W_{\mathrm{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\mathrm{pre}}\big) = \gamma_{\mathrm{pre}}\,\frac{\rho_{\mathrm{pre}}}{\sqrt{\rho^2_{\mathrm{pre}} + \sum_{j=1}^T \alpha_j^2 \Theta^2_{\lambda,j}}}, \tag{97}$$

$$\gamma\big(W_{\mathrm{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_i\big) = \gamma_i\,\frac{\alpha_i \Theta_{\lambda,i}}{\sqrt{\rho^2_{\mathrm{pre}} + \sum_{j=1}^T \alpha_j^2 \Theta^2_{\lambda,j}}}, \quad i = 1, \ldots, T. \tag{98}$$

Then, our goal is to solve the following optimization problem

$$\max_{\alpha_1,\cdots,\alpha_T \ge 0} \min\left(\gamma_{\mathrm{pre}}\,\frac{\rho_{\mathrm{pre}}}{\sqrt{\rho^2_{\mathrm{pre}} + \sum_{j=1}^T \alpha_j^2 \Theta^2_{\lambda,j}}}\,,\quad \min_{i\le T} \gamma_i\,\frac{\alpha_i \Theta_{\lambda,i}}{\sqrt{\rho^2_{\mathrm{pre}} + \sum_{j=1}^T \alpha_j^2 \Theta^2_{\lambda,j}}} \right)$$

For convenience, we introduce the following notation $x_i = \alpha_i \Theta_{\lambda,i} \geq 0, i \in [T]$, then the above optimization problem can be rewritten as

$$\max_{x_1,\cdots,x_T \geq 0} \min\left(\frac{\rho_{\text{pre}}\gamma_{\text{pre}}}{\sqrt{\rho_{\text{pre}}^2 + \sum_{j=1}^T x_j^2}}, \quad \min_{i\in[T]} \frac{\gamma_i x_i}{\sqrt{\rho_{\text{pre}}^2 + \sum_{j=1}^T x_j^2}} \right). \tag{99}$$

Let $S := \sqrt{\rho_{\text{pre}}^2 + \sum_{j=1}^T x_j^2}$ and $t := \min\{\frac{\gamma_i x_i}{S}, \frac{\gamma_{\text{pre}}\rho_{\text{pre}}}{S}\}$, then Problem 99 is equivalent to

$$\begin{aligned} \max_{t,S,x_1,\cdots,x_T} \quad & t \\ \text{s.t.} \quad & x_i \geq 0, \quad \gamma_i x_i \geq tS, \quad i \in [T] \\ & \gamma_{\text{pre}}\rho_{\text{pre}} \geq tS\,, \\ & \sum_{i=1}^T x_i^2 + \rho_{\text{pre}}^2 = S^2, \quad S > 0\,. \end{aligned} \tag{100}$$

We first claim at optimum, the inequalities w.r.t. x_i will be tight, i.e., $x_i = \frac{tS}{\gamma_i}, \forall i \in [T]$. In this case, one can see that

$$S^2 = \rho_{\text{pre}}^2 + \sum_{j=1}^T x_j^2 \geq S^2 t^2 \cdot \left(\sum_{i=1}^T \frac{1}{\gamma_i^2} + \frac{1}{\gamma_{\text{pre}}^2} \right).$$

Thus, the optimal value of the objective is $t \leq \frac{1}{\sqrt{\frac{1}{\gamma_{\text{pre}}^2} + \sum_{j=1}^T \frac{1}{\gamma_j^2}}}$, and the equality is achieved under the condition $\gamma_i x_i = \gamma_{\text{pre}}\rho_{\text{pre}}$, which is equivalent to $\alpha_i = \frac{\rho_{\text{pre}}\gamma_{\text{pre}}}{\gamma_i \Theta_{\lambda,i}},\ i = 1, \ldots, T$.

Finally, we show that at the optimum of Problem 100, both the adapter and pre-training inequalities are tight.

Case 1: some adapter inequality is slack. Assume for the optimal solution $(t^*, S^*, \boldsymbol{x}^*)$ there exists $j \in [T]$ such that $\gamma_j x_j^* > t^* S^*$. Decrease x_j^* slightly to $x_j^* - \epsilon$ with $\epsilon > 0$. The new S is

$$\tilde{S} = \sqrt{\rho_{\text{pre}}^2 + \sum_{i\neq j} (x_i^*)^2 + (x_j^* - \epsilon)^2} < S^*\,.$$

For ϵ sufficiently small, every other adapter inequality remains satisfied (since $\tilde{S} < S^*$), the pre-training inequality remains satisfied ($\gamma_{\text{pre}}\rho_{\text{pre}} \geq t^* S^* > t^* \tilde{S}$), and $\gamma_j(x_j^* - \epsilon) > t^* \tilde{S}$ by continuity. Therefore one can increase t^* until one inequality becomes tight, raising the objective and contradicting optimality.

Case 2: all adapter inequalities are tight but the pre-training inequality is slack. Assume $\gamma_i x_i^* = t^* S^*$ for every $i \in [T]$ but $\gamma_{\text{pre}}\rho_{\text{pre}} > t^* S^*$. Take $\tilde{x}_i := c x_i^*$ for some $c > 1$, leaving ρ_{pre} unchanged. Then

$$\tilde{S} = \sqrt{\rho_{\text{pre}}^2 + c^2 \sum_i (x_i^*)^2} = \sqrt{c^2 (S^*)^2 - \rho_{\text{pre}}^2 (c^2 - 1)} < c\, S^*\,.$$

Setting $\tilde{t} := c t^* S^* / \tilde{S}$ keeps all adapter inequalities tight (since $\gamma_i \tilde{x}_i = c\gamma_i x_i^* = c t^* S^* = \tilde{t}\tilde{S}$), and $\tilde{t} > t^*$ from $\tilde{S} < cS^*$. For c close enough to 1, the pre-training inequality $\gamma_{\text{pre}}\rho_{\text{pre}} \geq \tilde{t}\tilde{S} = c t^* S^*$ remains satisfied by continuity. Hence $(\tilde{t}, \tilde{S}, \tilde{\boldsymbol{x}})$ is feasible with $\tilde{t} > t^*$, again contradicting optimality.

Combining the two cases, all inequalities must be tight at any optimum. □

F.2 Proof of General Version

In this subsection we present the general version of Theorem 3.3: under a near-orthogonal multi-task assumption (defined below), the margins of the merged model $W_{\text{LoRA}}(\boldsymbol{\alpha})$ on the pre-training data and on each fine-tuning task remain close to the strict-orthogonal closed-form values from Theorem 3.3,

with deviations controlled by the perturbation error E defined below. Consequently, the optimal mixing coefficients of Theorem 3.3 continue to (approximately) maximize the merged margin.

Near-orthogonal multi-task assumption. Let $X_i \in \mathbb{R}^{d\times \bar{K}_i n_i}$ denote the data matrix of task $i \in [T]$. We replace Assumption 3.1 by the following near-orthogonal version: there exist $\epsilon_1, \epsilon_2 \in [0, 1)$ such that

(a) Every fine-tuning sample has unit norm: $\|\bar{x}\|_2 = 1$ for every $\bar{x} \in \bigcup_{i=1}^{T} \mathcal{D}_i$.

(b) *Within-task:* for each $i \in [T]$ and each pair of distinct $\bar{x}, \bar{x}' \in \mathcal{D}_i$, $|\bar{x}^\top \bar{x}'| \le \epsilon_1$.

(c) *Cross-task and pre-training:* for each $i \neq j$ in $[T]$, every $\bar{x} \in \mathcal{D}_i$ and $\bar{x}' \in \mathcal{D}_j$ satisfy $|\bar{x}^\top \bar{x}'| \le \epsilon_2$; and for every $\bar{x} \in \bigcup_i \mathcal{D}_i$ and every $x \in \mathcal{D}_{\mathrm{pre}}$, $|\bar{x}^\top x| \le \epsilon_2$.

The strict-orthogonal Assumption 3.1 is recovered at $\epsilon_1 = \epsilon_2 = 0$.

Notation. For each task $i \in [T]$ we adopt the per-task analogues of the notation of Theorem D.2: $G_i := X_i^\top X_i$, $R_i := G_i^{1/2}$, $Q_i := X_i R_i^{-1}$, $S_i := W_{\mathrm{pre}} X_i$, $\delta_i := (\bar{K}_i n_i - 1)\epsilon_1$, and the unique optimum $Z^*_{\lambda,i}(\bar{S}, \bar{R})$ of $\Phi_i(Z; \bar{S}, \bar{R})$ defined as in Theorem D.2 but with $X_i, \bar{K}_i, n_i$ in place of $X_{\mathrm{ft}}, \bar{K}, n$. We take δ_0, s_0, μ_0 as the minima of their per-task counterparts and C_0 as the maximum. By the per-task reformulation, $B_i^* A_i^* = Z^*_{\lambda,i}(S_i, R_i) Q_i^\top$, where the rank constraint $\mathrm{rank}(Z) \le r$ is non-binding for $\delta_i, \|S_i\|_F$ sufficiently small by Lemma D.5 applied per task. Write

$$\Theta^{(0)}_{\lambda,i} := \|Z^*_{\lambda,i}(0, I)\|_F\,, \qquad \gamma_i^{(0)} := \gamma_i\big|_{\epsilon_1=\epsilon_2=0}\,, \tag{101}$$

for the strict-orthogonal per-task adapter Frobenius norm and per-task margin (so $\Theta^{(0)}_{\lambda,i} = \Theta_{\lambda,i}$ and $\gamma_i^{(0)} = \gamma_i$ from Theorem 3.3 when $\epsilon_1 = \epsilon_2 = 0$). Set

$$E_{\lambda,i} := \frac{C_0}{\mu_0}\big(\|S_i\|_F + \delta_i(1 + \Theta^{(0)}_{\lambda,i})\big)\,, \qquad E := \max_{i\in[T]} E_{\lambda,i} + \epsilon_2\,. \tag{102}$$

The common constants satisfy $C_0/\mu_0 \ge 1$, so $\|S_i\|_F \le E_{\lambda,i} \le E$. Under a uniform positive lower bound on $\sigma_{\min}(X_{\mathrm{pre}})$, $\|S_i\|_F = O(\epsilon_2)$ uniformly in i (Remark D.2), so $E \to 0$ as $\epsilon_1, \epsilon_2 \to 0$.

Theorem F.1 (General version of Theorem 3.3)**.** *Adopt the near-orthogonal multi-task assumption above with $\delta_i \le \delta_0$ and $\|S_i\|_F \le s_0$ for every $i \in [T]$, using the common thresholds chosen above. Suppose each task is trained in the intermediate or low-penalty regime, and the pre-training samples have unit norm. There exist constants $E_0 \in (0, 1]$ and $C > 0$, depending on $\lambda, K, \{\bar{K}_i, n_i\}_{i=1}^T, \rho_{\mathrm{pre}}, \mu_0, C_0, \{\|Z^*_{\lambda,i}(0, I)\|_*\}_{i=1}^T$ but not on ϵ_1, ϵ_2 or $\boldsymbol{\alpha}$, such that whenever $E \le E_0$, the following bounds hold uniformly for all nonnegative coefficient vectors $\boldsymbol{\alpha} \in \mathbb{R}^T_{\ge 0}$:*

$$\left|\gamma\big(W_{\mathrm{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\mathrm{pre}}\big) - \frac{\gamma_{\mathrm{pre}}\,\rho_{\mathrm{pre}}}{\sqrt{\rho_{\mathrm{pre}}^2 + \sum_{i=1}^T \alpha_i^2 (\Theta^{(0)}_{\lambda,i})^2}}\right| \le CE\,, \tag{103}$$

$$\left|\gamma\big(W_{\mathrm{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_j\big) - \frac{\gamma_j^{(0)}\,\alpha_j\,\Theta^{(0)}_{\lambda,j}}{\sqrt{\rho_{\mathrm{pre}}^2 + \sum_{i=1}^T \alpha_i^2 (\Theta^{(0)}_{\lambda,i})^2}}\right| \le CE\,, \quad j \in [T]\,. \tag{104}$$

Write $F(\boldsymbol{\alpha}) := \gamma(W_{\mathrm{LoRA}}(\boldsymbol{\alpha}); \mathcal{D}_{\mathrm{pre}} \cup \bigcup_i \mathcal{D}_i)$ and let $F^{(0)}(\boldsymbol{\alpha})$ be the minimum of the reference margins displayed above. The 1-Lipschitz property of min *gives $|F(\boldsymbol{\alpha}) - F^{(0)}(\boldsymbol{\alpha})| \le CE$ uniformly over $\boldsymbol{\alpha} \ge 0$. In particular, the strict-orthogonal optimum $\alpha_i^* = \rho_{\mathrm{pre}}\gamma_{\mathrm{pre}}/(\gamma_i^{(0)}\Theta^{(0)}_{\lambda,i})$ satisfies*

$$0 \le \sup_{\boldsymbol{\alpha}\in\mathbb{R}^T_{\ge 0}} F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^*) \le 2CE.$$

Thus $\boldsymbol{\alpha}^*$ remains optimal up to an additive $O(E)$ error in achieved margin, and all margins converge uniformly to their strict-orthogonal reference values as $E \to 0$.

Proof. Throughout, C denotes a positive constant whose value may change line to line, but always depends only on the quantities listed in the theorem statement. We work with $E \le 1$ and choose $E_0 \le 1$ below.

We first establish leakage bounds. Since $S_i = W_{\mathrm{pre}} X_i$, for every task $i \in [T]$, sample $\bar{x}_j^{(i)} \in \mathcal{D}_i$, and row w_k of W_{pre},

$$|w_k^\top \bar{x}_j^{(i)}| = |(S_i)_{kj}| \le \|S_i\|_F \le E_{\lambda,i}\,. \tag{105}$$

For each i, write $\delta_{\lambda,i,k} := (B_i^* A_i^*)_k^\top$ for the k-th row of $B_i^* A_i^*$. Using $B_i^* A_i^* = Z_{\lambda,i}^*(S_i, R_i) Q_i^\top$, we obtain (i) for any unit-norm $x \in \mathcal{D}_{\mathrm{pre}}$,

$$|\delta_{\lambda,i,k}^\top x| \le \|(Z_{\lambda,i}^*(S_i,R_i))_k\|_2 \, \|Q_i^\top x\|_2 \le C\,\epsilon_2\,, \tag{106}$$

where the last step uses $\|Q_i^\top x\|_2 \le \sqrt{\bar{K}_i n_i}\,\epsilon_2/\sqrt{1-\delta_i}$ (from $|\bar{x}_l^{(i),\top} x| \le \epsilon_2$) and $\|Z_{\lambda,i}^*(S_i,R_i)\|_F \le \Theta_{\lambda,i}^{(0)} + E_{\lambda,i} \le C$. (ii) For any $\bar{x} \in \mathcal{D}_j$ with $j \ne i$, the same argument with X_j in place of $\{x\}$ gives

$$|\delta_{\lambda,i,k}^\top \bar{x}| \le C\,\epsilon_2\,, \tag{107}$$

since each entry of $X_i^\top \bar{x}$ is bounded by ϵ_2 (cross-task assumption). The adapter leakage bounds are $O(\epsilon_2)$ uniformly in i, k, and the pre-trained classifier leakage is controlled by $\|S_i\|_F$.

Then, we bound the perturbation of the merged-model Frobenius norm. Expanding,

$$\|W_{\mathrm{LoRA}}(\boldsymbol{\alpha})\|_F^2 = \rho_{\mathrm{pre}}^2 + \sum_{i=1}^T \alpha_i^2 \, \|B_i^* A_i^*\|_F^2 + 2\sum_{i=1}^T \alpha_i \big\langle W_{\mathrm{pre}}, B_i^* A_i^* \big\rangle + \sum_{i \ne j} \alpha_i \alpha_j \big\langle B_i^* A_i^*, B_j^* A_j^* \big\rangle\,. \tag{108}$$

The identity $W_{\mathrm{pre}} Q_i = S_i R_i^{-1}$ and the same projection argument as in the proof of Theorem E.2 give

$$\big|\big\langle W_{\mathrm{pre}}, B_i^* A_i^* \big\rangle\big| \le \frac{\|S_i\|_F \, \|B_i^* A_i^*\|_F}{\sqrt{1-\delta_i}} \le C\,(\Theta_{\lambda,i}^{(0)} + E_{\lambda,i})\,\|S_i\|_F\,. \tag{109}$$

For the cross-task term with $i \ne j$, each row $(B_i^* A_i^*)_k$ lies in $\mathrm{span}(X_i)$ and $(B_j^* A_j^*)_k \in \mathrm{span}(X_j)$. Writing $u \in \mathrm{span}(X_i)$ as $u = X_i \alpha$ with $\|\alpha\|_2 \le \|u\|_2/\sigma_{\min}(R_i) \le \|u\|_2/\sqrt{1-\delta_i} \le C\,\|u\|_2$ (and similarly for $v = X_j \beta$), we have $u^\top v = \alpha^\top X_i^\top X_j \beta$ with each entry of $X_i^\top X_j$ bounded by ϵ_2 in absolute value, hence $|u^\top v| \le \|X_i^\top X_j\|_{\mathrm{op}}\,\|\alpha\|_2\,\|\beta\|_2 \le C\,\epsilon_2\,\|u\|_2\,\|v\|_2$. Summing this bound over rows k and applying Cauchy–Schwarz,

$$\big|\big\langle B_i^* A_i^*, B_j^* A_j^* \big\rangle\big| \le C\,\epsilon_2\,\|B_i^* A_i^*\|_F\,\|B_j^* A_j^*\|_F \le C\,\epsilon_2\,(\Theta_{\lambda,i}^{(0)} + E_{\lambda,i})(\Theta_{\lambda,j}^{(0)} + E_{\lambda,j})\,. \tag{110}$$

Set $a_{\boldsymbol{\alpha}} := \|W_{\mathrm{LoRA}}(\boldsymbol{\alpha})\|_F$ and $b_{\boldsymbol{\alpha}}^2 := \rho_{\mathrm{pre}}^2 + \sum_i \alpha_i^2 (\Theta_{\lambda,i}^{(0)})^2$. Combining (109), (110) with the per-task bound $\big|\|B_i^* A_i^*\|_F - \Theta_{\lambda,i}^{(0)}\big| \le E_{\lambda,i} + O(\delta_i \Theta_{\lambda,i}^{(0)})$ from Theorem D.2 and using $E \le 1$ yields, for some $D > 0$ independent of $\boldsymbol{\alpha}, \epsilon_1, \epsilon_2$,

$$|a_{\boldsymbol{\alpha}}^2 - b_{\boldsymbol{\alpha}}^2| \le DE(1 + \|\boldsymbol{\alpha}\|_2^2).$$

Here the linear and mixed coefficient terms are controlled using $\|\boldsymbol{\alpha}\|_1 \le \sqrt{T}\|\boldsymbol{\alpha}\|_2$ and $\|\boldsymbol{\alpha}\|_1^2 \le T\|\boldsymbol{\alpha}\|_2^2$. The stated regimes ensure $\Theta_{\lambda,i}^{(0)} > 0$ for every i, so

$$c := \min\{\rho_{\mathrm{pre}}, \min_i \Theta_{\lambda,i}^{(0)}\} > 0, \qquad b_{\boldsymbol{\alpha}}^2 \ge c^2(1 + \|\boldsymbol{\alpha}\|_2^2).$$

Thus, setting $\kappa := D/c^2$, we obtain the uniform relative bound

$$|a_{\boldsymbol{\alpha}}^2 - b_{\boldsymbol{\alpha}}^2| \le \kappa E\, b_{\boldsymbol{\alpha}}^2. \tag{111}$$

Choose $E_0 := \min\{1, (2\kappa)^{-1}\}$. For $E \le E_0$, this gives, uniformly over $\boldsymbol{\alpha} \ge 0$,

$$a_{\boldsymbol{\alpha}} \ge \frac{b_{\boldsymbol{\alpha}}}{\sqrt{2}} > 0, \qquad \frac{|a_{\boldsymbol{\alpha}} - b_{\boldsymbol{\alpha}}|}{b_{\boldsymbol{\alpha}}} = \frac{|a_{\boldsymbol{\alpha}}^2 - b_{\boldsymbol{\alpha}}^2|}{b_{\boldsymbol{\alpha}}(a_{\boldsymbol{\alpha}} + b_{\boldsymbol{\alpha}})} \le \kappa E.$$

Next, we bound the perturbation of the unnormalized margins on each dataset. For $(x, y_c) \in \mathcal{D}_{\mathrm{pre}}$ and $k \ne c$, decomposing

$$\Big[\big(w_c + \textstyle\sum_i \alpha_i \delta_{\lambda,i,c}\big) - \big(w_k + \textstyle\sum_i \alpha_i \delta_{\lambda,i,k}\big)\Big]^\top x = (w_c - w_k)^\top x + \textstyle\sum_{i=1}^T \alpha_i (\delta_{\lambda,i,c} - \delta_{\lambda,i,k})^\top x\,, \tag{112}$$

and applying (106) to each adapter term,

$$\left|C(W_{\mathrm{LoRA}}(\boldsymbol{\alpha});\mathcal{D}_{\mathrm{pre}})-\rho_{\mathrm{pre}}\gamma_{\mathrm{pre}}\right|\le C\,\|\boldsymbol{\alpha}\|_1\,\epsilon_2\,. \tag{113}$$

For $(\bar{x},y_c)\in\mathcal{D}_j$ and $k\neq c$,

$$\big[\cdots\big]^\top\bar{x}=\alpha_j(\delta_{\lambda,j,c}-\delta_{\lambda,j,k})^\top\bar{x}+\sum_{i\neq j}\alpha_i(\delta_{\lambda,i,c}-\delta_{\lambda,i,k})^\top\bar{x}+(w_c-w_k)^\top\bar{x}\,. \tag{114}$$

The first term in (114) is the within-task contribution. Since $\alpha_j\ge 0$, it can be taken outside the minimum. By the same argument as in the proof of Theorem E.2 (Step 3 there), $\big|\min(w.r.t.\,(\bar{x},c,k))\,\alpha_j(\delta_{\lambda,j,c}-\delta_{\lambda,j,k})^\top\bar{x}-\alpha_j\,\Theta^{(0)}_{\lambda,j}\gamma^{(0)}_j\big|\le C\,|\alpha_j|\,(E_{\lambda,j}+\delta_j)$. The cross-task adapter terms are bounded via (107): $|\alpha_i(\delta_{\lambda,i,c}-\delta_{\lambda,i,k})^\top\bar{x}|\le 2\,C\,|\alpha_i|\,\epsilon_2$, and the pre-training leakage on $\mathcal{D}_j$ contributes at most $2\|S_j\|_F\le 2E_{\lambda,j}$ via (105). Combining via the triangle inequality,

$$\left|C(W_{\mathrm{LoRA}}(\boldsymbol{\alpha});\mathcal{D}_j)-\alpha_j\,\Theta^{(0)}_{\lambda,j}\gamma^{(0)}_j\right|\le C\,(1+\|\boldsymbol{\alpha}\|_1)\,E\,, \tag{115}$$

where we have absorbed δ_j into E via $\delta_j\le(\mu_0/C_0)E_{\lambda,j}\le(\mu_0/C_0)E$.

Finally, fix any dataset $\mathcal{D}\in\{\mathcal{D}_{\mathrm{pre}},\mathcal{D}_1,\ldots,\mathcal{D}_T\}$ and set $u_{\boldsymbol{\alpha}}:=C(W_{\mathrm{LoRA}}(\boldsymbol{\alpha});\mathcal{D})$. Its reference numerator $u^{(0)}_{\boldsymbol{\alpha}}$ is $\rho_{\mathrm{pre}}\gamma_{\mathrm{pre}}$ for $\mathcal{D}_{\mathrm{pre}}$, and $\alpha_j\Theta^{(0)}_{\lambda,j}\gamma^{(0)}_j$ for $\mathcal{D}_j$. Equations (113)–(115) give $|u_{\boldsymbol{\alpha}}-u^{(0)}_{\boldsymbol{\alpha}}|\le C(1+\|\boldsymbol{\alpha}\|_1)E$. Since

$$\frac{1+\|\boldsymbol{\alpha}\|_1}{b_{\boldsymbol{\alpha}}}\le\frac{\sqrt{T+1}\sqrt{1+\|\boldsymbol{\alpha}\|_2^2}}{b_{\boldsymbol{\alpha}}}\le\frac{\sqrt{T+1}}{c},$$

the numerator error divided by $b_{\boldsymbol{\alpha}}$ is uniformly $O(E)$. All inputs have unit norm and $a_{\boldsymbol{\alpha}}>0$, so $|u_{\boldsymbol{\alpha}}/a_{\boldsymbol{\alpha}}|\le\sqrt{2}$. Applying the ratio estimate (91) and the relative norm bound above,

$$\left|\frac{u_{\boldsymbol{\alpha}}}{a_{\boldsymbol{\alpha}}}-\frac{u^{(0)}_{\boldsymbol{\alpha}}}{b_{\boldsymbol{\alpha}}}\right|\le\frac{|u_{\boldsymbol{\alpha}}-u^{(0)}_{\boldsymbol{\alpha}}|+\sqrt{2}\,|a_{\boldsymbol{\alpha}}-b_{\boldsymbol{\alpha}}|}{b_{\boldsymbol{\alpha}}}\le CE.$$

This proves (103)–(104) uniformly in $\boldsymbol{\alpha}$. Taking the minimum across datasets gives $|F(\boldsymbol{\alpha})-F^{(0)}(\boldsymbol{\alpha})|\le CE$ uniformly.

For the optimality claim, Theorem 3.3 gives $F^{(0)}(\boldsymbol{\alpha})\le F^{(0)}(\boldsymbol{\alpha}^*)$ for every $\boldsymbol{\alpha}\ge 0$. Therefore,

$$F(\boldsymbol{\alpha})\le F^{(0)}(\boldsymbol{\alpha})+CE\le F^{(0)}(\boldsymbol{\alpha}^*)+CE\le F(\boldsymbol{\alpha}^*)+2CE.$$

Taking the supremum over all nonnegative coefficient vectors proves the stated $2CE$ near-optimality bound. □

G Lemmas

This section first lists the preliminary lemmas (cited from external sources) used throughout the proofs, and then gives full proofs of the paper-specific supporting lemmas.

G.1 Preliminary Lemmas

In this subsection, we provide several preliminary lemmas that will be used in the proof.

Lemma G.1 (Theorem 3 in Riedel [1992]). *Let $A\in\mathbb{R}^{l\times l}$ with $\mathrm{rank}(A)<l$, and $v_i,w_i,i=1,2$ be vectors in $\mathbb{R}^l$. Let $v_1\in M(A),w_1\perp M(A)$, and $v_2\in M(A^*),w_2\perp M(A^*)$, where $M(A)$ denotes the range of A. Assume $w_2\parallel w_1$ and $w_i\neq 0,i=1,2$. Let $\Omega=A+(v_1+w_1)(v_2+w_2)^*$, Then*

$$\Omega^\dagger=A^\dagger-\frac{w_2v_2^*A^\dagger}{\|w_2\|^2}-\frac{A^\dagger v_1w_1^*}{\|w_1\|^2}+(1+v_2^*A^\dagger v_1)\cdot\frac{w_2w_1^*}{\|w_1\|^2\|w_2\|^2}\,.$$

We refer the readers to Riedel [1992] for the proof.

Lemma G.2 (Variational form of the nuclear norm). *For any fixed $Z\in\mathbb{R}^{K\times m}$, we have*

$$\|Z\|_*=\min_{Z=BA}\frac{1}{2}(\|B\|_F^2+\|A\|_F^2)\,.$$

We refer the readers to Recht et al. [2010] for the proof.

Lemma G.3. *For any $Z \in \mathbb{R}^{K\times m}$,*

$$\|Z\|_* = \max_{V\in\mathbb{R}^{m\times K},\, \|V\|_{\mathrm{op}}\le 1} \mathrm{Tr}(ZV)\,.$$

For any orthogonal projection $P \in \mathbb{R}^{K\times K}$ and any $Z \in \mathbb{R}^{K\times m}$,

$$\|PZ\|_* \le \|P\|_{\mathrm{op}}\,\|Z\|_* \le \|Z\|_*\,,$$

Equality $\|PZ\|_* = \|Z\|_*$ holds if and only if $PZ = Z$. Moreover, for any two matrices $A, B \in \mathbb{R}^{K\times m}$, one has

$$\sum_{i=1}^{\min(K,m)} \sigma_i(AB^\top) \le \sum_{i=1}^{\min(K,m)} \sigma_i(A)\sigma_i(B)\,.$$

We refer the readers to Horn and Johnson [2012] for the proof.

Lemma G.4 (Berge's Maximum Theorem)**.** *Let X, Θ be topological spaces, $f : X \times \Theta \to \mathbb{R}$ be a continuous function on the product $X \times \Theta$, and $C : \Theta \to X$ be a compact-valued correspondence such that $C(\theta) \neq \emptyset$ for all $\theta \in \Theta$. Define the marginal function $f^* : \Theta \to \mathbb{R}$ and the set of minimizers $C^* : \Theta \to X$ by*

$$\begin{aligned} f^*(\theta) &= \sup\{f(x,\theta) : x \in C(\theta)\} \\ C^*(\theta) &= \arg\max\{f(x,\theta) : x \in C(\theta)\} = \{x \in C(\theta) : f(x;\theta) = f^*(\theta)\}\,. \end{aligned}$$

If C is continuous at θ, then the value function $f^*(\theta)$ is continuous, and the set of maximizers $C^*(\theta)$ is upper-hemicontinuous with nonempty and compact values. Moreover, if $C^*(\theta)$ is single-valued, and thus is a continuous function rather than a correspondence.

We refer the readers to Sundaram [1996] for the proof.

Lemma G.5 (Invariance under Permutation)**.** *Suppose X satisfies $\Pi X \Pi^\top = X$ for any permutation Π, then $X = a\boldsymbol{I} + c\mathbf{1}\mathbf{1}^\top$ for any constants $a, c \in \mathbb{R}$.*

We refer the readers to Lemma 10 in Hong and Ling [2023] for the proof.

Proposition G.1 (Representer Theorem for Hard-margin SVM Problem)**.** *Let $\{(\boldsymbol{x}_{c,i}, \boldsymbol{y}_c)\}_{i=1,c=1}^{n_c,K}$ be the dataset with features $\boldsymbol{x}_{c,i} \in \mathbb{R}^d$ and class labels $\boldsymbol{y}_c \in \mathbb{R}^K$. Then, let W^{mm} be the solution of the following hard-margin SVM problem, i.e.,*

$$\min_{W\in\mathbb{R}^{K\times d}} \ \tfrac{1}{2}\|W\|_F^2 \quad \textit{s.t.} \quad \forall\, i \in [K],\ \forall\, j \in [n_i],\ \forall\, k \neq i:\ \boldsymbol{w}_i^\top \boldsymbol{x}_{i,j} \ \ge\ \boldsymbol{w}_k^\top \boldsymbol{x}_{i,j} + 1\,,$$

where $W = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K]^\top$. Then, there exists a set of scalars $\{\alpha_{i,c,j}\}_{i=1,c=1,j=1}^{K,K,n_i}$ such that each row of W^{mm} has the following structure

$$\boldsymbol{w}_i^{\mathrm{mm}} = \sum_{c=1}^{K}\sum_{j=1}^{n_c} \alpha_{i,c,j}\boldsymbol{x}_{c,j}\,.$$

We refer the readers to Scholkopf and Smola [2018] for the proof.

G.2 Proof of Lemma D.1

Proof. By Proposition G.1 and Assumption 2.4, each row of W_{pre} lies in the span of the pre-training features. Under Assumption 2.1(c) with $\epsilon_2 = 0$, every fine-tuning feature is orthogonal to this span, so

$$W_{\mathrm{pre}} X_{\mathrm{ft}} = 0\,. \tag{116}$$

Moreover, Assumption 2.1(a)–(b) with $\epsilon_1 = 0$ gives $X_{\rm ft}^\top X_{\rm ft} = \boldsymbol{I}_{\bar{K}n}$. We apply change of variable as follows: $\tilde{A} = AX_{\rm ft} \in \mathbb{R}^{r\times \bar{K}n}$.

$$
\begin{aligned}
&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,(W_{\rm pre}+BA)x_{c,j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|A\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,W_{\rm pre}X_{\rm ft}e_{(c-1)n+j}+BAX_{\rm ft}e_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|A\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,W_{\rm pre}X_{\rm ft}e_{(c-1)n+j}+BAQRe_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|AQ\|_F^2+\|AQ_\perp\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,W_{\rm pre}X_{\rm ft}e_{(c-1)n+j}+B\tilde{A}Re_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|\tilde{A}\|_F^2+\|AQ_\perp\|_F^2\right).
\end{aligned}
$$

Under this transformation, the objective in (5) can be rewritten as follows

$$
\begin{aligned}
&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,(W_{\rm pre}+BA)x_{c,j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|A\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,BAX_{\rm ft}e_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|A\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,B\tilde{A}e_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|AX_{\rm ft}\|_F^2+\|AX_{{\rm ft},\perp}\|_F^2\right)\\
=&\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\left(y_c,B\tilde{A}e_{(c-1)n+j}\right)+\frac{\lambda}{2}\left(\|B\|_F^2+\|\tilde{A}\|_F^2+\|AX_{{\rm ft},\perp}\|_F^2\right).
\end{aligned}
$$

Notice that the component $AX_{{\rm ft},\perp}$ contributes only to the *regularization* term and does not affect the cross-entropy loss. Hence at any global minimizer (B^*, A^*) of (5) we must have

$$AX_{{\rm ft},\perp} = 0.$$

Moreover, if (B^*, A^*) is a global minimizer of the original problem, then

$$(B^*,\ A^*X_{\rm ft})$$

is a global minimizer of the objective in (25). Conversely, let $(B^*, \tilde{A}^*)$ be any global minimizer of (25), then one can solve the following equation to obtain (B^*, A^*) which is a global minimizer of the original objective in (5).

$$A^*X_{\rm ft} = \tilde{A}^*, \quad A^*X_{{\rm ft},\perp} = 0\,. \tag{117}$$

□

G.3 Proof of Lemma D.2

Proof. To show that there exists a unique value ρ_λ such that the (Problem one) and (Problem two) enjoys the same set of solutions, we first introduce two additional optimization problems.

$$\min_Z\ \frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\big(y_c,\ Ze_{(c-1)n+j}\big)+\lambda\|Z\|_*\,, \tag{Problem three}$$

$$\min_Z\ \frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\rm CE}\big(y_c,\ Ze_{(c-1)n+j}\big)\quad \text{s.t.}\|Z\|_*\le\rho_\lambda\,. \tag{Problem four}$$

where $Z\in\mathbb{R}^{K\times\bar{K}n}$. Let $S_i(\lambda), i=1,2,3,4$ be the solution sets of the above four optimization problems for the same fixed λ. We will show that

- $S_1(\lambda) = S_2(\lambda), S_3(\lambda) = S_4(\lambda)$.
- $S_3(\lambda), S_4(\lambda)$ contains only one element for any fixed $\lambda > 0$.
- $\forall B^*, \tilde{A}^* \in S_1(\lambda)$, one has $B^*\tilde{A}^* \in S_3(\lambda)$.
- Let $S_3(\lambda) = S_4(\lambda) = \{Z_\lambda^*\}$, then there exists $B^*\tilde{A}^* = Z_\lambda^*$ such that $(B^*, \tilde{A}^*) \in S_1(\lambda)$.

We first establish that (Problem three) has a unique solution. Our analysis is based on the following lemma.

Lemma G.6 (Cross entropy loss with nuclear norm has a unique minimizer)**.** *For any $\lambda > 0$, define*

$$\phi(Z;\lambda) = \frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\mathrm{CE}}\big(y_c,\, Ze_{(c-1)n+j}\big) + \lambda\|Z\|_*,$$

where y_c is a one-hot vector with c-th index equals one. Then $\phi(Z;\lambda)$ has a unique solution.

We refer the readers to later sections in Appendix G.4 for the proof.

Based on Lemma G.6, (Problem three) has a unique solution, denoted by Z_λ^*.

Before relating this solution to the fixed-width factorization in (Problem one), we verify its rank feasibility. Let $\Pi = \mathrm{diag}(\Pi_1, \ldots, \Pi_{\bar{K}})$, where each $\Pi_c \in \mathbb{R}^{n\times n}$ is a permutation matrix. Permuting columns within a class leaves the mean cross entropy unchanged, and $\|Z\Pi\|_* = \|Z\|_*$ by unitary invariance. Thus $\phi(Z\Pi;\lambda) = \phi(Z;\lambda)$, and uniqueness implies $Z_\lambda^*\Pi = Z_\lambda^*$ for every such Π. Consequently, columns within each class coincide, so

$$Z_\lambda^* = V_\lambda \otimes \mathbf{1}_n^\top, \qquad V_\lambda \in \mathbb{R}^{K\times\bar{K}}, \qquad \mathrm{rank}(Z_\lambda^*) \le \bar{K} \le r,$$

where the final inequality follows from Assumption 2.3.

Set $q = \mathrm{rank}(Z_\lambda^*)$. If $q > 0$, take the compact SVD $Z_\lambda^* = U_q\Sigma_q V_q^\top$, with $U_q \in \mathbb{R}^{K\times q}$ and $V_q \in \mathbb{R}^{\bar{K}n\times q}$, and define

$$B_0 = \begin{bmatrix} U_q\Sigma_q^{1/2} & \mathbf{0}_{K\times(r-q)}\end{bmatrix} \in \mathbb{R}^{K\times r},$$
$$\tilde{A}_0 = \begin{bmatrix}\Sigma_q^{1/2}V_q^\top \\ \mathbf{0}_{(r-q)\times\bar{K}n}\end{bmatrix} \in \mathbb{R}^{r\times\bar{K}n}.$$

If $q = 0$, take $B_0 = 0$ and $\tilde{A}_0 = 0$ with the same dimensions. In either case,

$$B_0\tilde{A}_0 = Z_\lambda^*, \qquad \|B_0\|_F^2 = \|\tilde{A}_0\|_F^2 = \|Z_\lambda^*\|_*.$$

By Lemma G.2, the objective of (Problem one) at any $(B, \tilde{A})$ is at least $\phi(B\tilde{A};\lambda) \ge \phi(Z_\lambda^*;\lambda)$, and $(B_0, \tilde{A}_0)$ attains this lower bound. Hence the minimum values of (Problem one) and (Problem three) agree. Every $(B^*, \tilde{A}^*) \in S_1(\lambda)$ must therefore satisfy

$$B^*\tilde{A}^* = Z_\lambda^*, \qquad 2\|Z_\lambda^*\|_* = \|B^*\|_F^2 + \|\tilde{A}^*\|_F^2.$$

Equality in $\|B^*\tilde{A}^*\|_* \le \|B^*\|_F\|\tilde{A}^*\|_F \le (\|B^*\|_F^2 + \|\tilde{A}^*\|_F^2)/2$ also gives $\|B^*\|_F^2 = \|\tilde{A}^*\|_F^2 = \|Z_\lambda^*\|_*$.

Then, one can choose $\rho_\lambda = \|Z_\lambda^*\|_*$, and based on strong duality, (Problem three) and (Problem four) admits the same set of solutions, i.e., $S_3(\lambda) = S_4(\lambda)$.

Next, we show that $\forall(B^*, \tilde{A}^*) \in S_2(\lambda)$, it must satisfy

$$B^*\tilde{A}^* \in S_4(\lambda), \qquad 2\|B^*\tilde{A}^*\|_* = \|B^*\|_F^2 + \|\tilde{A}^*\|_F^2\,.$$

This is because first, the minimum objective value of (Problem two) and (Problem four) must be equal. Given $Z_\lambda^* \in S_4(\lambda)$, one can do balanced factorization, and obtain the corresponding $B^*, \tilde{A}^*$. This implies that the minimum objective value of (Problem two) is less than or equal to that of (Problem four). On the other hand, $\forall(B^*, \tilde{A}^*) \in S_2(\lambda)$, one has $\|B^*\tilde{A}^*\|_* \le \frac{\|B^*\|_F^2 + \|\tilde{A}^*\|_F^2}{2} \le$

ρ_λ which implies $B^*\tilde{A}^*$ is a feasible solution of (Problem four). Thus, the minimum objective value of (Problem four) is less than or equal to that of (Problem two). Combine these together, we conclude that the minimum objective value of (Problem two) and (Problem four) must be equal, and $\forall(B^*, \tilde{A}^*) \in S_2(\lambda)$, $B^*\tilde{A}^* \in S_4(\lambda)$ must hold. Since $S_4(\lambda) = \{Z^*_\lambda\}$, the preceding norm inequalities are equalities, and $\|B^*\|_F^2 = \|\tilde{A}^*\|_F^2 = \rho_\lambda$.

Finally, we show $S_1(\lambda) = S_2(\lambda)$.

This is because, on one hand, $\forall B_1^*, \tilde{A}_1^* \in S_1(\lambda)$, from previous reasoning, they must satisfy

$$B_1^*\tilde{A}_1^* \in S_3(\lambda) = S_4(\lambda), \qquad \|B_1^*\|_F^2 = \|\tilde{A}_1^*\|_F^2 = \rho_\lambda\,,$$

and this indicates $B_1^*, \tilde{A}_1^* \in S_2(\lambda)$.

On the other hand, $\forall B_2^*, \tilde{A}_2^* \in S_2(\lambda)$, they must satisfy

$$B_2^*\tilde{A}_2^* \in S_3(\lambda) = S_4(\lambda), \qquad \|B_2^*\|_F^2 = \|\tilde{A}_2^*\|_F^2 = \rho_\lambda\,,$$

and this indicates $(B_2^*, \tilde{A}_2^*) \in S_1(\lambda)$. Thus, we finish the proof for $S_1(\lambda) = S_2(\lambda)$.

The last step is to show ρ_λ is a continuous and non-increasing function of λ that satisfies

$$\lim_{\rho\to 0+} \rho_\lambda = \infty, \qquad \lim_{\rho\to\infty} \rho_\lambda = 0\,.$$

Continuity of ρ_λ. In this part, we will use Lemma G.4 to prove the solution to (Problem three) is continuous, which implies that ρ_λ is a continuous function w.r.t. λ. Notice Lemma G.4 is presented for maximization problem, one can replace the objective f to $-f$ to extend it to minimization problem. Moreover, it is obvious that the objective function is continuous, convex, and the set of minimizers is single-valued for any fixed $\lambda > 0$. Thus, in order to apply Lemma G.4 to show the continuity of ρ_λ, we only need to check $C(\lambda)$, which is the range of Z, is a compact-valued correspondence. At first glance, this seems wrong since Z can take any values in $\mathbb{R}^{K\times \bar{K}n}$ in (Problem three). However, we will show that one can constrain the domain of Z which leads to the same minimizer. Notice that when $Z = \mathbf{0}$, the objective in (Problem three) takes the value $\log K$. Thus, we can choose $C(\lambda) = \left\{Z \middle| \|Z\|_* \le \frac{\log K}{\lambda}\right\}$, which is a compact set. Therefore, for any fix λ, one can apply Lemma G.4 to show Z^*_λ is continuous w.r.t. λ, which implies $\rho_\lambda := \|Z^*_\lambda\|_*$ is also continuous w.r.t. λ.

ρ_λ is a non-increasing function. For any $0 < \lambda_1 < \lambda_2$, let $Z_{\lambda_1}, Z_{\lambda_2}$ be the solutions for (Problem three) with the corresponding regularization penalty. Then, our first observation is

$$\begin{aligned}
&\phi(Z_{\lambda_2};\lambda_2) \le \phi(Z_{\lambda_1};\lambda_2)\\
\iff &\phi(Z_{\lambda_2};\lambda_1) + (\lambda_2-\lambda_1)\|Z_{\lambda_2}\|_* \le \phi(Z_{\lambda_1};\lambda_1) + (\lambda_2-\lambda_1)\|Z_{\lambda_1}\|_*\\
\iff &\phi(Z_{\lambda_2};\lambda_1) - \phi(Z_{\lambda_1};\lambda_1) \le (\lambda_2-\lambda_1)(\|Z_{\lambda_1}\|_* - \|Z_{\lambda_2}\|_*)\\
\iff &\phi(Z_{\lambda_2};\lambda_1) - \phi(Z_{\lambda_1};\lambda_1) \le (\lambda_2-\lambda_1)(\rho(\lambda_1) - \rho(\lambda_2))\,.
\end{aligned}$$

Notice $Z_{\lambda_1} \in \arg\min \phi(Z;\lambda_1)$, thus, we have $\phi(Z_{\lambda_2};\lambda_1) - \phi(Z_{\lambda_1};\lambda_1) \ge 0$, and it leads to $\rho(\lambda_1) \ge \rho(\lambda_2)$. Therefore, ρ_λ is non-increasing in λ. If $\rho(\lambda_1) = \rho(\lambda_2)$, the preceding comparison and uniqueness imply $Z_{\lambda_1} = Z_{\lambda_2} =: Z$. Write $\ell(Z) := \phi(Z;0)$. By first-order optimality, choose $G_i \in \partial\|Z\|_*$ such that

$$\nabla\ell(Z) + \lambda_i G_i = 0, \qquad i = 1, 2.$$

Subtracting these optimality conditions gives

$$\lambda_1 G_1 = \lambda_2 G_2. \tag{118}$$

Every nuclear-norm subgradient satisfies $\langle G_i, Z\rangle = \|Z\|_*$. Taking Frobenius inner products of (118) with Z therefore yields

$$(\lambda_1 - \lambda_2)\|Z\|_* = 0.$$

Since $0 < \lambda_1 < \lambda_2$, this forces $\|Z\|_* = 0$, hence $Z = 0$. Thus, when $\rho(\lambda_1) = \rho(\lambda_2)$ holds, one must have $Z_{\lambda_1} = Z_{\lambda_2} = 0$.

Asymptotic behaviour of ρ_λ. In this part, we aim to show

$$\lim_{\lambda\to 0}\rho_\lambda=\infty,\ \lim_{\lambda\to\infty}\rho_\lambda=0\,.$$

We first use proof by contradiction to show that $\lim_{\lambda\to\infty}\rho_\lambda=0$. Assume there exists $M>0$ such that $\forall\lambda>0$, one has $\rho_\lambda\geq M$. Then, we take a series λ_i such that $\lim_{i\to\infty}\lambda_i=\infty$, and let $S=\{\rho(\lambda_i)\}_{i=1}^{\infty}$. Based on the assumption, we know all the elements in S has a lower bound M. Moreover, we know $\forall\lambda_i$, $\rho(\lambda_i)\leq\frac{\log K}{\lambda_i}$. This is because $\phi(0;\lambda)=\log K$ which is independent of the choices of λ_i. Thus, there exists M' such that $\forall\rho_\lambda\in S$, we have $\rho_\lambda\leq M'$. Based on Bolzano–Weierstrass theorem, there exists a subsequence λ_{i_k} such that $\lim_{k\to\infty}\rho(\lambda_{i_k})=M^*$ where $M\leq M^*\leq M'$. However, in this case, $\lim_{k\to\infty}\phi(Z^*_{\lambda_{i_k}};\lambda_{i_k})\geq\lim_{k\to\infty}\lambda_{i_k}\rho(\lambda_{i_k})=\infty$. This is in contradictory to the assumption that $Z^*_{\lambda_{i_k}}$ minimizes $\phi(Z;\lambda_{i_k})$ and $\phi(0;\lambda)=\log K$. Thus, one must have $\lim_{\lambda\to\infty}\rho_\lambda=0$.

Next, we first use proof by contradiction to show that $\lim_{\lambda\to 0}\rho_\lambda=\infty$. Assume there exists $N>0$ such that $\forall\lambda>0$, one has $\rho_\lambda\leq N$. For every finite $R>0$, $\phi(Z;0)$ is continuous and strictly positive on the compact nuclear-norm ball $\{Z:\|Z\|_*\leq R\}$. Consequently,

$$\epsilon_R:=\min_{\|Z\|_*\leq R}\phi(Z;0)>0\,.$$

Then, we pick a series $\lambda_k=\frac{1}{k^2}$, $R_k=k$, and choose

$$\tilde{Z}_k=k\begin{pmatrix}\mathbf{1}_n^\top & \mathbf{0}_n^\top & \mathbf{0}_n^\top & \cdots, & \mathbf{0}_n^\top\\ \mathbf{0}_n^\top & \mathbf{1}_n^\top & \mathbf{0}_n^\top & \cdots & \mathbf{0}_n^\top\\ \mathbf{0}_n^\top & \mathbf{0}_n^\top & \cdots & \mathbf{0}_n^\top & \mathbf{1}_n^\top\\ & & \mathbf{0}_{(K-\bar{K})\times\bar{K}n} & & \end{pmatrix}$$

Based on the optimality of $Z^*_{\lambda_k}$, one must have

$$\phi(Z^*_{\lambda_k};\lambda_k)\leq\phi(\tilde{Z}_k;\lambda_k)=\log\bigl(1+(K-1)\exp(-k)\bigr)+\frac{\bar{K}\sqrt{n}}{k}\,.$$

On the other hand,

$$\phi(Z^*_{\lambda_k};\lambda_k)\geq\phi(Z^*_{\lambda_k};0)\geq\epsilon_N>0\,.$$

Combine these two inequalities together, one has

$$\epsilon_N\leq\log\bigl(1+(K-1)\exp(-k)\bigr)+\frac{\bar{K}\sqrt{n}}{k}\,.$$

However, one can choose k sufficiently large that the above inequality breaks. Thus, there cannot exist a $N>0$ such that $\rho_\lambda\leq N$ holds $\forall\lambda>0$. Therefore, one must have $\lim_{\lambda\to 0}\rho_\lambda=\infty$. □

G.4 Proof of Lemma G.6

Proof. Our starting point is the following lemma which is developed in Hong and Ling [2023].

Lemma G.7 (Cross entropy loss is locally strongly convex in restricted direction)**.** *Define*

$$\phi(Z)=\frac{1}{\bar{K}n}\sum_{c=1}^{\bar{K}}\sum_{j=1}^{n}\mathcal{L}_{\mathrm{CE}}\bigl(y_c,\,Ze_{(c-1)n+j}\bigr)\,,$$

where y_c is a one-hot vector with c-th entry equal to one. For any $R>0$, there exists $\mu(R)>0$ such that on the compact set $\{Z\in\mathbb{R}^{K\times\bar{K}n}:\|Z\|_F\leq R\}$, $\phi(Z)$ is $\mu(R)$-strongly convex in any direction $\Delta_Z\in\{\Delta_Z:\mathbf{1}_K^\top\Delta_Z=0\}$.

This bounded-set statement follows from the pointwise restricted Hessian bound in the proof of Lemma 5.1 in Hong and Ling [2023].

Fix $\lambda>0$ and write $\phi(Z;\lambda):=\phi(Z)+\lambda\|Z\|_*$. Since cross entropy is nonnegative,

$$\phi(Z;\lambda)\geq\lambda\|Z\|_*\geq\lambda\|Z\|_F.$$

Hence $\phi(\cdot;\lambda)$ is continuous and coercive, so it attains a global minimum. Every global minimizer Z^* satisfies

$$\lambda\|Z^*\|_F \le \phi(Z^*;\lambda) \le \phi(0;\lambda) = \log K,$$

and therefore $\|Z^*\|_F \le \log K/\lambda$.

We next show that every global minimizer is centered. Let Z^* be any global minimizer and decompose it as

$$Z^* = \frac{1}{K}1_K 1_K^\top Z^* + (I_K - \frac{1}{K}1_K 1_K^\top)Z^*\,.$$

For simplicity, let $P = I_K - \frac{1}{K}1_K 1_K^\top$ be the projection matrix onto the space orthogonal to 1_K. The removed component $\frac{1}{K}1_K 1_K^\top Z^*$ has constant entries within each column. Since cross entropy is invariant under such logit shifts, $\phi(Z^*) = \phi(PZ^*)$. By the contraction property in Lemma G.3 applied to the orthogonal projection P,

$$\|PZ^*\|_* \le \|Z^*\|_*\,,$$

with equality if and only if $PZ^* = Z^*$. Combined with $\phi(Z^*) = \phi(PZ^*)$, we have $\phi(PZ^*;\lambda) \le \phi(Z^*;\lambda)$, and optimality of Z^* forces equality, hence $PZ^* = Z^*$ and

$$\mathbf{1}_K^\top Z^* = 0\,.$$

Finally, choose $R > \log K/\lambda$. All global minimizers belong to the convex set

$$\mathcal{B}_R := \{Z \in \mathbb{R}^{K\times \bar{K}n} : \mathbf{1}_K^\top Z = 0,\ \|Z\|_F \le R\}.$$

By Lemma G.7, $\phi(Z)$ is $\mu(R)$-strongly convex on $\mathcal{B}_R$. Since the nuclear norm is convex, the regularized objective $\phi(Z;\lambda)$ is also $\mu(R)$-strongly convex on this set. Thus, for any two global minimizers Z_1, Z_2,

$$\phi\left(\frac{Z_1+Z_2}{2};\lambda\right) \le \frac{\phi(Z_1;\lambda)+\phi(Z_2;\lambda)}{2} - \frac{\mu(R)}{8}\|Z_1 - Z_2\|_F^2.$$

If $Z_1 \neq Z_2$, their midpoint would have a strictly smaller objective value, contradicting global optimality. Therefore the global minimizer is unique. □

H Experiment

In this section, we present the detailed setup of the experiments in §4.

Code availability. Code is available at `https://github.com/simonxu97/MaLoRA`.

Hardware. All experiments are run on a single NVIDIA RTXA5000 GPU (24GB) per (benchmark, seed), allocated through an academic SLURM cluster. The backbones—ViT-Base/16 pre-trained on ImageNet-21k and XLM-RoBERTa Base—are loaded directly from public checkpoints; we do not pretrain any model from scratch. Per-domain LoRA finetuning runs for 20 fixed epochs with AdamW and takes 5–30 minutes per domain depending on the dataset, totalling roughly 0.5–3GPU-hours per (benchmark, seed) across all per-domain adapters and the pretrain-task adapter. Computational overhead and merge-selection runtime are reported in Appendix H.2.2, using a dedicated matched-budget timing protocol.

H.1 Details for experiments in §H.1.2 and §4.1

Definition of superclasses of CIFAR-100. CIFAR-100 groups its 100 fine categories into **20 coarse *superclasses*** as follows

Construction of pre-training and fine-tuning tasks. For every superclass S, let $[c_1^S, \dots, c_5^S]$ be a fixed random permutation of its five fine labels. We build three disjoint labelled sets

$$\mathcal{D}_{\text{pre}} = \bigcup_S \{c_1^S, c_2^S, c_3^S\}, \quad \mathcal{D}_1 = \bigcup_S \{c_4^S\}, \quad \mathcal{D}_2 = \bigcup_S \{c_5^S\}.$$

Hence $\mathcal{D}_{\text{pre}}$ contains 60 fine classes (three per super-class) while each fine-tuning task $\mathcal{D}_i$ contains exactly one *new* class per superclass, preserving maximal diversity yet zero overlap with $\mathcal{D}_{\text{pre}}$.

Table 2: CIFAR-100 coarse superclasses and their five fine labels.

Superclass	Fine classes (5 per superclass)
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchid, poppy, rose, sunflower, tulip
food containers	bottle, bowl, can, cup, plate
fruit & vegetables	apple, mushroom, orange, pear, sweet pepper
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores & herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Frozen feature extractors. We evaluate four widely used backbones, noting their different pre-training dataset:

- **ResNet-50** [He et al., 2016] (torchvision, *supervised ImageNet-1K*).
- **ViT-Base/16** [Dosovitskiy et al., 2021] (timm: supervised pre-training on *ImageNet-21K*, then fine-tuned on ImageNet-1K).
- **ConvNeXt-Tiny** [Liu et al., 2022] pre-trained on ImageNet-12K, then fine-tuned on ImageNet-1K.
- **CLIP ViT-B/32** [Radford et al., 2021] (OpenAI's contrastive pre-training on *web-scale image–text* pairs; no ImageNet supervision).

All models are kept frozen; we extract the CLS token (ViT/CLIP) or the global-average-pooled penultimate tensor (CNNs) at 224×224 resolution for every CIFAR-100 image.

Pre-training stage. A linear classifier $W_{\text{pre}} \in \mathbb{R}^{20\times d}$ (d = embedding dimension) is trained from scratch on $\mathcal{D}_{\text{pre}}$ for 600 epochs with Adam ($\eta = 0.2$). The 20 rows correspond to the super-classes, not the 60 fine labels; this matches our theoretical model where each task's labels share a common output.

H.1.1 Empirical Motivation for the Task Geometry Assumption

In this section, we conduct a comprehensive feature correlation analysis on CIFAR-100. This analysis quantifies both intra-class and inter-class correlations in the feature space to provide empirical motivation for the near-orthogonal task geometry in Assumption 2.1.

Let $\mathcal{F}_\theta(x)$ denote the feature extraction function of a neural network with parameters θ, where $x \in \mathbb{R}^{3\times32\times32}$ represents an input image from CIFAR-100. For each architecture (ResNet-50, ViT-Base/16, CLIP ViT-B/32, and ConvNeXt-Tiny), we extract features from the penultimate layer, obtaining feature vectors $f_i = \mathcal{F}_\theta(x_i) \in \mathbb{R}^d$ for each training sample x_i.

We organize the extracted features by class, creating sets $\mathcal{S}_c = \{f_i : y_i = c\}$ for each class $c \in \{0, 1, \ldots, 99\}$, where y_i is the ground truth label of sample x_i.

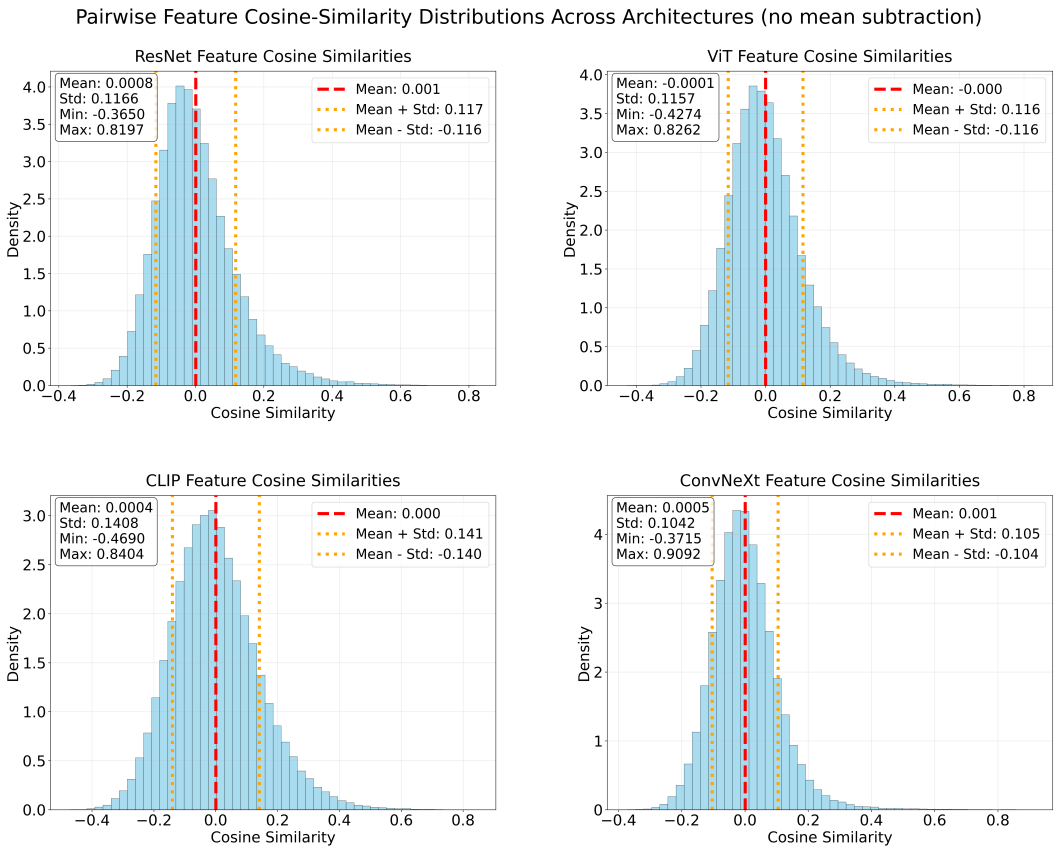


Figure 2: Distributions of pairwise feature correlation across four pre-trained models. The red dashed line represent the mean value of correlations, and the dashed lines represent mean plus and minus one std.

Cosine-Similarity Computation. For any two feature vectors $f_i, f_j \in \mathbb{R}^d$, we compute the cosine similarity:

$$s(f_i, f_j) = \frac{\langle f_i, f_j \rangle}{\|f_i\|_2 \, \|f_j\|_2} = \frac{\sum_{k=1}^{d} f_i^{(k)} f_j^{(k)}}{\sqrt{\sum_{k=1}^{d} \left(f_i^{(k)}\right)^2} \sqrt{\sum_{k=1}^{d} \left(f_j^{(k)}\right)^2}}, \tag{119}$$

where $f_i^{(k)}$ denotes the k-th coordinate of f_i.

Intra-class Similarity. For each class c, we compute the average intra-class cosine similarity:

$$s_{\text{intra}}(c) = \frac{1}{|\mathcal{S}_c|(|\mathcal{S}_c| - 1)/2} \sum_{f_i, f_j \in \mathcal{S}_c,\ i<j} s(f_i, f_j). \tag{120}$$

This measures how aligned features are within the same class, indicating the consistency of learned representations for semantically similar samples.

Inter-class Similarity. For any two distinct classes $c_1 \neq c_2$, we compute the average inter-class cosine similarity:

$$s_{\text{inter}}(c_1, c_2) = \frac{1}{|\mathcal{S}_{c_1}| \cdot |\mathcal{S}_{c_2}|} \sum_{f_i \in \mathcal{S}_{c_1}} \sum_{f_j \in \mathcal{S}_{c_2}} s(f_i, f_j). \tag{121}$$

This quantifies the alignment between features from different classes, with values closer to zero indicating more orthogonal (i.e., better separated) classes.

	CLIP	RESNET	VIT	CONVNEXT
Intra-class mean (std)	0.2713(0.0769)	0.2312(0.0689)	0.2536(0.0652)	0.2731(0.0784)
Inter-class mean (std)	-0.0024(0.0670)	-0.0021(0.0621)	-0.0021(0.0563)	-0.0027(0.0439)

Table 3: Mean and standard deviation of intra-class and inter-class feature cosine similarities. Values closer to zero indicate greater orthogonality between features.

Similarity Matrix Construction. We construct a 100×100 similarity matrix $\mathbf{S}$ where:

$$S_{ij} = \begin{cases} s_{\text{intra}}(i) & \text{if } i = j \\ s_{\text{inter}}(i,j) & \text{if } i \neq j. \end{cases} \tag{122}$$

Implementation Details. To ensure computational efficiency while maintaining statistical reliability, we subsample feature pairs for similarity computation. Specifically, for each class we randomly sample 100 points (10,000 in total). We first report the distribution of pairwise cosine similarities across all sampled points in Figure 2. To further distinguish intra-class and inter-class relationships, we visualize the corresponding similarities as a heatmap in Figure 3. For intra-class similarities, we randomly sample up to 1,000 distinct sample pairs per class (drawn without replacement to exclude trivial self-pairs whose cosine would be 1); for inter-class similarities, we sample up to 1,000 pairs for each class pair. This subsampling strategy yields robust similarity estimates while keeping the computation tractable.

Figure 2 highlights two key observations: (1) most pairwise cosine similarities are concentrated around zero; and (2) the mean similarity $\pm$ one standard deviation remains below 0.15 for every architecture. These descriptive statistics provide qualitative motivation for studying a near-orthogonal reference geometry.

Figure 3 highlights three observations: (1) higher diagonal values indicate stronger within-class alignment; (2) near-zero off-diagonal values indicate weak average inter-class alignment; and (3) Table 3 quantifies this contrast, with inter-class means close to zero and intra-class means around 0.23–0.27. The weak average inter-class and within-class alignment provides qualitative motivation for studying a near-orthogonal reference geometry. Extending the analysis to richer within-class structure and stronger correlations remains an important direction for future work.

H.1.2 Effect of regularization

In this section, we numerically verify Theorem 3.2.

We first pre-train a linear classifier on the pretraining task, and then fine-tune it using LoRA across 50 logarithmically spaced regularization strengths, with $\lambda \in [10^{-5}, 10^{-1}]$. Figure 4 shows that LoRA's performance exhibits a non-monotonic relationship with λ, peaking at a moderate value. For ViT and ResNet, the theoretically predicted optimal λ is within a factor of 1.8 of the best empirical choice for every $\bar{K}$, and for ConvNeXt within a factor of 1.2 when $\bar{K} \in \{10, 20\}$; for CLIP, the empirical optimum is 1.8–4.4 times larger than predicted. Detailed comparisons are reported in Appendix H.

For each fine-tuning task, we freeze the pre-trained weights W_{pre} and attach a rank-r LoRA block, where r is set to the number of classes in the fine-tuning task. This satisfies Assumption 2.3. We evaluate performance across three fine-tuning tasks with varying class counts: $\bar{K} = 5, 10, 20$. Each task is constructed by selecting the top $\bar{K}$ superclasses from $\mathcal{D}_2$.

We fine-tune the model on each task using SGD for 3000 steps with a learning rate of 0.5, sweeping over regularization values $\lambda \in [10^{-5}, 10^{-1}]$ with 50 logarithmically spaced regularization strengths. For each $\bar{K}$ and each frozen feature extractor, we record both the empirically optimal regularization parameter and the theoretically predicted one from Theorem 3.2. The detailed figures are given in Table 4.

H.1.3 Details for experiments in §4.1

For the experiments in §4.1, we fine-tune $\mathcal{D}_1, \mathcal{D}_2$ using different LoRA adapters $(B_i, A_i), i = 1, 2$ with LoRA rank 20. We fine-tune the LoRA adapters using Adam ($\eta = 0.1$) with regularization parameter $\lambda = 5 \times 10^{-7}$. Let $(B_i^*, A_i^*), i = 1, 2$ be the optimal LoRA adapters we achieve at the

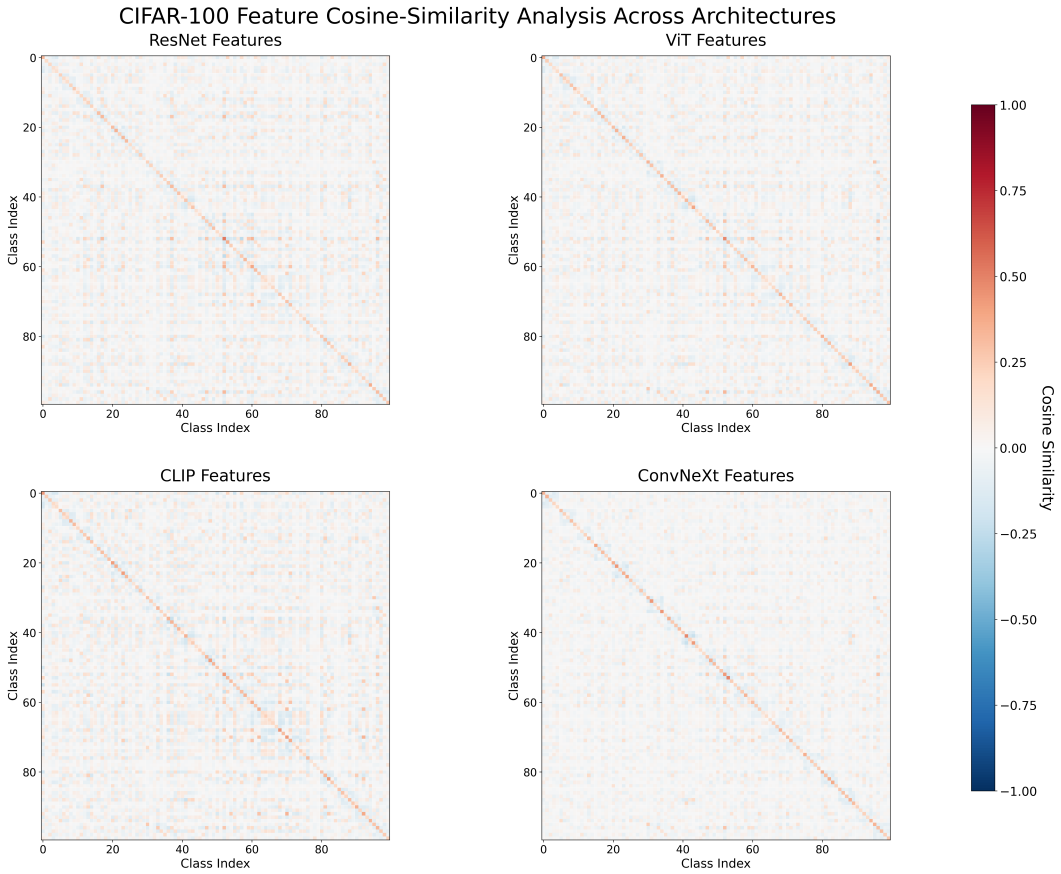


Figure 3: Feature correlation analysis across different pretrained models.

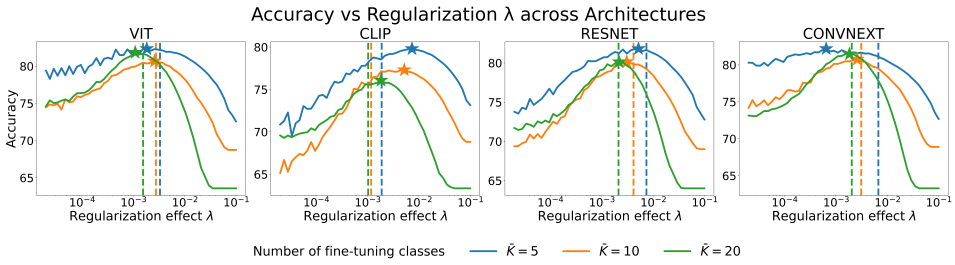


Figure 4: Accuracy of the fine-tuned model across varying regularization levels for four different pre-trained models. Each line represents a fine-tuning task, indicated by color. Stars highlight the best-performing regularization value for each task. Vertical dashed lines indicate the theoretically predicted optimal regularization parameter for each setting. Alignment between the star and the dashed line of the same color reflects how well our theory predicts the empirically optimal regularization level.

end of fine-tuning, we merge them together as follows

$$W_{\mathrm{LoRA}}^{\lambda}(\alpha_1, \alpha_2) = W_{\mathrm{pre}} + \alpha_1 B_1^* A_1^* + \alpha_2 B_2^* A_2^*$$

To seek for the optimal mixing coefficients (α_1, α_2), we run grid search over a 50×50 lattice on $(0, 1)^2$, and compare it with the optimal theoretical mixing coefficients presented in Theorem 3.3. We test different number of classes in the fine-tuning datasets, i.e., $\bar{K} = 5, 10, 20$, and reports their results here.

Table 4: Empirical (Emp) and theoretical (Thm) optimal regularization parameters λ for different pre-trained models and fine-tuning task sizes ($\bar{K}$).

Model	$\bar{K} = 5$	$\bar{K} = 10$	$\bar{K} = 20$
`ResNet-50`	Emp: 0.005208 Thm: 0.007424	Emp: 0.003092 Thm: 0.004158	Emp: 0.002184 Thm: 0.002121
`ViT-B/16`	Emp: 0.001835 Thm: 0.003304	Emp: 0.002599 Thm: 0.002763	Emp: 0.001090 Thm: 0.001551
`ConvNeXt`	Emp: 0.000647 Thm: 0.006709	Emp: 0.002599 Thm: 0.003114	Emp: 0.001835 Thm: 0.002039
`CLIP`	Emp: 0.007373 Thm: 0.001889	Emp: 0.005208 Thm: 0.001178	Emp: 0.001835 Thm: 0.001030

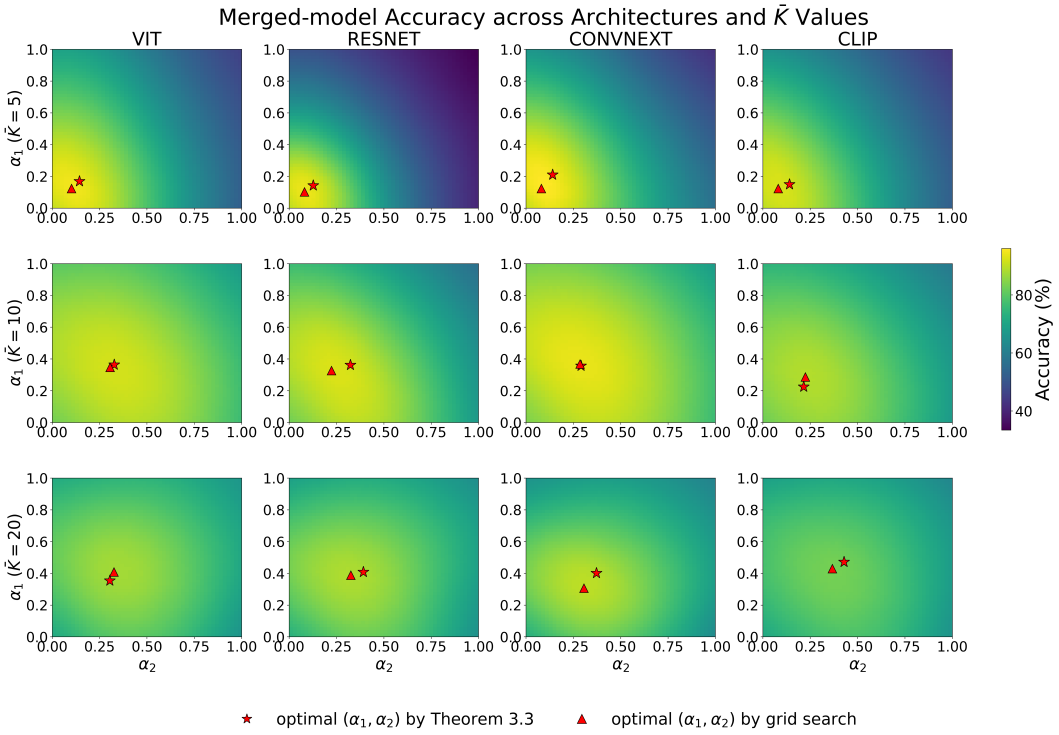


Figure 5: Merged-model accuracy across architectures and different number of classes in the fine-tuning tasks. Each panel shows the accuracy of the merged model evaluated on the combined dataset, across a 50×50 grid of mixing coefficients $(\alpha_1, \alpha_2) \in (0, 1)^2$. For each architecture, the red star indicates the theoretically predicted optimal coefficients from Theorem 3.3, while the red triangle marks the empirically optimal coefficients.

H.2 Details for experiments in §4.2

In §4.2, the datasets tested are all multi-domain benchmarks (See Table 7 for details).

Table 5: Average normalized accuracy (%) following Ilharco et al. [2022]: each task's merged-model accuracy is divided by the per-domain finetuned model's accuracy on that task, then averaged across domains within each benchmark. Mean $\pm$ std over 10 seeds. **Bold** = best per column. Values >100 indicate the merged model outperforms the per-task finetuned model on that benchmark.

Method	PACS	VLCS	Terra Inc.	MNLI	Amazon	CivilComm.	Avg
Soup	96.38 ± 0.77	96.89 ± 0.47	64.31 ± 2.53	101.16 ± 0.19	100.64 ± 0.54	$\mathbf{98.27 \pm 1.66}$	92.94 ± 0.53
TA	97.52 ± 0.97	97.19 ± 0.96	64.46 ± 2.65	$\mathbf{101.27 \pm 0.24}$	100.68 ± 0.64	97.80 ± 1.69	93.15 ± 0.58
TIES	97.52 ± 0.86	96.95 ± 0.72	63.78 ± 2.36	100.87 ± 0.21	100.64 ± 0.54	98.00 ± 1.98	92.96 ± 0.55
DARE-TIES	$\mathbf{97.57 \pm 0.85}$	97.23 ± 0.96	64.23 ± 2.27	100.90 ± 0.23	100.45 ± 0.65	97.92 ± 1.94	93.05 ± 0.55
DARE-Linear	96.48 ± 0.77	96.77 ± 0.51	64.73 ± 2.34	101.12 ± 0.21	100.50 ± 0.57	97.61 ± 1.88	92.87 ± 0.53
MaLoRA	96.69 ± 0.83	$\mathbf{97.54 \pm 0.62}$	$\mathbf{65.75 \pm 2.31}$	101.23 ± 0.22	$\mathbf{100.69 \pm 0.56}$	98.00 ± 1.99	$\mathbf{93.32 \pm 0.55}$

Table 6: Average relative accuracy drop (%), the complement of normalized accuracy ($1 -$ merged/oracle, averaged across domains within each benchmark and across benchmarks). Mean $\pm$ std over 10 seeds. **Bold** = smallest per column (least forgetting). Negative values indicate the merged model outperforms the per-task finetuned model on average. Avg is the equally weighted mean across benchmarks; its std is $\sqrt{\sum_b \sigma_b^2}/6$, where σ_b is the per-benchmark std, assuming independent benchmarks.

Method	PACS	VLCS	Terra Inc.	MNLI	Amazon	CivilComm.	Avg
Soup	3.62 ± 0.77	3.11 ± 0.47	35.69 ± 2.53	-1.16 ± 0.19	-0.64 ± 0.54	$\mathbf{1.73 \pm 1.66}$	7.06 ± 0.53
TA	2.48 ± 0.97	2.81 ± 0.96	35.54 ± 2.65	$-1.27 \pm \mathbf{0.24}$	-0.68 ± 0.64	2.20 ± 1.69	6.85 ± 0.58
TIES	2.48 ± 0.86	3.05 ± 0.72	36.22 ± 2.36	-0.87 ± 0.21	-0.64 ± 0.54	2.00 ± 1.98	7.04 ± 0.55
DARE-TIES	$\mathbf{2.43 \pm 0.85}$	2.77 ± 0.96	35.77 ± 2.27	-0.90 ± 0.23	-0.45 ± 0.65	2.08 ± 1.94	6.95 ± 0.55
DARE-Linear	3.52 ± 0.77	3.23 ± 0.51	35.27 ± 2.34	-1.12 ± 0.21	-0.50 ± 0.57	2.39 ± 1.88	7.13 ± 0.53
MaLoRA	3.31 ± 0.83	$\mathbf{2.46 \pm 0.62}$	$\mathbf{34.25 \pm 2.31}$	-1.23 ± 0.22	$-0.69 \pm \mathbf{0.56}$	2.00 ± 1.99	$\mathbf{6.68 \pm 0.55}$

Table 7: Benchmarks used in the main paper. "Domains" lists every domain in the dataset; the *pretrain* column marks the one folded into `base_with_pretrain`. "# merged" is the number of LoRA adapters that participate in the dense task-arithmetic merge (i.e. all non-pretrain adapters).

Dataset	Modality	# domains	# classes	# merged	Domains (pretrain in **bold**)
PACS	vision	4	7	3	**art painting**, cartoon, photo, sketch
VLCS	vision	4	5	3	Caltech101, LabelMe, SUN09, **VOC2007**
Terra Incognita	vision	4	10	3	location_38, location_43, location_46, **location_100**
MNLI	language	5	3	4	**fiction**, government, slate, telephone, travel
Amazon Reviews	language	4	2	3	**books**, electronics, kitchen, movies
CivilComments	language	5	2	4	identity_attack, **insult**, obscene, sexual_explicit, threat

Validation/test splits. For every domain, we hold out a **stratified** 10% **validation split** and a **stratified** 10% **test split** from the per-domain training set, with the remaining 80% used for finetuning. The split indices are deterministic in the configured random seed and are reused across all methods so that Soup, Task Arithmetic, and MaLoRA are evaluated on the same per-domain val/test sets within a seed. Validation accuracy averaged over all merged domains is the criterion for selecting the merge coefficient λ^*.

LoRA configuration. We use LoRA Hu et al. [2022] with the following layout, applied identically to every adapter (per-domain, per-seed):

- **Target modules:** `query` and `value` projections only (no key, no MLP, no embeddings, no normalization).
- **Layers transformed:** the last four transformer layers, indexed $\{8, 9, 10, 11\}$ (out of 12). This was identified as the best layer placement in our preliminary layer-sweep ablation.

- **Scaling factor:** $\alpha_{\text{LoRA}} = 16$ for all adapters.
- **LoRA dropout:** 0.05.
- **Classifier head:** a fresh classification head is initialized for each adapter and trained jointly with the LoRA parameters. At merge time, each merged adapter contributes *both* its classifier head and its LoRA delta to the weighted sum, so that the merged model has a coherent classifier without an external head-aggregation step.

Training and per-dataset hyperparameters. All adapters are trained with AdamW for 20 epochs (no early stopping; `lora_patience = 0`). Before enabling the LoRA branch, we warm up the classifier head for 5 epochs at learning rate 3×10^{-4} (`head_epochs = 5`, `head_lr = 3e-4`); each domain trains its own head (`train_shared_head = false`). We use default PyTorch gradient clipping. The four hyperparameters that vary across datasets, i.e., LoRA rank r, learning rate, batch size, and weight decay, are summarized in Table 8.

Table 8: Per-dataset training hyperparameters. All adapters use AdamW with LoRA $\alpha = 16$, dropout 0.05, query and value projections of the final four transformer layers, and 20 training epochs. The *pretrain domain* is the one folded into the `base_with_pretrain` model, with the remaining domains fine-tuned on top.

Dataset	Pretrain domain	LoRA rank r	Learning rate	Batch size	Weight decay
PACS	art painting	4	5×10^{-4}	32	5×10^{-3}
VLCS	VOC2007	4	5×10^{-4}	32	5×10^{-2}
Terra Incognita	location_100	4	5×10^{-4}	32	5×10^{-3}
MNLI	fiction	8	5×10^{-4}	32	5×10^{-3}
Amazon Reviews	books	8	5×10^{-4}	128	5×10^{-3}
CivilComments	insult	32	5×10^{-4}	32	5×10^{-3}

Baseline methods.. We compare the following methods for merging the LoRA adapters,

- **Uniform Soup** Wortsman et al. [2022]: $\alpha_i = 1/N$ (equal weight across all merged adapters). Equivalent to Task Arithmetic at $\lambda = 1/N$.
- **Task Arithmetic (TA)** Ilharco et al. [2022]: $\alpha_i = \lambda$ for every i, with λ tuned on validation.
- **DARE-Linear** Yu et al. [2024]: each task vector ΔW_i is randomly sparsified by a Bernoulli mask with keep rate d and rescaled by $1/d$, then linearly averaged with global scale λ, i.e. $\Delta W_{\text{merged}} = (\lambda/N) \sum_i \Delta \tilde{W}_i$ where $\Delta \tilde{W}_i$ is the dropped-and-rescaled task vector. Both d and λ are tuned on validation.
- **DARE-TIES** Yu et al. [2024]: applies the DARE drop-and-rescale step above to each task vector, then aggregates via the TIES sign-election and disjoint-merge procedure with global scale λ. Both d and λ are tuned on validation.
- **TIES** Yadav et al. [2023]: each task vector is first trimmed to its top-d fraction of parameters by magnitude; per parameter, the dominant sign is elected across adapters by total signed magnitude; the merged delta averages only the sign-aligned values and is scaled by λ. Both d and λ are tuned on validation.

All baselines that involve density and λ are tuned over $d \in \{0.2, 0.3, \ldots, 0.9\}$ and $\lambda \in \{0.1, 0.2, \ldots, 1.0\}$ (80 configurations); Task Arithmetic uses the same λ grid (10 configurations).

H.2.1 Feature correlations across domains

We extend the CIFAR-100 analysis in Appendix H.1.1 to the six multi-domain benchmarks in §4.2. For this analysis, we extract features using the stock pre-trained checkpoints `google/vit-base-patch16-224-in21k` for PACS, VLCS, and Terra Incognita, and `xlm-roberta-base` for MNLI, Amazon Reviews, and CivilComments. The backbones are kept fixed and used for inference only. Each domain is treated as a task, and the domain counts include both the pre-training domain and all fine-tuning domains.

In Table 9, cross-domain mean correlations range from -0.060 to 0.060, providing qualitative motivation for weak average cross-task alignment. Within domains, small mean correlations for

Table 9: Feature correlations within and across classes and domains. Entries report mean correlations, with $\pm$ denoting spread across class–domain cell pairs in each group. A cell is a class–domain combination.

Dataset	# classes	# domains	Same class		Different classes	
			Same domain	Different domains	Same domain	Different domains
PACS	7	4	$+0.376 \pm 0.155$	$+0.020 \pm 0.157$	$+0.139 \pm 0.100$	-0.060 ± 0.090
VLCS	5	4	$+0.183 \pm 0.169$	$+0.029 \pm 0.047$	$+0.008 \pm 0.042$	-0.019 ± 0.029
Terra Incognita	10	4	$+0.312 \pm 0.143$	$+0.060 \pm 0.085$	$+0.059 \pm 0.152$	-0.033 ± 0.095
MNLI	3	5	$+0.145 \pm 0.132$	-0.018 ± 0.048	$+0.143 \pm 0.131$	-0.019 ± 0.048
Amazon Reviews	2	4	$+0.073 \pm 0.024$	-0.003 ± 0.018	$+0.039 \pm 0.023$	-0.026 ± 0.020
CivilComments	2	5	$+0.053 \pm 0.024$	$+0.018 \pm 0.020$	$+0.044 \pm 0.020$	$+0.011 \pm 0.018$
Mean			$\mathbf{+0.191}$	$\mathbf{+0.018}$	$\mathbf{+0.072}$	$\mathbf{-0.025}$

different-class pairs on VLCS (0.008) and for both same-class and different-class pairs on Amazon Reviews and CivilComments (0.039–0.073) also motivate a near-orthogonal reference geometry. Stronger same-class alignment on PACS (0.376) and Terra Incognita (0.312) highlights structured departures from this geometry, showing that agreement with the reference geometry varies across benchmarks and pair categories.

H.2.2 Computational overhead

We measure end-to-end merge-selection runtime on all six benchmarks over ten random seeds, using one NVIDIA RTX A5000 GPU and a fixed budget of four CPU threads. For the six methods compared in Table 10, we use a matched budget of 49 configurations per method, separately from the accuracy-selection grids described above. The timings measure merging and coefficient selection from trained adapters; adapter-training costs are reported separately in the hardware description.

MaLoRA estimates margins with one forward pass over each domain's validation split, costing 1.05–1.48 times a single validation pass. The resulting margins are computed once and reused throughout the global-scale search. For fixed adapters and validation sets, this additional margin-computation cost is independent of the number of candidate scales.

Table 10 reports the mean runtime for each benchmark and the arithmetic mean across the six benchmarks. MaLoRA's aggregate mean runtime is 972.6 s, compared with 950.3 s for TA, an increase of 22.3 s or approximately 2.35%. Its aggregate mean runtime is lower than those of DARE-Linear, DARE-TIES, and TIES (980.2–991.0 s).

Table 10: Mean end-to-end merge-selection runtime (seconds) over ten seeds per benchmark under the matched-budget timing protocol. Avg. is the arithmetic mean across the six benchmarks; the final column divides each method's Avg. by MaLoRA's.

Method	PACS	VLCS	Terra Inc.	MNLI	Amazon	CivilComm.	Avg.	Ratio
Soup	287.6	752.2	904.1	3207.7	253.6	297.4	950.4	$0.98\times$
TA	287.6	752.8	903.8	3206.7	253.5	297.2	950.3	$0.98\times$
DARE-Linear	316.4	784.1	937.1	3230.9	280.4	332.4	980.2	$1.01\times$
DARE-TIES	323.2	787.3	940.8	3242.7	285.2	340.0	986.5	$1.01\times$
TIES	324.9	790.6	939.9	3254.4	290.0	346.2	991.0	$1.02\times$
MaLoRA	295.4	769.6	924.0	3277.4	261.7	307.3	972.6	$1.00\times$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claim made in the abstract can be found in Theorem 3.1, Theorem 3.2, Theorem 3.3, and Algorithm 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we discuss the limitations of the work in §5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: we list all the assumptions in §2.1, and provide proofs in Appendix D–Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we list all the model architectures, datasets, and training setups in AppendixH.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: all the experiments conducted in this paper use open source datasets, and we provide citations to them.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: in Appendix H, we provide details on the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: in Appendix H, we report the mean normalized accuracy and drop with std computed based on 10 repetitions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: we report the usage of computational resources in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: we have reviewed the NeurIPS Code of Ethics and followed it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: in §4, we cite the original paper that produces the datasets used in the experiment.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: we document how we use LLM in Appendix A.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.