GRADIENT-FLOW FRAMEWORK FOR LORA: FINITE-STEPSIZE GUARANTEES AND RANK-DEPENDENT IMPLICIT BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Previous empirical studies have shown that LoRA achieves model quality comparable to full-parameter methods on downstream fine-tuning tasks, even for rank-1 updates. By contrast, the theoretical underpinnings of LoRA remain relatively unexplored. In this work, we analyze the behavior of LoRA from a gradient flow perspective. We first rigorously derive the gradient-flow equations governing LoRA-adapted gradient descent in the limit that finetuning stepsize $\alpha \to 0$. In particular, we show that their form is agnostic to whether LoRA parameter updates are applied sequentially or simultaneously. We then show that for finite stepsize $\alpha > 0$, the discrete LoRA iterates track the gradient-flow trajectory to within $O(k\alpha)$ uniformly on finite finetuning horizons. Finally, we show how the gradient flow equations can be used to analyze the dependence of LoRA's implicit bias on rank parameter r. We do so by demonstrating the relative approximation error between the low-rank and full-rank minimizers on two analytically tractable objectives, obtaining closed-form, rank-explicit characterizations of the approximation error in each setting. The objective functions themselves serve as illustrative testbeds with well-behaved, closed-form solutions; the contribution is the gradient-flow framework and methodology, which extends in principle to other fine-tuning objectives.

1 INTRODUCTION

Parameter-efficient finetuning (PEFT) has become the norm for training deep learning models on downstream tasks. Low-Rank Adaptation (LoRA) (Hu et al., 2022) in particular has gained wide popularity among such methods. Consider a model with pre-trained weights $W^* \in \mathbb{R}^{n\times m}$, and let $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ denote a downstream finetuning objective. Keeping W^* frozen, finetuning consists of finding an update $\Delta W \in \mathbb{R}^{n\times m}$ which approximately solves

$$\min_{\Delta W} f\left(W^* - \Delta W\right) \tag{1}$$

for the downstream finetuning task. While full-parameter finetuning (FPFT) methods search all of $\mathbb{R}^{n\times m}$ for optimal ΔW, LoRA splits the classical gradient descent (GD) algorithm into separate updates which approximately solve $\min_{A,B} f(W^* - BA)$, where $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$ for $r << \min(n, m)$. By lowering the number of finetuning parameters from nm to $r(n + m)$, LoRA enables parameter efficiency crucial for models with hundreds of billions of trainable parameters.

While several empirical studies have verified the effectiveness of LoRA for fine-tuning models in both natural language processing (Li et al., 2024; Lazauskas, 2025; Hanindhito et al., 2025; Mao et al., 2025) and computer vision tasks (Zanella & Ayed, 2024; Agiza et al., 2024), the breadth of theoretical literature on LoRA remains sparse. Existing work in this area focuses on convergence (Malinovsky et al., 2024; Mu & Klabjan, 2025), expressiveness bounds (Zeng & Lee, 2024), initialization (Hayou et al., 2024a), and the learning dynamics of LoRA from a gradient-flow perspective (Xu et al., 2025). While prior studies (Xu et al., 2025; Zeng & Lee, 2024) have derived theoretical *bounds* on the loss and approximation error attainable by LoRA for certain *rank thresholds*, none to our knowledge have used gradient-flow analysis to derive *explicit, closed-form* relationships between LoRA's approximation error and chosen rank parameter r. By *approximation error*, we mean the

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

Frobenius-norm discrepancy between the converged full-rank GD minimizer and the converged rank-r LoRA minimizer. We establish such explicit rank-approximation tradeoffs via a gradient-flow methodology, illustrated on two analytically tractable loss functions.

We refer the reader to Appendix A for an explanation of the notation used in this work.

1.1 CONTRIBUTIONS

Our core contribution is methodological: we develop a gradient-flow framework for deriving closed-form, rank-explicit characterizations of LoRA's approximation error at convergence. Our work consists of two parts: a rigorous derivation of the GF equations governing LoRA-adapted GD, and a loss-specific analytical strategy for solving the resulting ODE systems to extract rank-approximation tradeoffs. We illustrate the framework on two analytically tractable loss functions; in each case, our novelty lies primarily in the characterization of LoRA's convergence behavior rather than the significance of the loss functions themselves. The loss functions serve as demonstrations of our methodology, which extends in principle to any sufficiently smooth fine-tuning objective.

Concretely:

1. **Gradient-flow derivation.** While the ODEs governing LoRA are stated in previous work (Xu et al., 2025), we provide the first (to our knowledge) rigorous derivation of these GF dynamics as the continuous-time (stepsize $\alpha \to 0$) limit of deterministic gradient descent with fixed step size under a low-rank parameterization. For updates of the form BA, we show that the resulting GF equations are invariant across a broad range of LoRA parameter update schemes. This invariance result is foundational to the framework: it ensures the rank-dependent characterizations we derive do not depend on implementation choices in the discrete optimization scheme. Finally, we bound the deviation between the gradient-flow trajectory and the iterates of LoRA-adapted gradient descent at finite step size $\alpha > 0$, showing the two agree to $O(k\alpha)$ uniformly on any bounded time interval, where k is the number of inner updates per iteration (see Algorithm 1).
2. **Application to the trace-squared objective.** We apply the framework to
$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \tfrac{1}{2}\operatorname{Tr}^2(W^* - BA), \tag{2}$$
where $r < n$. We show that the LoRA GF dynamics admit a closed-form solution to (2) for near-arbitrary initial conditions, and we use this solution to characterize LoRA's implicit bias, showing that its selected minimizer approaches the full-rank GD minimizer to this problem at rate $O(r^{-1/2})$.
3. **Application to low-rank approximation.** We apply the framework to
$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times m}}} \tfrac{1}{2}\left\|W^* - BA\right\|^2, \tag{3}$$
where $r < \min(n, m)$. Low-rank approximation is ubiquitous in applied mathematics and ML, with applications in areas such as recommender systems (Chen et al., 2020) and image processing (Liu et al., 2019). Under a spectral initialization scheme related to but distinct from prior work (Meng et al., 2024; Bałazy et al., 2024; Xu et al., 2025; Wang et al., 2025; Mu et al., 2025), we show that the LoRA GF dynamics admit a closed-form solution. Moreover, the nonzero singular values of the rank-r fine-tuning matrix BA converge to the top r singular values of W^*. Within our initialization scheme, this recovers the optimal rank-r solution characterized by the Eckart-Young-Mirsky theorem (Eckart & Young, 1936) and provides a continuous-time, dynamical-systems perspective on a classical static result. The resulting approximation error between the low-rank and full-rank minimizers is then characterized by W^*'s bottom $\min(n, m) - r$ singular values.

We emphasize that our contribution is methodological. The objectives in (2–3) serve as analytically tractable demonstrations of the GF framework, consistent with a tradition of using idealized objectives to expose structural behavior in optimization theory (Saxe et al., 2014; Du et al., 2018a; Arora et al., 2018).

1.2 Related Work

1.2.1 Expressive Power of LoRA

Zeng & Lee (2024) show that a rank-R adapter can exactly represent a smaller target network once R exceeds an explicit threshold. They give a rank-independent error bound for R below the threshold without analyzing a specific finetuning algorithm. In contrast, we derive rank-dependent approximation errors for LoRA-adapted GD directly.

1.2.2 LoRA Gradient Flow

To our knowledge, only Xu et al. (2025) have previously analyzed LoRA from a GF perspective. While their work studies the convergence of LoRA within a matrix factorization (MF) setting, they focus on how convergence *rates* and *behavior* depend on *initialization schema*. In particular, Xu et al. consider fine-tuning a perfectly factorized matrix $Y_{\text{pre}} = W_2 W_1$ toward an updated target matrix Y_{ft} via low-rank updates $B_2 A_2$ and $B_1 A_1$. They find that the low-rank updates drive the fine-tuning loss to zero, recovering Y_{ft} exactly.

By contrast, our work focuses on the distinct classical low-rank approximation problem. In particular, we analyze how well LoRA can approximate W^* *as a function of the rank of* BA. In this sense, our work connects more closely to foundational MF work by Eckart, Young, and Mirsky (Eckart & Young, 1936; Mirsky, 1960), providing a continuous-time extension of their theoretical findings from the perspective of LoRA optimization dynamics.

Finally, while Xu et al. state the gradient-flow equations governing LoRA, we provide their rigorous derivation and show that these equations are invariant to broad choice of parameter update scheme, resolving an ambiguity between theoretical formulations and common implementation practice.

1.2.3 Other Related Theoretical Work

Further theoretical work on LoRA focuses on convergence rates (Malinovsky et al., 2024; Mu & Klabjan, 2025), initialization (Hayou et al., 2024a), and learning-rate selection (Hayou et al., 2024b). Related work on MF (Ward & Kolda, 2023; Du et al., 2018a) studies convergence behavior of factorized optimization problems but does not consider low-rank approximation. Other studies propose spectral initialization methods for LoRA (Bałazy et al., 2024; Meng et al., 2024; Wang et al., 2025; Mu et al., 2025).

2 Results

2.1 LoRA Gradient Flow Equations

Denote the Cartesian product space $\Theta := \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m}$. Consider a continuously differentiable function $g : \Theta \to \mathbb{R}$ which we seek to minimize over Θ:

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times m}}} g\left(B, A\right) \tag{4}$$

We solve (4) via LoRA-adapted GD given in Algorithm 1 of Appendix B. Algorithm 1 updates A and B and is agnostic to how these updates are performed. For $\lambda = 1$, the updates of A and B are performed simultaneously, with both parameters evaluated using data from the previous iteration. In contrast, $\lambda = 0$ corresponds to sequential updates, where B is first updated using its prior iterate, after which A is updated using the newly computed value of B from the current iteration. For $0 < \lambda < 1$, the update of A uses a convex combination of the values of B from the current and previous iterations. Algorithm 1 further generalizes the update scheme by allowing A and B to be updated a total of k times within each iteration. While $\lambda = k = 1$ is used in practice (Hu & contributors, 2021), we allow arbitrary $\lambda \in [0, 1]$ and $k \in \mathbb{N}$ as a proof device to show that the resulting GF equations for LoRA are invariant to the choice of update scheme.

The update equations at any given iteration i of Algorithm 1 are given by

$$B^{(\alpha)}_{ik+j+1} = B^{(\alpha)}_{ik+j} - \alpha \nabla_B g\left(B^{(\alpha)}_{ik+j}, A^{(\alpha)}_{ik}\right) \tag{5}$$

$$A^{(\alpha)}_{ik+j+1} = A^{(\alpha)}_{ik+j} - \alpha \nabla_A g\left(\left[\lambda B^{(\alpha)}_{ik} + (1-\lambda) B^{(\alpha)}_{(i+1)k}\right], A^{(\alpha)}_{ik+j}\right). \tag{6}$$

Note α is the fixed learning rate.

The discrete sets of updates for A and B motivate us to define their continuous affine interpolations:

$$Y_\alpha(t) = B^{(\alpha)}_{\lfloor \frac{t}{\alpha} \rfloor} - \left(t - \left\lfloor \frac{t}{\alpha} \right\rfloor \alpha\right) \nabla_B g\left(B^{(\alpha)}_{\lfloor \frac{t}{\alpha} \rfloor}, A^{(\alpha)}_{h(t)}\right) \tag{7}$$

$$X_\alpha(t) = A^{(\alpha)}_{\lfloor \frac{t}{\alpha} \rfloor} - \left(t - \left\lfloor \frac{t}{\alpha} \right\rfloor \alpha\right) \nabla_A g\left(\left[\lambda B^{(\alpha)}_{h(t)} + (1-\lambda) B^{(\alpha)}_{h(t)+k}\right], A^{(\alpha)}_{\lfloor \frac{t}{\alpha} \rfloor}\right), \tag{8}$$

where

$$h(t) = \left\lfloor \frac{t}{\alpha} \right\rfloor - \left(\left\lfloor \frac{t}{\alpha} \right\rfloor \mod k\right). \tag{9}$$

We proceed by defining the candidate GF dynamics *a priori*. Specifically, let $X : \mathbb{R}^+_0 \to \mathbb{R}^{r \times m}$ and $Y : \mathbb{R}^+_0 \to \mathbb{R}^{n \times r}$ be matrix-valued functions satisfying the initial-value problem (IVP)

$$\frac{dY(t)}{dt} = -\nabla_Y g\Big(Y(t), X(t)\Big) \tag{10}$$

$$\frac{dX(t)}{dt} = -\nabla_X g\Big(Y(t), X(t)\Big) \tag{11}$$

$$Y(0) = B_0 \tag{12}$$

$$X(0) = A_0. \tag{13}$$

We justify the existence of solutions $Y(t)$ and $X(t)$ under the following assumptions:

Assumption 2.1 (Boundedness of Iterates)**.** *Let $T > 0$ be any finite number. We assume that the iterates $\left\|B^{(\alpha)}_{ik+j}\right\|$ and $\left\|A^{(\alpha)}_{ik+j}\right\|$ do not escape to ∞ during finite finetuning horizons. That is, there exists $R' > 0$ such that*

$$\max\left(\left\|B^{(\alpha)}_{ik+j}\right\|, \left\|A^{(\alpha)}_{ik+j}\right\|\right) \leq R' \tag{14}$$

for all $ik + j \in \mathbb{N}_0$ such that $(ik + j)\alpha \leq T$.

Note that any nonnegative objective satisfies Assumption 2.1 for sufficiently small stepsize on finite finetuning horizons, as we state in the remark below:

Remark 2.2 (Sufficient Condition for Assumption 2.1)**.** *Suppose the objective function g is Lipschitz smooth and bounded below. That is, g satisfies Assumption 2.4, and there exists some $M \in \mathbb{R}$ such that*

$$g(B, A) \geq M \quad \text{for all } (B, A) \in \Theta. \tag{15}$$

Then for any finite $T > 0$, there exist $R'_T > 0$ and $\alpha_0 > 0$, both depending only on T, k, $\theta(0)$, and g, such that for every $\alpha < \alpha_0$ the iterates of Algorithm 1 satisfy

$$\max\left(\left\|B^{(\alpha)}_{ik+j}\right\|, \left\|A^{(\alpha)}_{ik+j}\right\|\right) \leq R'_T \tag{16}$$

for all $ik + j \in \mathbb{N}_0$ such that $(ik + j)\alpha \leq T$. We provide proof for this remark in Appendix L.

Remark 2.3 (Uniform Boundedness of Gradient)**.** *The gradient $\|\nabla g(B, A)\|$ is uniformly bounded on bounded subsets of Θ. That is, for every bounded set $S \subset \Theta$, there exists a constant $M_S < \infty$ such that*

$$\|\nabla g(B, A)\| \leq M_S \quad \text{for all } (B, A) \in S.$$

This remark follows from Assumptions 2.1 and 2.4. Definitions of norms and gradients on Θ are provided in Remark B.1 of Appendix B.

In subsequent portions of our derivation, we assume that g is Lipschitz smooth:

Assumption 2.4 (Lipschitz Smoothness of g)**.** *The function g is Lipschitz smooth on bounded subsets of Θ. That is, for any $(B_1, A_1), (B_2, A_2) \in \Theta$ satisfying*

$$\max\Big(\|B_1\|, \|A_1\|, \|B_2\|, \|A_2\| \Big) \leq R, \tag{17}$$

there exists a constant $L_R > 0$ such that

$$\|\nabla g(B_2, A_2) - \nabla g(B_1, A_1)\| \leq L_R \|(B_2, A_2) - (B_1, A_1)\|. \tag{18}$$

Under these assumptions, we show that

$$\lim_{\alpha \to 0} ||Y_\alpha(t) - Y(t)|| = 0 \tag{19}$$

$$\lim_{\alpha \to 0} ||X_\alpha(t) - X(t)|| = 0 \tag{20}$$

for all $t \geq 0$. The uniform convergence in (19–20) implies that the continuous affine interpolations defined in (7–8), and thus the underlying discrete iterates themselves, evolve according to the dynamical system in (10–11) in the limit where $\alpha \to 0$. This brings us to our final result:

Theorem 2.5. *Consider an objective function $g : \Theta \to \mathbb{R}$ satisfying Assumptions 2.1 and 2.4 which is minimized via Algorithm 1. During finetuning, the iterates produced by Algorithm 1 evolve according to the dynamical system in (10–11) for arbitrary $\lambda \in [0, 1]$ and $k \in \mathbb{N}$ in the limit that $\alpha \to 0$. For finite step size $\alpha > 0$, the error between the iterates produced by Algorithm 1 and the solution to (10–11) is at most $O(k\alpha)$; this is an upper bound, and the error may be smaller.*

The full proof of Theorem 2.5 is provided in Appendix B. Numerical studies of the error bound's tightness in α and k appear in Appendices M and O, respectively. We now present two problems to demonstrate use of this result to analyze how the approximation error of LoRA depends on update rank.

2.2 Trace-Squared Objective

Consider the finetuning problem

$$\min_{\Delta W \in \mathbb{R}^{n \times n}} \frac{1}{2} \operatorname{Tr}^2(W^* - \Delta W), \tag{21}$$

as well as its rank-r constrained equivalent:

$$\min_{\substack{B \in \mathbb{R}^{n \times r} \\ A \in \mathbb{R}^{r \times n}}} \frac{1}{2} \operatorname{Tr}^2(W^* - BA), \tag{22}$$

where $r < n$. We analyze the implicit bias of LoRA-adapted GD on this loss function by quantifying the rank-dependent approximation error between the respective minimizers of (21) and (22) obtained by full-rank GD and LoRA.

From the GF equation for full-rank GD, we know that the learning dynamics of (21) during finetuning are governed by

$$\frac{dU(t)}{dt} = -\nabla \left(\frac{1}{2} \operatorname{Tr}^2 \Big(W^* - U(t) \Big) \right) = \operatorname{Tr} \big(W^* - U(t) \big) I_n. \tag{23}$$

Here, $U : \mathbb{R}_0^+ \to \mathbb{R}^{n \times n}$ denotes the full-rank iterate produced by full-rank GF applied to (21) at time $t \geq 0$. We show that the unique closed-form solution to (23) is given by

$$U(t) = \frac{1 - e^{-nt}}{n} \operatorname{Tr} (W^* - Y_0 X_0) I_n + Y_0 X_0, \tag{24}$$

where we initialize $U(0) = Y_0 X_0$ for consistency with the LoRA setting. Following LoRA initialization (Hu et al., 2022), we set $Y_0 = \mathbf{0}_{n \times r}$ and draw the entries of X_0 i.i.d. as

$$x_{ij}(0) \sim \mathcal{N}(0, \sigma^2), \quad \sigma^2 > 0, \tag{25}$$

so that the full-rank iterates converge to

$$\lim_{t\to\infty} U(t) = \frac{1}{n}\operatorname{Tr}(W^*)\, I_n, \tag{26}$$

which yields zero final loss in (21). Full details for our calculation of (26) are provided in Appendix G.

For the rank-r problem in (22), we assume uniform boundedness of the iterates as stated in Assumption 2.1 and justified in Remark 2.2. Under this assumption, we show that the gradient of the trace-squared objective in (22) is uniformly bounded and Lipschitz continuous on bounded subsets of Θ. Furthermore, we assume

Assumption 2.6 (Nonzero Initialization for $\|X_0\|$)**.** *We assume $\|X_0\| \neq 0$, which holds almost surely for the Gaussian initialization in* (25)*.*

Assumption 2.7 (Nonzero Trace for Target Matrix)**.** *We assume* $\operatorname{Tr}(W^*) \neq 0$.

If either Assumption 2.6 or 2.7 is violated, the initialization is a stationary point of (27–28) and $Y(t)X(t) = \mathbf{0}_{n\times n}$ for all $t \geq 0$.

With all assumptions from Section 2.1 satisfied, we know that the learning dynamics of (22) during finetuning are governed by

$$\frac{dY(t)}{dt} = -\nabla_Y \left(\frac{1}{2}\operatorname{Tr}^2\left(W^* - Y(t)X(t)\right)\right) = \operatorname{Tr}(W^* - YX)X^T \tag{27}$$

$$\frac{dX(t)}{dt} = -\nabla_X \left(\frac{1}{2}\operatorname{Tr}^2\left(W^* - Y(t)X(t)\right)\right) = \operatorname{Tr}(W^* - YX)Y^T \tag{28}$$

Here, $Y : \mathbb{R}_0^+ \to \mathbb{R}^{n\times r}$ and $X : \mathbb{R}_0^+ \to \mathbb{R}^{r\times n}$ denote the low-rank iterates produced by LoRA-adapted GF applied to (22) at time $t \geq 0$. We find the following closed-form solution to (27–28):

$$Y(t) = q(t)\operatorname{Tr}(W^*)\, X_0^T \tag{29}$$

$$X(t) = p(t)X_0 \tag{30}$$

where $p(t)$ and $q(t)$ are smooth scalar functions on $[0,\infty)$ given in Appendix D. Note that, while we draw the entries of X_0 from the distribution in (25), the solution in (29–30) holds for arbitrary $X_0 \in \mathbb{R}^{r\times n}$ with nonzero norm.

Right multiply (29) by (30) to find the rank-r iterate produced by LoRA-adapted GF applied to (22) at time $t \geq 0$:

$$Y(t)X(t) = p(t)q(t)\operatorname{Tr}(W^*)X_0^T X_0. \tag{31}$$

We find that the rank-r iterates converge to

$$\lim_{t\to\infty} Y(t)X(t) = \lim_{t\to\infty} p(t)q(t)\operatorname{Tr}\left(W^*\right) X_0^T X_0 = \frac{\operatorname{Tr}(W^*)}{\|X_0\|^2} X_0^T X_0, \tag{32}$$

which yields zero final loss in (22). Since full-rank GD also attains zero loss for this problem, the two methods differ only in which minimizer they select. Full details for our calculation of (32) are provided in Appendix D.

In Appendix H, we show that the relative approximation error between $U(t)$ and $Y(t)X(t)$ converges to

$$\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|}{\|U(t)\|} = \frac{\sqrt{n\left\|X_0^T X_0\right\|^2 - \|X_0\|^4}}{\|X_0\|^2}, \tag{33}$$

for arbitrary $X_0 \in \mathbb{R}^{r\times n}$ with nonzero norm. If we draw the independent entries of X_0 from (25), we can calculate the expectation of the *square* of our relative approximation error with respect to the initial conditions as a function of n and r:

$$\mathbb{E}\left[\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|^2}{\|U(t)\|^2}\right] = \frac{n^2 + n - 2}{nr + 2}. \tag{34}$$

By Jensen's Inequality (Dekking et al., 2005), we obtain an upper bound on the expectation of the relative approximation error itself:

$$\mathbb{E}\left[\lim_{t\to\infty}\frac{\|Y(t)X(t)-U(t)\|}{\|U(t)\|}\right]\leq\sqrt{\frac{n^2+n-2}{nr+2}}, \tag{35}$$

which shows that, for fixed n, the expected relative approximation error decays at least as fast as $O(r^{-1/2})$ as r increases. We thus have our next result:

Theorem 2.8. *Consider the trace-squared optimization problems in* (21) *and* (22)*. For* (22)*, assume uniform boundedness of the iterates as in Assumption 2.1 as well as the nonzero initialization in Assumptions 2.6–2.7. Under GF, the relative approximation error between the converged full-rank and low-rank minimizers is given exactly by* (33)*.*

When the entries of X_0 are drawn independently according to (25)*, the expectation of the squared approximation error at convergence with respect to the initial conditions is given by* (34)*. An upper bound on the expectation of the converged approximation error is given by* (35)*.*

A complete derivation of these results is provided in Appendix H. A numerical study on the tightness of the bound in (35) is given in Appendix N. We have thus shown that, using the LoRA GF equations derived in Section 2.1, one can characterize the implicit bias of LoRA by bounding the expected relative approximation error at convergence for the rank-r constrained finetuning problem in (22) as a function of r.

2.3 Low-Rank Approximation

Consider the low-rank approximation problem

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\A\in\mathbb{R}^{r\times m}}}\frac{1}{2}\left\|W^*-BA\right\|^2, \tag{36}$$

where $r<\min(n,m)$. Using the GF equations derived in Section 2.1, we show that LoRA-adapted GF converges to the optimal rank-r solution to (36) given by the EYM Theorem (Eckart & Young, 1936) under our spectral initialization scheme. Our contribution here is not to low-rank approximation itself–for which the optimal solution is well-known–but rather to the characterization of LoRA's convergence behavior in this setting, and in particular how that behavior depends on the chosen rank r.

We again assume uniform boundedness for the iterates as stated in Assumption 2.1 and justified in Remark 2.2. Under this assumption, we show that the gradient of the objective in (36) is uniformly bounded and Lipschitz continuous on bounded subsets of Θ. Furthermore, we assume

Assumption 2.9 (Rank of W^*)**.** *We assume that*

$$r<\operatorname{rank}(W^*)\leq\min(n,m). \tag{37}$$

Violation of Assumption 2.9 renders (36) an over-parameterized matrix factorization problem rather than a low-rank approximation problem.

With all assumptions from Section 2.1 satisfied, the learning dynamics for (36) under LoRA are governed by

$$\frac{dY(t)}{dt}=-\nabla_Y\left(\frac{1}{2}\left\|W^*-Y(t)X(t)\right\|^2\right)=(W^*-YX)X^T \tag{38}$$

$$\frac{dX(t)}{dt}=-\nabla_X\left(\frac{1}{2}\left\|W^*-Y(t)X(t)\right\|^2\right)=Y^T(W^*-YX) \tag{39}$$

Under the standard initialization scheme for LoRA described in Section 2.2, we are unaware of any closed-form solution to the ODE system above. However, under a spectral initialization scheme related to that described for LoRA in previous work (Meng et al., 2024; Bałazy et al., 2024; Xu et al., 2025; Wang et al., 2025; Mu et al., 2025), the dynamics in (38–39) decouple into scalar ODEs governing the evolution of the singular values of $Y(t)X(t)$ during training. We then solve the

resulting scalar equations to demonstrate the convergence of the nonzero singular values of $Y(t)X(t)$ to the top r singular values of W^*.

We begin by writing the singular value decomposition (SVD) of W^*:

$$W^* = U\Sigma^* V^T. \tag{40}$$

where the singular values in Σ^* are in non-increasing order along the diagonal. We pause to note that the SVD of W^* and the spectral initialization serve only to analyze the dynamics of LoRA for this problem; neither is computed or used by Algorithm 1 in practice. When W^* has distinct singular values, every local minimum of (36) has BA equal to the EYM truncation (Baldi & Hornik, 1989, Fact 4), so any minimizer reached by LoRA GD yields the same product BA regardless of initialization. This does not guarantee that LoRA GD, as a first-order method, converges to a minimizer. Rather, writing out the SVD allows a closed-form characterization of the LoRA GF trajectory under spectral initialization.

Using the matrices of left and right singular vectors for W^*, we define the following transformations of X and Y:

$$\tilde{Y} = U^T Y \tag{41}$$

$$\tilde{X} = XV. \tag{42}$$

Under this change of variables, (38–39) become

$$\frac{d\tilde{Y}(t)}{dt} = \left(\Sigma^* - \tilde{Y}\tilde{X}\right)\tilde{X}^T \tag{43}$$

$$\frac{d\tilde{X}(t)}{dt} = \tilde{Y}^T\left(\Sigma^* - \tilde{Y}\tilde{X}\right). \tag{44}$$

We initialize $\tilde{Y}_0 = \mathbf{0}_{n\times r}$. The initial $\tilde{X}_0$ has zero off-diagonal entries, while its diagonal entries are drawn independently from a centered Gaussian distribution with variance $\sigma^2 > 0$, i.e.,

$$\tilde{x}_{ij}(0) = \begin{cases} 0 & i \neq j, \\ \mathcal{N}(0, \sigma^2) & i = j. \end{cases} \tag{45}$$

To prevent stationary point initialization in (46–47), we assume the diagonal entries of $\tilde{X}$ are initially nonzero:

Assumption 2.10 (Nonzero Diagonal Initialization)**.** *We assume $\tilde{x}_{ii}(0) \neq 0$ for all $i \in \{1, \ldots, r\}$, which holds almost surely for the Gaussian initialization in* (45)*.*

This initialization scheme preserves diagonality during training, which decouples (38–39) into r independent scalar dynamical systems:

$$\frac{dy}{dt} = (s^* - yx)\, x \tag{46}$$

$$\frac{dx}{dt} = (s^* - yx)\, y \tag{47}$$

Here, s^* denotes an arbitrary singular value of W^*, while y and x represent the corresponding diagonal entries of $\tilde{Y}$ and $\tilde{X}$.

For $s^* = 0$ (when $\operatorname{rank}(W^*) < \min(n, m)$), we initialize at a stationary point of (46–47), and $y(t)x(t) = 0$ for all $t \geq 0$. As a result, the zero singular values of W^* and $\tilde{Y}(t)\tilde{X}(t)$ are aligned throughout training.

When $s^* \neq 0$, the ODE system admits the closed-form solution

$$y(t) = q_s(t)s^* x_0 \tag{48}$$

$$x(t) = p_s(t)x_0 \tag{49}$$

where $x_0 := x_{ii}(0)$, and $p_s(t)$ and $q_s(t)$ are smooth scalar functions on $[0, \infty)$ given in Appendix J. Taking the product of (48) and (49) yields the corresponding diagonal entry of $\tilde{Y}(t)\tilde{X}(t)$ associated with s^*:

$$y(t)x(t) = p_s(t)q_s(t)s^* x_0^2. \tag{50}$$

Then (50) converges to

$$\lim_{t\to\infty} y(t)x(t) = s^*. \tag{51}$$

We thus find that the nonzero diagonal entries of $\tilde{Y}(t)\tilde{X}(t)$ align with the top r singular values of W^* as $t \to \infty$. The remaining diagonal entries are zero since $\operatorname{rank}\left(\tilde{Y}(t)\tilde{X}(t)\right) = r$. Hence, after convergence, any discrepancy between Σ^* and $\tilde{Y}(t)\tilde{X}(t)$ lies in the extra $\operatorname{rank}(W^*) - r$ nonzero singular values of W^*. Note that

$$YX = U\tilde{Y}\tilde{X}V^T. \tag{52}$$

Transforming back to Y and X, any discrepancy between $Y(t)X(t)$ and W^* after convergence can be attributed to the bottom $\min(n, m) - r$ singular values of W^*, i.e.

$$\lim_{t\to\infty} \|W^* - Y(t)X(t)\|^2 = \sum_{i=r+1}^{\min(n,m)} {s_i^*}^2 \tag{53}$$

where s_i^* denotes the ith largest singular value of W^*. The result in (53) matches that given by EYM for the loss between W^* and its optimal rank-r minimizer (Eckart & Young, 1936). Since the bottom $\min(n, m) - r$ singular values of $Y(t)X(t)$ are zero after convergence, we can write the final rank-r minimizer for (36) obtained under LoRA as

$$\lim_{t\to\infty} Y(t)X(t) = U_r\Sigma_r^*V_r^T \tag{54}$$

where U_r and V_r (truncations of U and V) are the matrices of left and right singular vectors for the top r singular values of W^*, and $\Sigma_r^* \in \mathbb{R}^{n\times m}$ is a diagonal matrix containing the top r singular values of W^*, with all remaining diagonal entries equal to zero. The minimization problem equivalent to (36) with unconstrained rank is trivially solved by W^* itself. Consequently, the converged approximation error between the full-rank and low-rank solutions coincides with the final loss in this setting. Our GF analysis therefore shows that, for low-rank approximation, the approximation error of LoRA scales with r at a rate given by the tail of the singular-value spectrum of W^*, with each increment of r reducing the loss by the next squared singular value. We thus arrive at our final result:

Theorem 2.11. *Consider the optimization problem in* (36)*. Assume uniform boundedness of the iterates as in Assumption 2.1 (justified in Remark 2.2) and nonzero initialization as in Assumption 2.10. Assume* $r < \operatorname{rank}(W^*)$ *as in Assumption 2.9.*

Under LoRA GF, the converged rank-r *minimizer of* (36) *is given by* (54)*. Furthermore, the final loss and approximation error between the low-rank and full-rank solutions are given by* (53)*. These results coincide with the optimal rank-*r *minimizer and loss characterized by the Eckart–Young–Mirsky theorem.*

A detailed calculation for this section is provided in Appendix J.

3 Conclusion

This paper presents a GF method for determining the variability of LoRA's approximation error with chosen update rank r. The form of the GF equations is rigorously proven in a manner agnostic to update scheme. From the resulting dynamical system, our GF method is illustrated through rank-approximation analysis of two tractable loss functions. We are clear that our contribution lies in the methodology itself rather than a groundbreaking result for the loss functions themselves–consistent with previous work in optimization theory (Saxe et al., 2014; Du et al., 2018b; Arora et al., 2018). Future work will use our method as a stepping-stone to study more complex and task-specific loss functions.

Limitations in our work include the spectral initialization required for our closed-form trajectory in the low-rank approximation setting, which cannot be used in practice since it depends on the singular vectors of W^*; while the final minimizer BA does not depend on this choice, our trajectory characterization does.

Reproducibility Statement

This work contains two main types of results: theoretical and experimental (Appendices M,N,O). To aid in reproducibility for the short experimental results, a .zip folder has been included with the supplementary materials of this submission along with a README documenting how to run the code used to reproduce the experiments as well as the environment/dependency requirements for running the code. To aid in reproducibility for the theoretical results, detailed mathematical calculations have been included in the paper's appendix showing how we arrived at the theoretical results.

AI use statement

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, implement numerical experiments, and interpret results. We have not used generative AI tools for generating synthetic datasets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, or support qualitative and thematic data analysis. Additionally, we used generative AI tools for writing Python code for numerical experiments, editing the paper to improve readability, brainstorming, identifying relevant literature, and propose a title for the research paper. We have reviewed all AI-assisted work. We have independently verified all mathematical proofs by hand. We have verified all generated secondary sources were real and really contained the claims that AI said they did. We have drafted all parts of the paper ourselves and only used AI for suggestions to make the readability clearer. We have examined in detail graphs produced by the numerical experiments whose code was written using AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Ahmed Agiza, Marina Neseem, and Sherief Reda. Mtlora: A low-rank adaptation approach for efficient multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16196–16205, 2024. doi: 10.1109/CVPR52733.2024.01533. URL https://doi.org/10.1109/CVPR52733.2024.01533.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin, Heidelberg, 3 edition, 2006. ISBN 978-3-540-29587-7. doi: 10.1007/3-540-29587-9. See p. 552 for the definition of sequence-space p-sums and product norms.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/arora18a.html.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. Lora-xs: Low-rank adaptation with extremely small number of parameters. *arXiv preprint arXiv:2405.17604*, 2024. URL https://arxiv.org/abs/2405.17604.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58, 1989. URL https://api.semanticscholar.org/CorpusID:14333248.

Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Low-rank quaternion approximation for color image processing. *IEEE Transactions on Image Processing*, 29:1426–1439, 2020. doi: 10.1109/TIP.2019.2941319.

Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaa, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer Texts in Statistics. Springer, London, 2005. ISBN 978-1-85233-896-1. See Exercise 8.11s for Jensen's Inequality on concave functions.

Simon Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/fe131d7f5a6b38b23cc967316c13dae2-Paper.pdf.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018b.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 1999. ISBN 9780471317166. URL https://books.google.com/books?id=N8jVDwAAQBAJ. See p. 153 for the discussion of product spaces and norms.

Bagus Hanindhito, Bhavesh Patel, and Lizy K. John. Large language model fine-tuning with low-rank adaptation: A performance exploration. In *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering (ICPE '25)*, pp. 92–104. ACM, 2025. doi: 10.1145/3676151.3719377. URL https://doi.org/10.1145/3676151.3719377.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on lora finetuning dynamics. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9798331314385.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Edward J. Hu and contributors. Lora: Official implementation. https://github.com/microsoft/LoRA, 2021. Accessed: 2025-11-17.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tomas Lazauskas. Empirical evaluation of low-rank adaptation for efficient fine-tuning of large language models. Technical Report 9, The Alan Turing Institute, 2025. URL https://doi.org/10.5281/zenodo.16417805.

Junjie Li, Aseem Sangalay, Cheng Cheng, Yuan Tian, and Jinqiu Yang. Fine tuning large language model for secure code generation. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering (Forge)*, pp. 86–90, Lisbon, Portugal, 2024. IEEE/ACM. doi: 10.1145/3650105.3652299. URL https://doi.org/10.1145/3650105.3652299.

Huafeng Liu, Liping Jing, Yuhua Qian, and Jian Yu. Adaptive local low-rank matrix approximation for recommendation. *ACM Trans. Inf. Syst.*, 37(4), October 2019. ISSN 1046-8188. doi: 10.1145/3360488. URL https://doi.org/10.1145/3360488.

Grigory Malinovsky et al. Randomized asymmetric chain of lora: The first meaningful theoretical framework for low-rank adaptation. *arXiv preprint arXiv:2410.08305v1*, 2024. URL https://arxiv.org/abs/2410.08305v1.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19:197605, 2025. doi: 10.1007/s11704-024-40663-9. URL https://doi.org/10.1007/s11704-024-40663-9.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: principal singular values and singular vectors adaptation of large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50. URL https://doi.org/10.1093/qmath/11.1.50.

Lin Mu, Xiaoyu Wang, Li Ni, Yang Li, Zhize Wu, Peiquan Jin, and Yiwen Zhang. Denselora: Dense low-rank adaptation of large language models. *arXiv preprint*, arXiv:2505.23808, 2025. doi: 10.48550/arXiv.2505.23808. URL https://arxiv.org/abs/2505.23808.

Siqiao Mu and Diego Klabjan. On the convergence rate of lora gradient descent. *arXiv preprint arXiv:2512.18248*, 2025. URL https://arxiv.org/abs/2512.18248.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. doi: 10.1017/9781108231596. See Exercise 3.3.7.

Wolfgang Walter. *Ordinary Differential Equations*, volume 182 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1998. ISBN 978-0-387-98459-9. doi: 10.1007/978-1-4612-0601-9. See p. 62-83.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4823–4836, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.248. URL https://aclanthology.org/2025.naacl-long.248/.

Rachel A. Ward and Tamara G. Kolda. Convergence of alternating gradient descent for matrix factorization. In *Advances in Neural Information Processing Systems*, NeurIPS 2023, pp. —, 2023. doi: 10.5555/3598907.3599026. URL https://proceedings.neurips.cc/paper/2023/hash/46c10f6c8ea5aa6f267bcdabcb123f97-Abstract.html.

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and René Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, Mai Khao, Thailand, 2025. PMLR.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1593–1603, 2024. URL https://api.semanticscholar.org/CorpusID:270094592.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=likXVjmh3E.

A NOTATION

Throughout this paper, capital letters denote matrices, and lowercase letters denote scalars. Boldface letters denote vectors. For a matrix A, we write A^T for its transpose and a_{ij} for its (i, j)-th entry. We use $\|A\|$ to denote the Frobenius norm of A and $\|\mathbf{a}\|_2$ to denote the Euclidean norm of a vector $\mathbf{a}$. We use $\lfloor \cdot \rfloor$ to denote the integer floor function. We denote $\mathrm{Tr}^2(W) := \Big(\mathrm{Tr}(W)\Big)^2$.

For a time-dependent function $f(t)$ we write $f'(t) := \frac{d}{dt} f(t)$ for its time derivative. Uppercase $A(t)$ refers to a matrix-valued function of time. When emphasizing time dependence, we write $a_{ij}(t)$ for the (i, j)-th entry of $A(t)$, with $a_{ij}(0)$ denoting its value at initialization.

We refer to the set of all natural numbers using $\mathbb{N}$ and to the set of non-negative real numbers using $\mathbb{R}_0^+$.

Let $B \in \mathbb{R}^{n\times r}$. For a scalar-valued function $g(B, A)$ of *two matrices*, we define the partial gradient $\nabla_B g(B, A) := \frac{\partial g}{\partial B}$ to be the matrix in $\mathbb{R}^{n\times r}$ whose (i, j)-th entry is $\frac{\partial g}{\partial b_{ij}}$.

Similarly, let $W \in \mathbb{R}^{n\times m}$. For a scalar-valued function $f(W)$ of a *single matrix*, we define the gradient $\nabla f(W) := \frac{\partial f}{\partial W}$ to be the matrix in $\mathbb{R}^{n\times m}$ whose (i, j)-th entry is $\frac{\partial f}{\partial w_{ij}}$.

The $n \times n$ identity matrix is denoted by I_n. The $n \times m$ matrix of all zeroes is denoted by $\mathbf{0}_{n\times m}$.

B Derivation of LoRA Gradient Flow ODEs

We wish to analyze the learning dynamics of LoRA by deriving a set of so-called gradient flow ODEs which describe how our parameters $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$ evolve during finetuning. We present below the LoRA-adapted gradient descent algorithm for a continuously differentiable objective $g : \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m} \to \mathbb{R}$:

Algorithm 1 Deterministic Gradient Descent adapted via LoRA (Fixed Stepsize)

1: **Initialize:**
2: Choose an initial $B_0 \in \mathbb{R}^{n\times r}$ and $A_0 \in \mathbb{R}^{r\times m}$
3: Set $B^* = B_0$, $A^* = A_0$, and $i = 0$
4: Choose $\alpha > 0$ and $\lambda \in [0, 1]$

5: **while** not converged **do**

6: **Stage 1: Update B while keeping A^* fixed**
7: **for** $j = 0, \ldots, k-1$ **do**
8: $B_{j+1} = B_j - \alpha \nabla_B g(B_j, A^*)$
9: **end for**
10: Set $B^* \longleftarrow \lambda B^* + (1-\lambda) B_k$
11: Set $B_0 \longleftarrow B_k$

12: **Stage 2: Update A while keeping B^* fixed**
13: **for** $j = 1, \ldots, k$ **do**
14: $A_{j+1} = A_j - \alpha \nabla_A g(B^*, A_j)$
15: **end for**
16: Set $A^* \longleftarrow A_k$
17: Set $A_0 \longleftarrow A_k$
18: Set $B^* \longleftarrow B_k$

19: **Stage 3: Convergence Check**
20: $i \leftarrow i + 1$
21: **if** stopping criteria met **then**
22: **break**
23: **end if**

24: **end while**

25: **Output:** (B^*, A^*), a stationary point of g

Note that the updates of A in stage 2 of the algorithm above are performed using the convex combination given on line 10. This definition for B^* incorporates information about B from both the current and previous iterations and makes our analysis agnostic to whether updates between B and A are performed sequentially ($\lambda = 0$) or simultaneously ($\lambda = 1$) using data from the previous iteration. Although updates are applied simultaneously in practice for parallelism Hu & contributors (2021), we derive the gradient-flow dynamics from the algorithm above to show that our analysis does not depend on whether the updates are performed sequentially or simultaneously.

We wish to show that, in the limit where α approaches zero, the learning dynamics of Algorithm 1 are described by the solution $\Big(Y(t), X(t)\Big)$ of the ODE system

$$\frac{dY(t)}{dt} = -\nabla_Y g\Big(Y(t), X(t)\Big) \tag{55}$$

$$\frac{dX(t)}{dt} = -\nabla_X g\Big(Y(t), X(t)\Big) \tag{56}$$

$$Y(0) = B_0, \tag{57}$$

$$X(0) = A_0, \tag{58}$$

defined for $t \in [0, T)$, where $T > 0$ is arbitrary.

To begin, define the Cartesian product Θ as

$$\Theta := \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m} \tag{59}$$

and denote the elements $\theta \in \Theta$ as ordered pairs

$$\theta = (B, A) \tag{60}$$

for $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$. For the Frobenius norm on $\mathbb{R}^{n\times r}$ and $\mathbb{R}^{r\times m}$, define the function $\|\cdot\| : \Theta \to \mathbb{R}_0^+$ as

$$||\theta|| := \left(||B||^2 + ||A||^2\right)^{1/2} \tag{61}$$

Remark B.1 below shows that $\|\cdot\|$ is a product norm on Θ.

Remark B.1 (Product Norm on Θ)**.** *Define the space*

$$\Theta := \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m},$$

with elements $\theta \in \Theta$ given as ordered pairs of matrices:

$$\theta = (B, A), \tag{62}$$

$$B \in \mathbb{R}^{n\times r} \tag{63}$$

$$A \in \mathbb{R}^{r\times m} \tag{64}$$

equipped with the norm

$$\|\theta\| := \left(\|B\|^2 + \|A\|^2\right)^{1/2}.$$

Then for any $\theta \in \Theta$, we have:

1. $\|\theta\| \geq 0$
2. $\|\theta\| = 0$ *iff* $\theta = (\mathbf{0}_{n\times r}, \mathbf{0}_{r\times m})$
3. *Define scalar multiplication on Θ as $\eta\theta = (\eta B, \eta A)$. Then we have $\|\eta\theta\| = |\eta|\,\|\theta\|$ for all $\eta \in \mathbb{R}$*
4. *For any two elements $\theta_1 = (B_1, A_1)$ and $\theta_2 = (B_2, A_2)$ of Θ, define the binary addition operation $\Theta \times \Theta \to \Theta$ as*

$$\theta_1 + \theta_2 = (B_1 + B_2,\ A_1 + A_2) \tag{65}$$

Then $\|\theta_1 + \theta_2\| \leq \|\theta_1\| + \|\theta_2\|$

While previous works Folland (1999); Aliprantis & Border (2006) note that Cartesian products of normed spaces are also normed spaces, we were unable to find a complete proof that the particular function defined above acts as a norm on Θ. For completeness, we present the proof of Remark B.1 in Appendix C, as our derivation depends on this structure.

Before beginning our derivation of (55-58), note also that for any $\theta \in \Theta$, we define the gradient

$$\nabla g(\theta) = \nabla g(B, A) = \Big(\nabla_B g(B, A),\ \nabla_A g(B, A)\Big) \tag{66}$$

For a time-dependent path $\theta(t) = \Big(Y(t), X(t)\Big)$, we define

$$\frac{d}{dt}\left[\theta(t)\right] = \frac{d}{dt}\left[\Big(Y(t), X(t)\Big)\right] = \left(\frac{dY(t)}{dt}, \frac{dX(t)}{dt}\right) \tag{67}$$

as well as

$$\int_{t_1}^{t_2} \theta(s)\, ds = \left(\int_{t_1}^{t_2} Y(s)ds\,,\ \int_{t_1}^{t_2} X(s)ds\right). \tag{68}$$

We are now ready to begin our derivation.

Consider a continuously differentiable function $g : \Theta \to \mathbb{R}$, and suppose we wish to minimize g on Θ. This is equivalent to solving

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times m}}} g(B, A) \tag{69}$$

Consider the iterates for A and B output by Algorithm 1 applied to (69). At any given iteration i of Algorithm 1, stages 1 and 2 of the algorithm produce the sequences of iterates as

$$B^{(\alpha)}_{ik+j+1} = B^{(\alpha)}_{ik+j} - \alpha\nabla_B g\left(B^{(\alpha)}_{ik+j}, A^{(\alpha)}_{ik}\right) \tag{70}$$

$$A^{(\alpha)}_{ik+j+1} = A^{(\alpha)}_{ik+j} - \alpha\nabla_A g\left(\left[\lambda B^{(\alpha)}_{ik} + (1-\lambda)B^{(\alpha)}_{(i+1)k}\right], A^{(\alpha)}_{ik+j}\right) \tag{71}$$

We assume all iterates produced by Algorithm 1 lie within some compact domain, as illustrated below:

Assumption B.2 (Uniform Boundedness of Iterates)**.** *Define the domain $\mathcal{D}_R \subseteq \Theta$ by*

$$\mathcal{D}_R := \{(B, A) \in \Theta : \|B\| \le R,\ \|A\| \le R\}.$$

Then there exists $R' > 0$ such that, for any $ik + j \in \mathbb{N}_0$, we have

$$\left(B^{(\alpha)}_{ik+j},\ A^{(\alpha)}_{ik+j}\right) \in \mathcal{D}_{R'}.$$

The assumption above ensures all iterates remain within compact subsets of Θ on which all subsequent regularity assumptions for our analysis are imposed. While our analysis holds for any objective function satisfying Assumptions B.2 and B.7, we note in Remark B.3 below that Assumption B.2 holds for any Lipschitz smooth and bounded below objective given a sufficiently small finetuning stepsize:

Remark B.3 (Sufficient Condition for Assumption B.2)**.** *Suppose the objective function g is Lipschitz smooth and bounded below. That is, g satisfies Assumption B.7, and there exists some $M \in \mathbb{R}$ such that*

$$g(B, A) \ge M \quad \text{for all } (B, A) \in \Theta. \tag{72}$$

Then for any finite $T > 0$, there exist $R'_T > 0$ and $\alpha_0 > 0$, both depending only on T, k, $\theta(0)$, and g, such that for every $\alpha < \alpha_0$ the iterates of Algorithm 1 satisfy

$$\max\left(\left\|B^{(\alpha)}_{ik+j}\right\|, \left\|A^{(\alpha)}_{ik+j}\right\|\right) \le R'_T \tag{73}$$

for all $ik + j \in \mathbb{N}_0$ such that $(ik + j)\alpha \le T$. We provide proof for this remark in Appendix L.

To illustrate how the iterates evolve during training, construct their analogous continuous affine interpolations:

$$Y_\alpha(t) = B^{(\alpha)}_{\lfloor \frac{t}{\alpha}\rfloor} - \left(t - \left\lfloor\frac{t}{\alpha}\right\rfloor \alpha\right)\nabla_B g\left(B^{(\alpha)}_{\lfloor \frac{t}{\alpha}\rfloor}, A^{(\alpha)}_{h(t)}\right) \tag{74}$$

$$X_\alpha(t) = A^{(\alpha)}_{\lfloor \frac{t}{\alpha}\rfloor} - \left(t - \left\lfloor\frac{t}{\alpha}\right\rfloor \alpha\right)\nabla_A g\left(\left[\lambda B^{(\alpha)}_{h(t)} + (1-\lambda)B^{(\alpha)}_{h(t)+k}\right], A^{(\alpha)}_{\lfloor \frac{t}{\alpha}\rfloor}\right) \tag{75}$$

where

$$h(t) = \left\lfloor\frac{t}{\alpha}\right\rfloor - \left(\left\lfloor\frac{t}{\alpha}\right\rfloor \mod k\right) \tag{76}$$

Under the following boundedness assumption on the objective gradient, the continuous-time interpolations capture the behavior of Algorithm 1.

Remark B.4 (Uniform Boundedness of Gradient)**.** *For every finite $R > 0$, there exists $M_R > 0$ such that, for every $(B, A) \in \mathcal{D}_R$,*

$$\|\nabla g(B, A)\|^2 = \|\nabla_B g(B, A)\|^2 + \|\nabla_A g(B, A)\|^2 \le M_R^2, \tag{77}$$

This remark follows from Assumption B.7.

The remark above enables us to illustrate how $Y_\alpha(t)$ and $X_\alpha(t)$ capture the behavior of the iterates during training. Under Remark B.4, we will have, for any $ik + j \in \mathbb{N}_0$,

$$
\begin{aligned}
Y_\alpha\Big(t = (ik+j)\alpha\Big) &= B^{(\alpha)}_{\lfloor \frac{(ik+j)\alpha}{\alpha} \rfloor} \\
&\quad - \left((ik+j)\alpha - \left\lfloor \frac{(ik+j)\alpha}{\alpha} \right\rfloor \alpha\right) \nabla_B g\left(B^{(\alpha)}_{\lfloor \frac{(ik+j)\alpha}{\alpha} \rfloor}, A^{(\alpha)}_{h((ik+j)\alpha)}\right) && (78)\\
&= B^{(\alpha)}_{ik+j} - \Big((ik+j)\alpha - (ik+j)\alpha\Big)\nabla_B g\left(B^{(\alpha)}_{ik+j}, A^{(\alpha)}_{ik}\right) && (79)\\
&= B^{(\alpha)}_{ik+j} && (80)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&X_\alpha\Big(t = (ik+j)\alpha\Big) \\
&\quad = A^{(\alpha)}_{\lfloor \frac{(ik+j)\alpha}{\alpha} \rfloor} - \Big((ik+j)\alpha - \left\lfloor \frac{(ik+j)\alpha}{\alpha} \right\rfloor \alpha\Big) \\
&\qquad \cdot \nabla_A g\left(\lambda B^{(\alpha)}_{h((ik+j)\alpha)} + (1-\lambda) B^{(\alpha)}_{h((ik+j)\alpha)+k},\, A^{(\alpha)}_{\lfloor \frac{(ik+j)\alpha}{\alpha} \rfloor}\right) && (81)\\
&\quad = A^{(\alpha)}_{ik+j} - \Big((ik+j)\alpha - (ik+j)\alpha\Big)\nabla_A g\left(\lambda B^{(\alpha)}_{ik} + (1-\lambda) B^{(\alpha)}_{(i+1)k},\, A^{(\alpha)}_{ik+j}\right) && (82)\\
&\quad = A^{(\alpha)}_{ik+j} && (83)
\end{aligned}
$$

Thus, the iterates output by Algorithm 1 for B and A correspond exactly to the values of the continuous affine interpolations at the discretization points. Note also that $Y_\alpha(t)$ and $X_\alpha(t)$ lie within $\mathcal{D}_{R'}$ for all finetuning times.

Let $T > 0$ be any finite number. To show that the ODEs in (55–58) describe the dynamics of $Y_\alpha(t)$ and $X_\alpha(t)$ in the limit that α approaches zero for any $t \in [0, T]$, define $Y : [0, T] \to \mathbb{R}^{n\times r}$ and $X : [0, T] \to \mathbb{R}^{r\times m}$ as matrix-valued functions which satisfy (55–58). That is, $Y(t)$ and $X(t)$ are solutions on $[0, T]$ to

$$
\begin{aligned}
\frac{dY(t)}{dt} &= -\nabla_Y g\Big(Y(t), X(t)\Big) && (84)\\
\frac{dX(t)}{dt} &= -\nabla_X g\Big(Y(t), X(t)\Big) && (85)\\
Y(0) &= B_0 && (86)\\
X(0) &= A_0 && (87)
\end{aligned}
$$

which we can rewrite using our notation as

$$
\begin{aligned}
\frac{d\theta(t)}{dt} &= -\nabla g\Big(\theta(t)\Big) && (88)\\
\theta(0) &= \Big(B^{(\alpha)}_0, A^{(\alpha)}_0\Big) && (89)
\end{aligned}
$$

where $\theta(t) = \Big(Y(t), X(t)\Big)$. Remark B.5 explains how Remark B.4 guarantees existence of the solution on $[0, T]$.

Remark B.5 (Existence of ODE Solution)**.** *Since ∇g is continuous and bounded on any $\mathcal{D}_R$, the initial-value problem in (88–89) has a solution for all finite $t \geq 0$ Walter (1998). Moreover, any such solution remains bounded on finite time intervals. That is, for any finite $T > 0$, there exists some $R_T > 0$ such that*

$$
\Big(Y(t), X(t)\Big) \in \mathcal{D}_{R_T} \text{ for all } t \in [0, T]. \tag{90}
$$

The boundedness of ∇g on any $\mathcal{D}_R$, together with the existence and boundedness of solutions to (88–89), implies Remark B.6:

Remark B.6 (Uniform Boundedness of Gradient for Finite Time)**.** *Let finite $T > 0$ be given. For any $t \in [0, T]$, we will have that*

$$\Big(Y(t), X(t)\Big) \in \mathcal{D}_{R_t} \tag{91}$$

where $R_t = \max\Big(\, \|Y(t)\|, \|X(t)\| \,\Big)$. By Remark B.4, there exists $M_{R_t} > 0$ such that

$$\Big\|\nabla g\Big(\theta(t)\Big)\Big\| \leq M_{R_t} \tag{92}$$

Denote

$$M_T = \sup_{t\in[0,T]} \{M_{R_t}\} \tag{93}$$

Then we will have

$$\Big\|\nabla g\Big(\theta(t)\Big)\Big\| \leq M_T \text{ for all } t \in [0, T]. \tag{94}$$

To demonstrate that the gradient flow ODEs in (88–89) give the behavior of $Y_\alpha(t)$ and $X_\alpha(t)$ as α approaches zero, we will show that

$$\lim_{\alpha\to 0} ||Y_\alpha(t) - Y(t)|| = 0 \tag{95}$$

$$\lim_{\alpha\to 0} ||X_\alpha(t) - X(t)|| = 0 \tag{96}$$

for all $t \in [0, T]$. Or, equivalently, that

$$\lim_{\alpha\to 0} ||\theta_\alpha(t) - \theta(t)|| = 0 \tag{97}$$

where $\theta_\alpha(t) = \Big(Y_\alpha(t), X_\alpha(t)\Big)$. Begin by letting $q = ik + j$ be any non-negative integer satisfying $0 \leq q + 1 \leq \frac{T}{\alpha}$, and denote $t_q = q\alpha$. Note that $t_{q+1} - t_q = \alpha$. Integrating both sides of (88) with respect to t, we have (for any $\alpha > 0$)

$$\int_{t_q}^{t_{q+1}} \frac{d\theta(s)}{ds}\, ds = -\int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) ds \tag{98}$$

$$\theta(t_{q+1}) - \theta(t_q) = -\int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) ds \tag{99}$$

Add and subtract $\alpha\nabla g\Big(\theta(t_q)\Big)$ from the right-hand side to get

$$\theta(t_{q+1}) - \theta(t_q) = -\alpha\nabla g\Big(\theta(t_q)\Big) - \int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big)\, ds \tag{100}$$

$$\theta(t_{q+1}) - \Big[\theta(t_q) - \alpha\nabla g\Big(\theta(t_q)\Big)\Big] = -\int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big)\, ds \tag{101}$$

Define $\tau_q \in \Theta$, where

$$\tau_q := \theta(t_{q+1}) - \Big[\theta(t_q) - \alpha\nabla g\Big(\theta(t_q)\Big)\Big] \tag{102}$$

and note from (101) that

$$\tau_q = -\int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big)\, ds \tag{103}$$

To continue with our derivation, we need Assumption B.7 and Lemma B.8 below:

Assumption B.7 (Lipschitz Smoothness of g)**.** *Assume g is Lipschitz smooth on any $\mathcal{D}_R$. That is, for any finite $R > 0$, there exists $L_R > 0$ such that, for all $\theta_1, \theta_2 \in \mathcal{D}_R$, we have*

$$\|\nabla g(\theta_2) - \nabla g(\theta_1)\| \leq L_R ||\theta_2 - \theta_1|| \tag{104}$$

Lipschitz smoothness of ∇g on any $\mathcal{D}_R$ gives

$$\left\|\nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big)\right\| \leq L_{R_T} \|\theta(s) - \theta(t_q)\| \tag{105}$$

The following lemma allows us to bound $\|\theta(s) - \theta(t_q)\|$:

Lemma B.8 (Changes in $\theta(t)$ in Time are Bounded)**.** *Let $T > 0$ be an arbitrary finite number. For the solution to (88–89) which exists on $t \in [0, T]$, we will have*

$$\left\|\frac{d\theta(t)}{dt}\right\| = \left\|\nabla g\Big(\theta(t)\Big)\right\| \text{ for all } t \in [0, T]. \tag{106}$$

By Remark B.6, we have that

$$\left\|\frac{d\theta(t)}{dt}\right\| = \left\|\nabla g\Big(\theta(t)\Big)\right\| \leq M_T \text{ for all } t \in [0, T]. \tag{107}$$

Using Lemma B.8, the Fundamental Theorem of Calculus gives

$$\|\theta(s) - \theta(t_q)\| = \left\|\int_{t_q}^{s} \frac{d}{ds'}\theta(s')ds'\right\| \tag{108}$$

$$\leq \int_{t_q}^{s} \left\|\frac{d}{ds'}\theta(s')\right\| ds' \tag{109}$$

$$\leq \int_{t_q}^{t_{q+1}} \left\|\frac{d}{ds'}\theta(s')\right\| ds' \tag{110}$$

$$\leq \int_{t_q}^{t_{q+1}} M_T \, ds' \tag{111}$$

$$= M_T \alpha \tag{112}$$

and we have $\|\theta(s) - \theta(t_q)\| \leq M_T \alpha$.

Return to (103). Taking the norm of both sides, we have

$$\|\tau_q\| = \left\|\int_{t_q}^{t_{q+1}} \nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big) \, ds\right\| \tag{113}$$

$$\leq \int_{t_q}^{t_{q+1}} \left\|\nabla g\Big(\theta(s)\Big) - \nabla g\Big(\theta(t_q)\Big)\right\| ds \tag{114}$$

$$\leq L_{R_T} \int_{t_q}^{t_{q+1}} \|\theta(s) - \theta(t_q)\| \, ds \tag{115}$$

$$\leq L_{R_T} \int_{t_q}^{t_{q+1}} M_T \alpha \, ds \tag{116}$$

$$= L_{R_T} M_T \alpha^2 \tag{117}$$

So we have $\|\tau_q\| \leq L_{R_T} M_T \alpha^2$, where

1. R_T bounds $\|Y(t)\|$ and $\|X(t)\|$ for all $t \in [0, T]$.

2. L_{R_T} is the Lipschitz constant of $\nabla g(\theta)$ on $\mathcal{D}_{R_T}$.

3. M_T bounds $\left\|\nabla g\Big(\theta(t)\Big)\right\|$ for all $t \in [0, T]$.

Denote $C_T := L_{R_T} M_T$.

Now define

$$e_q := \theta(t_q) - \left(B_q^{(\alpha)}\,,\, A_q^{(\alpha)}\right) \tag{118}$$

$$\tag{119}$$

and rearrange (101–103) to get

$$\theta(t_{q+1}) = \theta(t_q) - \alpha \nabla g\Big(\theta(t_q)\Big) + \tau_q \tag{120}$$

The gradient descent iterations in (70–71) give

$$B_{q+1}^{(\alpha)} = B_q^{(\alpha)} - \alpha \nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right) \tag{121}$$

$$A_{q+1}^{(\alpha)} = A_q^{(\alpha)} - \alpha \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right) \tag{122}$$

We can then calculate e_{q+1} :

$$e_{q+1} := \theta(t_{q+1}) - \left(B_{q+1}^{(\alpha)}, A_{q+1}^{(\alpha)}\right) \tag{123}$$

$$\begin{aligned} &= \theta(t_q) - \alpha \nabla g\Big(\theta(t_q)\Big) + \tau_q \\ &- \left(B_q^{(\alpha)} - \alpha \nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right)\,,\, A_q^{(\alpha)} - \alpha \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) \end{aligned} \tag{124}$$

We can rewrite the ordered pair in (124) using the definitions of binary addition and scalar multiplication on Θ (see Remark B.1) as

$$\left(B_q^{(\alpha)} - \alpha \nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right)\,,\, A_q^{(\alpha)} - \alpha \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) \tag{125}$$

$$= \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \alpha \left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right), \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) \tag{126}$$

Then (124) becomes

$$\begin{aligned} e_{q+1} &= \theta(t_q) - \alpha \nabla g\Big(\theta(t_q)\Big) + \tau_q \\ &\quad - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) + \alpha \left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right), \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) \end{aligned} \tag{127}$$

$$\begin{aligned} &= \theta(t_q) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) + \tau_q \\ &\quad + \alpha \left[\left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right), \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) - \nabla g\Big(\theta(t_q)\Big)\right] \end{aligned} \tag{128}$$

$$\begin{aligned} &= e_q + \tau_q \\ &\quad + \alpha \left[\left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right), \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) - \nabla g\Big(\theta(t_q)\Big)\right] \end{aligned} \tag{129}$$

We can then bound the norm of e_{q+1}:

$$\|e_{q+1}\| \leq \|e_q\| + \|\tau_q\| + \alpha \left\| \left(\nabla_B g \left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)} \right), \nabla_A g \left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} \right], A_q^{(\alpha)} \right) \right) - \nabla g\Big(\theta(t_q)\Big) \right\| \tag{130}$$

For brevity, denote

$$B_1 := B_q^{(\alpha)} \tag{131}$$
$$A_1 := A_{q-j}^{(\alpha)} \tag{132}$$
$$B_2 := \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} \tag{133}$$
$$A_2 := A_q^{(\alpha)} \tag{134}$$
$$\delta_B := \nabla_B g(B_1, A_1) - \nabla_Y g\Big(Y(t_q), X(t_q)\Big) \tag{135}$$
$$\delta_A := \nabla_A g(B_2, A_2) - \nabla_X g\Big(Y(t_q), X(t_q)\Big) \tag{136}$$

We bound the norm of the third term in (130) above:

$$\left\| \left(\nabla_B g(B_1, A_1), \nabla_A g(B_2, A_2) \right) - \nabla g\Big(\theta(t_q)\Big) \right\| \tag{137}$$
$$= \Big\| \Big(\delta_B \, , \, \delta_A \Big) \Big\| \tag{138}$$
$$= \Big\| \Big(\delta_B \, , \, \mathbf{0}_{r \times m} \Big) + \Big(\mathbf{0}_{n \times r} \, , \, \delta_A \Big) \Big\| \tag{139}$$
$$\leq \Big\| \Big(\delta_B \, , \, \mathbf{0}_{r \times m} \Big) \Big\| + \Big\| \Big(\mathbf{0}_{n \times r} \, , \, \delta_A \Big) \Big\| \tag{140}$$
$$\leq \Big\| \nabla g(B_1, A_1) - \nabla g\Big(\theta(t_q)\Big) \Big\| + \Big\| \nabla g(B_2, A_2) - \nabla g\Big(\theta(t_q)\Big) \Big\| \tag{141}$$

By Assumption B.2 and Remark B.5, we know that

$$\left\| B_q^{(\alpha)} \right\| \leq R' \tag{142}$$
$$\left\| A_q^{(\alpha)} \right\| \leq R' \tag{143}$$
$$\left\| A_{q-j}^{(\alpha)} \right\| \leq R' \tag{144}$$
$$\left\| B_{q-j}^{(\alpha)} \right\| \leq R' \tag{145}$$
$$\left\| B_{q-j+k}^{(\alpha)} \right\| \leq R' \tag{146}$$
$$\|Y(t_q)\| \leq R_T \tag{147}$$
$$\|X(t_q)\| \leq R_T \tag{148}$$

We also have

$$\left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} \right\| \leq \lambda \left\| B_{q-j}^{(\alpha)} \right\| + (1-\lambda) \left\| B_{q-j+k}^{(\alpha)} \right\| \leq \lambda R' + (1-\lambda) R' = R' \tag{149}$$

Let $R_0 = \max(R', R_T)$. Then the ordered pairs above satisfy

$$\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)} \right) \in \mathcal{D}_{R_0} \tag{150}$$
$$\left(\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} \, , \, A_q^{(\alpha)} \right) \in \mathcal{D}_{R_0} \tag{151}$$
$$\Big(Y(t_q), X(t_q) \Big) \in \mathcal{D}_{R_0} \tag{152}$$

By the Lipschitz smoothness assumption in B.7, this gives the existence of finite $L_{R_0} > 0$ such that

$$\left\|\nabla g(B_1, A_1) - \nabla g\Big(\theta(t_q)\Big)\right\| \leq L_{R_0} \left\|(B_1, A_1) - \theta(t_q)\right\| \tag{153}$$

$$\left\|\nabla g(B_2, A_2) - \nabla g\Big(\theta(t_q)\Big)\right\| \leq L_{R_0} \left\|(B_2, A_2) - \theta(t_q)\right\| \tag{154}$$

Return to (141). We have

$$\left\|\left(\nabla_B g\left(B_1, A_1\right), \nabla_A g\left(B_2, A_2\right)\right) - \nabla g\Big(\theta(t_q)\Big)\right\| \tag{155}$$

$$\leq \left\|\nabla g(B_1, A_1) - \nabla g\Big(\theta(t_q)\Big)\right\| + \left\|\nabla g(B_2, A_2) - \nabla g\Big(\theta(t_q)\Big)\right\| \tag{156}$$

$$\leq L_{R_0} \left\|(B_1, A_1) - \theta(t_q)\right\| + L_{R_0} \left\|(B_2, A_2) - \theta(t_q)\right\| \tag{157}$$

$$= L_{R_0}\left(\left\|(B_1, A_1) - \theta(t_q)\right\| + \left\|(B_2, A_2) - \theta(t_q)\right\|\right) \tag{158}$$

To bound the normed differences above, we can calculate

$$\left\|(B_1, A_1) - \theta(t_q)\right\| = \left\|(B_1, A_1) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) + \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q)\right\| \tag{159}$$

$$\leq \left\|(B_1, A_1) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + \left\|\left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q)\right\| \tag{160}$$

$$= \left\|(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + \|e_q\| \tag{161}$$

$$= \left\|A_{q-j}^{(\alpha)} - A_q^{(\alpha)}\right\| + \|e_q\| \tag{162}$$

Similarly, we have

$$\left\|(B_2, A_2) - \theta(t_q)\right\| = \left\|(B_2, A_2) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) + \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q)\right\| \tag{163}$$

$$\leq \left\|(B_2, A_2) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + \left\|\left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q)\right\| \tag{164}$$

$$= \left\|\left(\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\,,\, A_q^{(\alpha)}\right) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + \|e_q\| \tag{165}$$

$$= \left\|\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)}\right\| + \|e_q\| \tag{166}$$

So we get

$$\left\|(B_1, A_1) - \theta(t_q)\right\| \leq \left\|A_{q-j}^{(\alpha)} - A_q^{(\alpha)}\right\| + \|e_q\| \tag{167}$$

$$\left\|(B_2, A_2) - \theta(t_q)\right\| \leq \left\|\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)}\right\| + \|e_q\| \tag{168}$$

Recall our update equations for the gradient descent iterates:

$$B_{q+1}^{(\alpha)} = B_q^{(\alpha)} - \alpha \nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right) \tag{169}$$

$$A_{q+1}^{(\alpha)} = A_q^{(\alpha)} - \alpha \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right) \tag{170}$$

By the bounded gradient assumption from B.4, there exists M_{R_0} such that

$$\left\|B_{q+1}^{(\alpha)} - B_q^{(\alpha)}\right\| \leq \alpha M_{R_0} \tag{171}$$

$$\left\|A_{q+1}^{(\alpha)} - A_q^{(\alpha)}\right\| \leq \alpha M_{R_0} \tag{172}$$

We can then calculate

$$\left\|A_q^{(\alpha)} - A_{q-j}^{(\alpha)}\right\| = \left\|A_q^{(\alpha)} - A_{q-j+1}^{(\alpha)} + A_{q-j+1}^{(\alpha)} - A_{q-j}^{(\alpha)}\right\| \tag{173}$$

$$\leq \left\|A_q^{(\alpha)} - A_{q-j+1}^{(\alpha)}\right\| + \left\|A_{q-j+1}^{(\alpha)} - A_{q-j}^{(\alpha)}\right\| \tag{174}$$

$$\leq \left\| A_q^{(\alpha)} - A_{q-j+1}^{(\alpha)} \right\| + \alpha M_{R_0} \tag{175}$$

Add and subtract the next iterate and apply the triangle inequality $j-1$ more times to obtain

$$\left\| A_q^{(\alpha)} - A_{q-j}^{(\alpha)} \right\| \leq j\alpha M_{R_0}. \tag{176}$$

Or, more explicitly, we have

$$A_q^{(\alpha)} - A_{q-j}^{(\alpha)} = \sum_{\ell=1}^{j} \left(A_{q-\ell+1}^{(\alpha)} - A_{q-\ell}^{(\alpha)} \right) \tag{177}$$

which gives

$$\|A_q^{(\alpha)} - A_{q-j}^{(\alpha)}\| \leq \sum_{\ell=1}^{j} \|A_{q-\ell+1}^{(\alpha)} - A_{q-\ell}^{(\alpha)}\| \tag{178}$$

$$\leq \sum_{\ell=1}^{j} \alpha M_{R_0} \tag{179}$$

$$= j\alpha M_{R_0}. \tag{180}$$

For $\left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\|$, we obtain

$$\left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| = \left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - \lambda B_q^{(\alpha)} - (1-\lambda) B_q^{(\alpha)} \right\| \tag{181}$$

$$\leq \lambda \left\| B_q^{(\alpha)} - B_{q-j}^{(\alpha)} \right\| + (1-\lambda) \left\| B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| \tag{182}$$

The same telescoping argument from (177–180) gives

$$\left\| B_q^{(\alpha)} - B_{q-j}^{(\alpha)} \right\| \leq j\alpha M_{R_0} \tag{183}$$

$$\left\| B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| \leq (k-j)\alpha M_{R_0} \tag{184}$$

Thus, we have

$$\left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| \leq \lambda \left\| B_q^{(\alpha)} - B_{q-j}^{(\alpha)} \right\| + (1-\lambda) \left\| B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| \tag{185}$$

$$\leq \lambda j \alpha M_{R_0} + (1-\lambda)(k-j)\alpha M_{R_0} \tag{186}$$

$$= \alpha M_{R_0} \left[\lambda j + (1-\lambda)(k-j) \right] \tag{187}$$

$$\leq \alpha M_{R_0} \left[\lambda k + (1-\lambda)(k-j) \right] \tag{188}$$

$$\leq \alpha M_{R_0} \left[\lambda k + (1-\lambda) k \right] \tag{189}$$

$$= k\alpha M_{R_0} \tag{190}$$

We have then obtained

$$\left\| A_{q-j}^{(\alpha)} - A_q^{(\alpha)} \right\| \leq j\alpha M_{R_0} \tag{191}$$

$$\left\| \lambda B_{q-j}^{(\alpha)} + (1-\lambda) B_{q-j+k}^{(\alpha)} - B_q^{(\alpha)} \right\| \leq k\alpha M_{R_0} \tag{192}$$

With the above bounds in mind, (167–168) become

$$\|(B_1, A_1) - \theta(t_q)\| \leq j\alpha M_{R_0} + \|e_q\| \tag{193}$$

$$\|(B_2, A_2) - \theta(t_q)\| \leq k\alpha M_{R_0} + \|e_q\| \tag{194}$$

Returning to (158), we finally obtain the upper bound on the third term in (130):

$$\left\| \left(\nabla_B g\left(B_1, A_1\right), \nabla_A g\left(B_2, A_2\right) \right) - \nabla g\Big(\theta(t_q) \Big) \right\| \tag{195}$$

$$\leq L_{R_0}\left(\|(B_1, A_1) - \theta(t_q)\| + \|(B_2, A_2) - \theta(t_q)\| \right) \tag{196}$$

$$\leq L_{R_0}\left(j\alpha M_{R_0} + \|e_q\| + k\alpha M_{R_0} + \|e_q\| \right) \tag{197}$$

$$= L_{R_0}\left(2\|e_q\| + (j+k)\alpha M_{R_0} \right) \tag{198}$$

The upper bound on $\|e_{q+1}\|$ finally becomes

$$\|e_{q+1}\| \leq \|e_q\| + \|\tau_q\| + \alpha L_{R_0}\left(2\|e_q\| + (j+k)\alpha M_{R_0} \right) \tag{199}$$

$$= \|e_q\| + \|\tau_q\| + 2\alpha L_{R_0}\|e_q\| + (j+k)M_{R_0}L_{R_0}\alpha^2 \tag{200}$$

$$= (1 + 2\alpha L_{R_0})\|e_q\| + \|\tau_q\| + (j+k)M_{R_0}L_{R_0}\alpha^2 \tag{201}$$

Denote $L' := 2L_{R_0}$ (note $L' > 0$) and recall the bound on $\|\tau_q\|$ from (117). We have

$$\|e_{q+1}\| \leq (1 + \alpha L')\|e_q\| + L_{R_T}M_T\alpha^2 + (j+k)M_{R_0}L_{R_0}\alpha^2 \tag{202}$$

$$= (1 + \alpha L')\|e_q\| + \Big(L_{R_T}M_T + (j+k)M_{R_0}L_{R_0} \Big)\alpha^2 \tag{203}$$

Denote

$$C \coloneqq L_{R_T}M_T + (j+k)M_{R_0}L_{R_0}, \tag{204}$$

and we arrive at

$$\|e_{q+1}\| \leq (1 + \alpha L')\|e_q\| + C\alpha^2 \tag{205}$$

We can then recursively find an upper bound on $\|e_q\|$ in terms of $\|e_0\|$:

$$\|e_q\| \leq (1 + \alpha L')\|e_{q-1}\| + C\alpha^2 \tag{206}$$

$$\leq (1 + \alpha L')\Big((1 + \alpha L')\|e_{q-2}\| + C\alpha^2 \Big) + C\alpha^2 \tag{207}$$

$$\leq (1 + \alpha L')\left((1 + \alpha L')\Big((1 + \alpha L')\|e_{q-3}\| + C\alpha^2 \Big) + C\alpha^2 \right) + C\alpha^2 \tag{208}$$

Applying the recursion q times and collecting the coefficients of each $C\alpha^2$ term produces

$$\|e_q\| \leq (1 + \alpha L')^q\|e_0\| + C\alpha^2\sum_{p=0}^{q-1}(1 + \alpha L')^{q-1-p} \tag{209}$$

Recall from (89) that the initial conditions for our iterates and the ODE solution are the same, yielding $\|e_0\| = 0$. We then have

$$\|e_q\| \leq C\alpha^2\sum_{p=0}^{q-1}(1 + \alpha L')^{q-1-p} \tag{210}$$

which we can re-index

$$\|e_q\| \leq C\alpha^2\sum_{p=0}^{q-1}(1 + \alpha L')^{p} \tag{211}$$

Sum the finite geometric series above to find

$$\sum_{p=0}^{q-1}(1 + \alpha L')^p = \frac{1 - (1 + \alpha L')^q}{1 - (1 + \alpha L')} = \frac{(1 + \alpha L')^q - 1}{\alpha L'} \leq \frac{e^{q\alpha L'} - 1}{\alpha L'} \tag{212}$$

where we obtain the inequality by treating $(1+\alpha L')$ as the truncated Taylor series for $e^{\alpha L'}$. Then (211) becomes

$$\|e_q\| \leq C\frac{e^{q\alpha L'}-1}{L'}\alpha \tag{213}$$

Denote

$$G \coloneqq C\frac{e^{q\alpha L'}-1}{L'}, \tag{214}$$

and we have

$$\left\|\theta(t_q) - \left(B_q^{(\alpha)}\,,\, A_q^{(\alpha)}\right)\right\| \leq G\alpha \tag{215}$$

for any $q = ik + j$ satisfying $0 \leq q+1 \leq \frac{T}{\alpha}$.

Divide $[0,T]$ into the sub-intervals $[t_q, t_{q+1}]$. For any $t \in [t_q, t_{q+1}]$, we have

$$\|\theta_\alpha(t) - \theta(t)\| = \left\|\theta_\alpha(t) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) + \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q) + \theta(t_q) - \theta(t)\right\| \tag{216}$$

$$\leq \left\|\theta_\alpha(t) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + \left\|\left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) - \theta(t_q)\right\| + \|\theta(t_q) - \theta(t)\| \tag{217}$$

$$\leq \left\|\theta_\alpha(t) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + G\alpha + \|\theta(t_q) - \theta(t)\| \tag{218}$$

Bound the first term by noting that $\theta_\alpha(t)$ is the affine interpolation between $\left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)$ and $\left(B_{q+1}^{(\alpha)}, A_{q+1}^{(\alpha)}\right)$. Thus, $\left\lfloor \frac{t}{\alpha} \right\rfloor \alpha = q\alpha = t_q$. From (74–75), we have

$$\theta_\alpha(t) = \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right) \tag{219}$$

$$- (t - t_q)\left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right)\,,\, \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda)B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right) \tag{220}$$

Note from Assumption B.2 that $\theta_a(t)$ and $\left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)$ are elements of $\mathcal{D}_{R'}$. By the boundedness of the gradient assumption in B.4, we will have

$$\left\|\left(\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right)\,,\, \nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda)B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right)\right\| \tag{221}$$

$$= \sqrt{\left\|\nabla_B g\left(B_q^{(\alpha)}, A_{q-j}^{(\alpha)}\right)\right\|^2 + \left\|\nabla_A g\left(\left[\lambda B_{q-j}^{(\alpha)} + (1-\lambda)B_{q-j+k}^{(\alpha)}\right], A_q^{(\alpha)}\right)\right\|^2} \tag{222}$$

$$\leq \sqrt{M_{R'}^2 + M_{R'}^2} \tag{223}$$

$$= \sqrt{2} M_{R'} \tag{224}$$

Since $t \in [t_q, t_{q+1}]$, we will also have $t - t_q \leq \alpha$. Thus, from (220), we get

$$\left\|\theta_\alpha(t) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| \leq \sqrt{2} M_{R'}\alpha \tag{225}$$

To bound the third term, use the fact that

$$\frac{d\theta(t)}{dt} = -\nabla g\Big(\theta(t)\Big) \tag{226}$$

From Remark B.6, we have that both sides of the ODE above are bounded by M_T. We then get

$$\|\theta(t_q) - \theta(t)\| = \|\theta(t) - \theta(t_q)\| = \left\|\int_{t_q}^{t} \frac{d\theta(s)}{ds}\, ds\right\| \leq \int_{t_q}^{t} \left\|\frac{d\theta(s)}{ds}\right\| ds \leq M_T \int_{t_q}^{t} ds \tag{227}$$

$$= M_T(t - t_q) \leq M_T\alpha \tag{228}$$

From (225) and (228), we arrive at

$$\|\theta_\alpha(t) - \theta(t)\| \leq \left\|\theta_\alpha(t) - \left(B_q^{(\alpha)}, A_q^{(\alpha)}\right)\right\| + G\alpha + \|\theta(t_q) - \theta(t)\| \tag{229}$$

$$\leq \sqrt{2}M_{R'}\alpha + G\alpha + M_T\alpha \tag{230}$$

Thus, we have

$$0 \leq \|\theta_\alpha(t) - \theta(t)\| \leq \left(\sqrt{2}M_{R'} + G + M_T\right)\alpha \tag{231}$$

for all $t \in [t_q, t_{q+1}]$, where $[t_q, t_{q+1}] \subseteq [0, T]$. Note that (231) shows that the error between the discrete iterates and the gradient flow is bounded by a term linear in α. Furthermore, since the constant in (204) is affine in k, this bound grows at most linearly in k. We therefore conclude that the error between the discrete iterates and the gradient flow is $O(k\alpha)$.

Since the intervals $[t_q, t_{q+1}]$ cover $[0, T]$, the inequality in (231) holds for all $t \in [0, T]$. Taking the limit of (231) as $\alpha \to 0$, we finally have

$$\lim_{\alpha \to 0} \|\theta_\alpha(t) - \theta(t)\| = 0 \tag{232}$$

for all $t \in [0, T]$, where $T > 0$ is any finite number. We have therefore demonstrated that the continuous affine interpolation connecting the iterates of Algorithm 1 follow the behavior described by the LoRA gradient flow ODEs given in (84–87) in the limit that α approaches zero.

C PROOF OF REMARK B.1

Define

$$\Theta := \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m},$$

with elements $\theta \in \Theta$ given as ordered pairs of matrices:

$$\theta = (B, A), \tag{233}$$

$$B \in \mathbb{R}^{n\times r} \tag{234}$$

$$A \in \mathbb{R}^{r\times m} \tag{235}$$

Define the product norm for elements $\theta \in \Theta$ via

$$\|\theta\| := \left(\|B\|^2 + \|A\|^2\right)^{1/2}.$$

We wish to show that for any $\theta \in \Theta$, $\|\theta\|$ is a norm on Θ. That is,

1. $\|\theta\| \geq 0$
2. $\|\theta\| = 0$ iff $\theta = (\mathbf{0}_{n\times r}, \mathbf{0}_{r\times m})$
3. Define scalar multiplication on Θ as $\eta\theta = (\eta B, \eta A)$. Then $\|\eta\theta\| = |\eta| \, \|\theta\|$ for all $\eta \in \mathbb{R}$
4. For any two elements $\theta_1 = (B_1, A_1)$ and $\theta_2 = (B_2, A_2)$ of Θ, define the binary addition operation $\Theta \times \Theta \to \Theta$ as

$$\theta_1 + \theta_2 = (B_1 + B_2,\ A_1 + A_2) \tag{236}$$

   Then $\|\theta_1 + \theta_2\| \leq \|\theta_1\| + \|\theta_2\|$

Proof:
Let $\theta = (A, B)$ be given. Then

$$\|\theta\| = \left(\|B\|^2 + \|A\|^2\right)^{1/2} \tag{237}$$

By properties of matrix norms, we have $\|B\| \geq 0$ and $\|A\| \geq 0$. Then $\|B\|^2 \geq 0$ and $\|A\|^2 \geq 0$. This gives

$$\|B\|^2 + \|A\|^2 \geq 0 \tag{238}$$

$$\left(\|B\|^2 + \|A\|^2\right)^{1/2} \geq 0 \tag{239}$$

and we have $\|\theta\| \geq 0$. This concludes the proof of 1.

Now suppose $\|\theta\| = 0$. Then

$$\left(\|B\|^2 + \|A\|^2\right)^{1/2} = 0 \tag{240}$$

$$\|B\|^2 + \|A\|^2 = 0 \tag{241}$$

$$\|B\|^2 = -\|A\|^2 \tag{242}$$

Now denote $c = \|B\|^2 = -\|A\|^2$. Since $\|B\|^2 \geq 0$, we have $c \geq 0$. On the other hand, $-\|A\|^2 \leq 0$ gives us $c \leq 0$. Thus, $c = \|B\|^2 = -\|A\|^2 = 0$. By properties of matrix norms, we then have

$$B = \mathbf{0}_{n\times r} \tag{243}$$

$$A = \mathbf{0}_{r\times m} \tag{244}$$

and $\theta = (\mathbf{0}_{n\times r}, \mathbf{0}_{r\times m})$.

Suppose now that $\theta = (\mathbf{0}_{n\times r}, \mathbf{0}_{r\times m})$. Then

$$\|\theta\| = \left(\|\mathbf{0}_{n\times r}\|^2 + \|\mathbf{0}_{r\times m}\|^2\right)^{1/2} \tag{245}$$

$$= \left(0^2 + 0^2\right)^{1/2} \tag{246}$$

$$= 0 \tag{247}$$

and we have proven 2.

For any real number $\eta \in \mathbb{R}$, we will have

$$\|\eta\theta\| = \left(\|\eta B\|^2 + \|\eta A\|^2\right)^{1/2} \tag{248}$$

$$= \left(\eta^2 \|B\|^2 + \eta^2 \|A\|^2\right)^{1/2} \tag{249}$$

$$= |\eta| \left(\|B\|^2 + \|A\|^2\right)^{1/2} \tag{250}$$

$$= |\eta| \|\theta\| \tag{251}$$

and we have 3.

Finally, for any $\theta_1, \theta_2 \in \Theta$, denote

$$\theta_1 = (B_1, A_1) \tag{252}$$

$$\theta_2 = (B_2, A_2) \tag{253}$$

We have

$$\|\theta_1 + \theta_2\| = \|(B_1 + B_2,\ A_1 + A_2)\| \tag{254}$$

$$= \left(\|B_1 + B_2\|^2 + \|A_1 + A_2\|^2\right)^{1/2} \tag{255}$$

Construct $u, v \in \mathbb{R}^2$, where

$$u = \left(\|B_1\|, \|A_1\|\right) \tag{256}$$

$$v = \left(\|B_2\|, \|A_2\|\right) \tag{257}$$

and note that

$$\|u\|_2 = \left(\|B_1\|^2 + \|A_1\|^2\right)^{1/2} \tag{258}$$

$$\|v\|_2 = \left(\|B_2\|^2 + \|A_2\|^2\right)^{1/2} \tag{259}$$

$$\|u + v\|_2 = \left((\|B_1\| + \|B_2\|)^2 + (\|A_1\| + \|A_2\|)^2\right)^{1/2} \tag{260}$$

where $\|\cdot\|_2$ denotes the vector Euclidean norm. The triangle inequality gives

$$\|u + v\|_2 \leq \|u\|_2 + \|v\|_2 \tag{261}$$

Putting this all together, we have

$$\|\theta_1 + \theta_2\| = \|(B_1 + B_2,\ A_1 + A_2)\| \tag{262}$$

$$= \left(\|B_1 + B_2\|^2 + \|A_1 + A_2\|^2\right)^{1/2} \tag{263}$$

$$\leq \left(\|B_1\|^2 + 2\|B_1\|\|B_2\| + \|B_2\|^2 + \|A_1\|^2 + 2\|A_1\|\|A_2\| + \|A_2\|^2\right)^{1/2} \tag{264}$$

$$= \left((\|B_1\| + \|B_2\|)^2 + (\|A_1\| + \|A_2\|)^2\right)^{1/2} \tag{265}$$

$$\leq \left(\|B_1\|^2 + \|A_1\|^2\right)^{1/2} + \left(\|B_2\|^2 + \|A_2\|^2\right)^{1/2} \tag{266}$$

$$= \|\theta_1\| + \|\theta_2\| \tag{267}$$

and the proof of Remark B.1 is complete.

D LEARNING DYNAMICS FOR TRACE-SQUARED LOSS (LOW-RANK)

We wish to analyze the learning dynamics of the low-rank optimizer $BA \in \mathbb{R}^{n\times n}$ produced by applying LoRA to the finetuning problem

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \frac{1}{2}\operatorname{Tr}^2(W^* - BA) \tag{268}$$

where $r << n$. The partial gradients for our objective $g(B,A) = \frac{1}{2}\operatorname{Tr}^2(W^* - BA)$ are given by

$$\nabla_A g(B,A) = -\operatorname{Tr}(W^* - BA)B^T \tag{269}$$

$$\nabla_B g(B,A) = -\operatorname{Tr}(W^* - BA)A^T \tag{270}$$

Assumption D.1 (Bounded Domain)**.** *We optimize (268) over the subspace $\mathcal{D}_{R'} \subseteq \Theta$, where $R' > 0$ is some finite number. In other words, we assume that there exists some $R' > 0$ such that the norms on both B and A remain bounded above by R' during training.*

Note that, since the trace-squared objective is bounded below by 0, that Assumption D.1 holds for this problem on finite finetuning horizons for sufficiently small stepsize as outlined in Remark 2.2. This assumption ensures that the objective gradient remains bounded during training:

Remark D.2 (Boundedness of Gradient for Trace Squared Loss)**.** *The objective gradient $\nabla g(B,A)$ remains bounded above during training. That is, for all $(B,A) \in \mathcal{D}_{R'}$, we have*

$$\begin{aligned}
\|\nabla g(B,A)\|^2 &= \|\nabla_B g(B,A)\|^2 + \|\nabla_A g(B,A)\|^2 && (271)\\
&= \left\|-\operatorname{Tr}(W^* - BA)A^T\right\|^2 + \left\|-\operatorname{Tr}(W^* - BA)B^T\right\|^2 && (272)\\
&\le \operatorname{Tr}^2(W^* - BA)\Big(\left\|A^T\right\|^2 + \left\|B^T\right\|^2\Big) && (273)\\
&\le 2R'^2\operatorname{Tr}^2(W^* - BA) && (274)\\
&\le 2R'^2\|W^* - BA\|^2 && (275)\\
&\le 2R'^2\Big(\|W^*\| + \|BA\|\Big)^2 && (276)\\
&\le 2R'^2\Big(\|W^*\| + \|B\|\,\|A\|\Big)^2 && (277)\\
&\le 2R'^2\Big(\|W^*\| + R'^2\Big)^2 && (278)
\end{aligned}$$

So $\|\nabla g(B,A)\| \le \sqrt{2}R'\Big(\|W^*\| + R'^2\Big)$ throughout training.

We then have Lipschitz smoothness during training:

Lemma D.3 (Lipschitz Smoothness for Trace Squared Objective)**.** *Our objective gradient ∇g is Lipschitz smooth in our training domain. Namely, there exists $L_{R'} > 0$ such that, for any $(B_1,A_1),(B_2,A_2) \in \mathcal{D}_{R'}$, we have*

$$\|\nabla g(B_1,A_1) - \nabla g(B_2,A_2)\| \le L_{R'}\left\|\Big(B_1,A_1\Big) - \Big(B_2,A_2\Big)\right\| \tag{279}$$

Proof of this lemma can be found in Appendix E.

Having shown that the trace squared objective in (268) satisfies Assumptions B.2 and B.7 from Appendix B, the ODEs describing the learning dynamics of $g(B,A)$ under LoRA are given by (see appendix B)

$$\frac{dY(t)}{dt} = \operatorname{Tr}(W^* - YX)X^T \tag{280}$$

$$\frac{dX(t)}{dt} = \operatorname{Tr}(W^* - YX)Y^T \tag{281}$$

for any $t \in [0, T]$, where $T > 0$ is arbitrary. Denote the initial conditions for this problem as

$$Y(0) = Y_0 \tag{282}$$

$$X(0) = X_0 \tag{283}$$

and note by application of Remark B.5 that a solution to the ODE above exists for all $t \in [0, T]$. Evaluating (280-281) at $t = 0$ yields the initial derivatives

$$\frac{dY(t)}{dt}\Big|_{t=0} = \operatorname{Tr}(W^* - Y_0 X_0) X_0^T \tag{284}$$

$$\frac{dX(t)}{dt}\Big|_{t=0} = \operatorname{Tr}(W^* - Y_0 X_0) Y_0^T \tag{285}$$

Following the initialization scheme commonly used in LoRA literature Hu et al. (2022); Xu et al. (2025), we initialize $Y_0 = \mathbf{0}_{n \times r}$, where $\mathbf{0}_{n \times r}$ denotes the $n \times r$ matrix of all zeroes. Thus, all calculations going forward assume $Y_0 = \mathbf{0}_{n \times r}$. We initialize the elements of X_0 from a random normal distribution centered at zero with variance σ^2. In other words,

$$x_{ij}(0) \sim \mathcal{N}(0, \sigma^2) \tag{286}$$

where x_{ij} are the individual elements of X.

Assumption D.4 (Nonzero X_0)**.** *Assume $x_{ij}(0) \neq 0$ for at least one $x_{ij}(0) \in X_0$, which holds almost surely for the Gaussian initialization in* (286)*.*

Assumption D.5 (Nonzero W^*)**.** *The trace of the matrix of prefrozen weights in nonzero. In other words,* $\operatorname{Tr}(W^*) \neq 0$*.*

Violation of either Assumption $D.4$ or D.5 causes the LoRA dynamics to stall at the saddle point $(\mathbf{0}_{n \times r}, \mathbf{0}_{r \times n})$, yielding $\mathbf{0}_{n \times n}$ as the optimizing matrix in either case. Our gradient flow analysis then proves trivial. To converge to the rank r optimizer, our weight matrix and initial conditions must satisfy Assumptions D.4 and D.5.

With proper initialization, we proceed with our solution to the problem described in (280-285) by first solving for the dynamics of $\operatorname{Tr}(W^* - YX)$ for $t \in [0, T]$. Right multiply both sides of (280) by X, and left multiply both sides of (281) by Y to get

$$\frac{dY}{dt} X = \operatorname{Tr}(W^* - YX) X^T X \tag{287}$$

$$Y \frac{dX}{dt} = \operatorname{Tr}(W^* - YX) Y Y^T \tag{288}$$

Add together the two equations above to get

$$\frac{dY}{dt} X + Y \frac{dX}{dt} = \operatorname{Tr}(W^* - YX) \left[X^T X + Y Y^T \right] \tag{289}$$

$$\frac{d(YX)}{dt} = \operatorname{Tr}(W^* - YX) \left[X^T X + Y Y^T \right] \tag{290}$$

$$-\frac{d(YX)}{dt} = -\operatorname{Tr}(W^* - YX) \left[X^T X + Y Y^T \right] \tag{291}$$

Since $\dfrac{dW^*}{dt} = 0$ (pretrained weights are frozen), this is equivalent to

$$\frac{d}{dt}(W^* - YX) = -\operatorname{Tr}(W^* - YX) \left[X^T X + Y Y^T \right] \tag{292}$$

Take the trace of the both sides of (292) to arrive at

$$\frac{d}{dt} \operatorname{Tr}(W^* - YX) = -\operatorname{Tr}(W^* - YX) \operatorname{Tr}\left(X^T X + Y Y^T \right) \tag{293}$$

Apply the $\frac{d}{dt}$ operator to both sides of the equation above to arrive at

$$\frac{d^2}{dt^2} \operatorname{Tr}(W^* - YX) \tag{294}$$

$$= -\frac{d}{dt}\left[\operatorname{Tr}(W^* - YX)\right]\operatorname{Tr}\left(X^TX + YY^T\right) - \operatorname{Tr}(W^* - YX)\frac{d}{dt}\left[\operatorname{Tr}\left(X^TX + YY^T\right)\right] \quad (295)$$

To derive an ODE for $\operatorname{Tr}(W^* - YX)$, it remains to find expressions for $\frac{d}{dt}\operatorname{Tr}\left(X^TX + YY^T\right)$ and $\operatorname{Tr}\left(X^TX + YY^T\right)$ in terms of $\operatorname{Tr}(W^* - YX)$. Return to (280). Right multiply both sides of our ODE by Y^T to find

$$\frac{dY}{dt}Y^T = \operatorname{Tr}(W^* - YX)X^TY^T \quad (296)$$

Take the transpose of both sides to get

$$Y\frac{dY^T}{dt} = \operatorname{Tr}(W^* - YX)YX \quad (297)$$

Add (296) and (297) together to find

$$\frac{dY}{dt}Y^T + Y\frac{dY^T}{dt} = \operatorname{Tr}(W^* - YX)\left(X^TY^T + YX\right) \quad (298)$$

$$\frac{d\left(YY^T\right)}{dt} = \operatorname{Tr}(W^* - YX)\left(X^TY^T + YX\right) \quad (299)$$

Taking the trace of both sides, we have

$$\frac{d}{dt}\operatorname{Tr}\left(YY^T\right) = \operatorname{Tr}(W^* - YX)\operatorname{Tr}\left(X^TY^T + YX\right) \quad (300)$$

$$= 2\operatorname{Tr}(W^* - YX)\operatorname{Tr}\left(YX\right) \quad (301)$$

The same calculation for (281) immediately gives

$$\frac{d}{dt}\operatorname{Tr}\left(X^TX\right) = 2\operatorname{Tr}(W^* - YX)\operatorname{Tr}\left(YX\right) \quad (302)$$

Add together (301) and (302) to get

$$\frac{d}{dt}\operatorname{Tr}\left(YY^T + X^TX\right) = 4\operatorname{Tr}(W^* - YX)\operatorname{Tr}(YX) \quad (303)$$

We now find an expression for $\operatorname{Tr}\left(X^TX + YY^T\right)$ in terms of $\operatorname{Tr}(W^* - YX)$. Return to (293) and give the following assumption:

Remark D.6 ($\operatorname{Tr}(W^* - YX)$ Nonzero during Training)**.** *We have that $\operatorname{Tr}(W^* - YX)$ is nonzero during training. That is, for all t prior to convergence, we have*

$$\operatorname{Tr}\Big(W^* - Y(t)X(t)\Big) \neq 0 \quad (304)$$

Note from the partial gradients of g in (269–270) that $\operatorname{Tr}(W^* - XY) = 0$ only at critical points of g. Therefore, (304) must hold during training, or, in other words, wherever

$$\left\|\frac{dY(t)}{dt}\right\| + \left\|\frac{dX(t)}{dt}\right\| \neq 0 \quad (305)$$

With this remark in mind, we can divide both sides of (293) by $-\operatorname{Tr}(W^* - YX)$ to find

$$\operatorname{Tr}\left(YY^T + X^TX\right) = -\frac{\frac{d}{dt}\operatorname{Tr}(W^* - YX)}{\operatorname{Tr}(W^* - YX)} \quad (306)$$

Finally, we substitute (303) and (306) into (295) to arrive at

$$\frac{d^2}{dt^2}\operatorname{Tr}(W^* - YX) = -\frac{d}{dt}\left[\operatorname{Tr}(W^* - YX)\right] \cdot -\frac{\frac{d}{dt}\operatorname{Tr}(W^* - YX)}{\operatorname{Tr}(W^* - YX)} \quad (307)$$

$$- \operatorname{Tr}(W^* - YX) \cdot 4\operatorname{Tr}(W^* - YX)\operatorname{Tr}(YX) \quad (308)$$

$$\frac{d^2}{dt^2}\operatorname{Tr}(W^* - YX) = \frac{\left(\frac{d}{dt}\operatorname{Tr}(W^* - YX)\right)^2}{\operatorname{Tr}(W^* - YX)} - 4\operatorname{Tr}^2(W^* - YX)\operatorname{Tr}(YX) \quad (309)$$

$$\frac{d^2}{dt^2}\operatorname{Tr}(W^* - YX) = \frac{\left(\frac{d}{dt}\operatorname{Tr}(W^* - YX)\right)^2}{\operatorname{Tr}(W^* - YX)} + 4\operatorname{Tr}^2(W^* - YX)\operatorname{Tr}(-YX) \tag{310}$$

$$\frac{d^2}{dt^2}\operatorname{Tr}(W^* - YX) = \frac{\left(\frac{d}{dt}\operatorname{Tr}(W^* - YX)\right)^2}{\operatorname{Tr}(W^* - YX)} + 4\operatorname{Tr}^2(W^* - YX)\Big(\operatorname{Tr}(W^* - YX) - \operatorname{Tr}(W^*)\Big) \tag{311}$$

$$\frac{d^2}{dt^2}\operatorname{Tr}(W^* - YX) = \frac{\left(\frac{d}{dt}\operatorname{Tr}(W^* - YX)\right)^2}{\operatorname{Tr}(W^* - YX)} + 4\operatorname{Tr}^3(W^* - YX) - 4\operatorname{Tr}(W^*)\operatorname{Tr}^2(W^* - YX) \tag{312}$$

Denote

$$a(t) := \operatorname{Tr}\Big(W^* - Y(t)X(t)\Big) \tag{313}$$

$$c := \operatorname{Tr}(W^*) \tag{314}$$

We then have the initial-value problem

$$a''(t) = \frac{\Big(a'(t)\Big)^2}{a(t)} + 4a^3(t) - 4ca^2(t) \tag{315}$$

$$a(0) = \operatorname{Tr}\Big(W^* - Y_0X_0\Big) = \operatorname{Tr}\left(W^*\right) \tag{316}$$

$$a'(0) = -\operatorname{Tr}\Big(W^* - Y_0X_0\Big)\operatorname{Tr}\left(Y_0Y_0^T + X_0^TX_0\right) \quad \Big(\text{from (293)}\Big) \tag{317}$$

$$= -\operatorname{Tr}\Big(W^*\Big)\left\|X_0\right\|^2 \tag{318}$$

In Appendix (F), we show that the closed-form solution to the IVP above is given on $[0, T]$ by

$$a(t) = \operatorname{sgn}(c)\frac{\left\|X_0\right\|^4 + 4c^2}{2\left\|X_0\right\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|} \tag{319}$$

where the constants κ_1 and κ_2 are given by

$$\kappa_1 = \left\|X_0\right\|^4 + 4c^2 \tag{320}$$

$$\kappa_2 = \frac{1}{\sqrt{\left\|X_0\right\|^4 + 4c^2}}\operatorname{arsinh}\left(\frac{\left\|X_0\right\|^2}{2|c|}\right) \tag{321}$$

Remark D.7 (Nonzero Denominator)**.** *Note that the denominator in (319) is strictly positive at initialization and increases monotonically in t. Consequently, $a(t)$ is well-defined for all $t \in [0, T]$.*

Returning to (280–281), our ODEs decouple to give

$$\frac{dY(t)}{dt} = a(t)X^T \tag{322}$$

$$\frac{dX(t)}{dt} = a(t)Y^T \tag{323}$$

Apply the $\dfrac{d}{dt}$ operator to both sides of (322). We have

$$\frac{d^2Y(t)}{dt^2} = a'(t)X^T + a(t)\frac{dX^T}{dt} \tag{324}$$

Take the transpose of (323) and plug it into (324) above to find

$$\frac{d^2Y(t)}{dt^2} = a'(t)X^T + a(t)\frac{dX^T}{dt} \tag{325}$$

$$\frac{d^2Y(t)}{dt^2} = a'(t)X^T + a^2(t)Y \tag{326}$$

Dividing both sides of (322) by $a(t)$ yields

$$X^T = \frac{1}{a(t)}\frac{dY}{dt} \tag{327}$$

Then (326) becomes

$$\frac{d^2Y(t)}{dt^2} = \frac{a'(t)}{a(t)}\frac{dY}{dt} + a^2(t)Y \tag{328}$$

We are thus left with the linear, decoupled ODE for $Y(t)$:

$$\frac{d^2Y}{dt^2} - \frac{a'(t)}{a(t)}\frac{dY}{dt} - a^2(t)Y = 0 \tag{329}$$

where $a(t)$ is given in (319). Taking a derivative of $a(t)$ yields

$$a'(t) = \left(\|X_0\|^4 + 4c^2\right) \cdot -\operatorname{sgn}(c)\frac{2\|X_0\|^2\sqrt{\kappa_1}\cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)}{\left(2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|\right)^2} \tag{330}$$

$$= -\operatorname{sgn}(c)\frac{2\|X_0\|^2\left(\|X_0\|^4 + 4c^2\right)^{3/2}\cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)}{\left(2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|\right)^2} \tag{331}$$

Then we can calculate

$$\frac{a'(t)}{a(t)} = -\operatorname{sgn}(c)\frac{2\|X_0\|^2\left(\|X_0\|^4 + 4c^2\right)^{3/2}\cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)}{\left(2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|\right)^2} \tag{332}$$

$$\cdot \operatorname{sgn}(c)\frac{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}{\|X_0\|^4 + 4c^2} \tag{333}$$

$$= -\frac{\|X_0\|^2\left(\|X_0\|^4 + 4c^2\right)^{1/2}\cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)}{\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 2|c|} \tag{334}$$

$$= -\frac{\|X_0\|^2\sqrt{\kappa_1}\cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)}{\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 2|c|} \tag{335}$$

We also have

$$a^2(t) = \left(\frac{\|X_0\|^4 + 4c^2}{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}\right)^2 \tag{336}$$

$$= \frac{\kappa_1^2}{\left(2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|\right)^2} \tag{337}$$

Our next task is to solve the following ODE for $Y(t)$:

$$\frac{d^2Y(t)}{dt^2} - \frac{a'(t)}{a(t)}\frac{dY(t)}{dt} - a^2(t)Y(t) = 0. \tag{338}$$

subject to the initial conditions

$$Y(0) = Y_0, \qquad \frac{dY(t)}{dt}\Big|_{t=0} = \operatorname{Tr}(W^* - Y_0X_0)X_0^T. \tag{339}$$

Although we adopt the initialization $Y_0 = \mathbf{0}_{n\times r}$, we keep Y_0 general in the calculation below to obtain a unified expression for the corresponding solution $X(t)$.

Let

$$s(t) = 2\left\|X_0\right\|^2 \sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c| \tag{340}$$

$$s'(t) = 2\sqrt{\kappa_1}\left\|X_0\right\|^2 \cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) \tag{341}$$

$$s''(t) = 2\kappa_1\left\|X_0\right\|^2 \sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) \tag{342}$$

Then (338) becomes

$$\frac{d^2Y}{dt^2} + \frac{s'(t)}{s(t)}\frac{dY}{dt} - \frac{\kappa_1^2}{s^2(t)}Y = 0 \tag{343}$$

Multiply both sides of the equation above by $s(t)$ to find

$$s(t)\frac{d^2Y}{dt^2} + s'(t)\frac{dY}{dt} - \frac{\kappa_1^2}{s(t)}Y = 0 \tag{344}$$

Note that $s(t) > 0$ on $[0, T]$ and make the substitution

$$Y(t) = \frac{U(t)}{\sqrt{s(t)}} = s^{-1/2}(t)U(t) \tag{345}$$

$$Y'(t) = -\frac{1}{2}s^{-3/2}(t)s'(t)U(t) + s^{-1/2}(t)U'(t) \tag{346}$$

$$\begin{aligned} Y''(t) = {} & \frac{3}{4}s^{-5/2}(t)s'^2(t)U(t) - \frac{1}{2}s^{-3/2}(t)s''(t)U(t) - \frac{1}{2}s^{-3/2}(t)s'(t)U'(t) \\ & - \frac{1}{2}s^{-3/2}(t)s'(t)U'(t) + s^{-1/2}(t)U''(t) \end{aligned} \tag{347}$$

Then we have

$$\begin{aligned} s(t)Y''(t) = {} & \frac{3}{4}s^{-3/2}(t)s'^2(t)U(t) - \frac{1}{2}s^{-1/2}(t)s''(t)U(t) - \frac{1}{2}s^{-1/2}(t)s'(t)U'(t) \\ & - \frac{1}{2}s^{-1/2}(t)s'(t)U'(t) + s^{1/2}(t)U''(t) \end{aligned} \tag{348}$$

as well as

$$s'(t)Y'(t) = -\frac{1}{2}s^{-3/2}(t)s'^2(t)U(t) + s^{-1/2}(t)s'(t)U'(t) \tag{349}$$

and

$$\frac{\kappa_1^2}{s(t)}Y = \kappa_1^2 s^{-3/2}(t)U(t) \tag{350}$$

Then (344) becomes

$$\begin{aligned} & \frac{3}{4}\,s^{-3/2}(t)\,s'^2(t)\,U(t) - \frac{1}{2}\,s^{-1/2}(t)\,s''(t)\,U(t) - \frac{1}{2}\,s^{-1/2}(t)\,s'(t)\,U'(t) - \frac{1}{2}\,s^{-1/2}(t)\,s'(t)\,U'(t) \\ & \quad + s^{1/2}(t)\,U''(t) - \frac{1}{2}\,s^{-3/2}(t)\,s'^2(t)\,U(t) + s^{-1/2}(t)\,s'(t)\,U'(t) - \kappa_1^2\,s^{-3/2}(t)\,U(t) = 0. \end{aligned} \tag{351}$$

or, after combining like terms,

$$\frac{1}{4}s^{-3/2}(t)\,s'^2(t)\,U(t) - \frac{1}{2}\,s^{-1/2}(t)\,s''(t)\,U(t) + s^{1/2}(t)\,U''(t) - \kappa_1^2\,s^{-3/2}(t)\,U(t) = 0 \tag{352}$$

Multiply both sides by $s^{-1/2}(t)$ and collect all the $U(t)$ terms to arrive at

$$U''(t) + \left(\frac{1}{4}s^{-2}(t)\,s'^2(t) - \frac{1}{2}\,s^{-1}(t)\,s''(t) - \kappa_1^2\,s^{-2}(t)\right)U(t) = 0 \tag{353}$$

$$U''(t) + \left(\frac{s'^2(t) - 2s(t)s''(t) - 4\kappa_1^2}{4s^2(t)}\right)U(t) = 0 \tag{354}$$

The coefficient on our $U(t)$ term simplifies. For brevity, write $\omega := \sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)$ and $\gamma := \cosh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)$. Then

$$\begin{aligned}
\frac{s'(t)^2 - 2s(t)s''(t) - 4\kappa_1^2}{4s(t)^2} &= \frac{4\kappa_1\|X_0\|^4\gamma^2 - 4\kappa_1\|X_0\|^2\omega\big(2\|X_0\|^2\omega + 4|c|\big) - 4\kappa_1^2}{4\big(2\|X_0\|^2\omega + 4|c|\big)^2} \\
&= \frac{\kappa_1}{4}\cdot\frac{4\|X_0\|^4\gamma^2 - 8\|X_0\|^4\omega^2 - 16|c|\|X_0\|^2\omega - 4\kappa_1}{4\|X_0\|^4\omega^2 + 16|c|\|X_0\|^2\omega + 16c^2} \\
&= \frac{\kappa_1}{4}\cdot\frac{4\|X_0\|^4(\gamma^2-\omega^2) - 4\|X_0\|^4\omega^2 - 16|c|\|X_0\|^2\omega - 4\kappa_1}{4\|X_0\|^4\omega^2 + 16|c|\|X_0\|^2\omega + 16c^2} \\
&= \frac{\kappa_1}{4}\cdot\frac{4\|X_0\|^4 - 4\|X_0\|^4\omega^2 - 16|c|\|X_0\|^2\omega - 4\kappa_1}{4\|X_0\|^4\omega^2 + 16|c|\|X_0\|^2\omega + 16c^2} \\
&= \frac{\kappa_1}{4}\cdot\frac{\|X_0\|^4 - \|X_0\|^4\omega^2 - 4|c|\|X_0\|^2\omega - \|X_0\|^4 - 4c^2}{\|X_0\|^4\omega^2 + 4|c|\|X_0\|^2\omega + 4c^2} \\
&= \frac{\kappa_1}{4}\cdot\frac{-4c^2 - \|X_0\|^4\omega^2 - 4|c|\|X_0\|^2\omega}{4c^2 + \|X_0\|^4\omega^2 + 4|c|\|X_0\|^2\omega} \\
&= -\frac{\kappa_1}{4}.
\end{aligned}$$

Then (354) becomes

$$U''(t) - \frac{\kappa_1}{4}U(t) = 0 \tag{355}$$

which is solved by

$$U(t) = A\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right) + B\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right) \tag{356}$$

We then have

$$Y(t)\sqrt{s(t)} = A\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right) + B\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right) \tag{357}$$

$$Y(t) = \frac{A\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right) + B\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{s(t)}} \tag{358}$$

$$Y(t) = \frac{A\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right) + B\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}} \tag{359}$$

Differentiating (359) with respect to t using the product and chain rules yields

$$Y'(t) = \frac{\sqrt{\kappa_1}\left[\|X_0\|^2\big(A\omega_1 - B\gamma_1\big) + 2|c|\big(B\omega_2 + A\gamma_2\big)\right]}{2^{3/2}\big(\|X_0\|^2\omega_3 + 2|c|\big)^{3/2}} \tag{360}$$

where $\omega_1 = \sinh\big(\frac{\sqrt{\kappa_1}(t+2\kappa_2)}{2}\big)$, $\gamma_1 = \cosh\big(\frac{\sqrt{\kappa_1}(t+2\kappa_2)}{2}\big)$, $\omega_2 = \sinh\big(\frac{\sqrt{\kappa_1}\,t}{2}\big)$, $\gamma_2 = \cosh\big(\frac{\sqrt{\kappa_1}\,t}{2}\big)$, and $\omega_3 = \sinh\big(\sqrt{\kappa_1}(t+\kappa_2)\big)$. Now use our initial conditions to solve for A and B. We have

$$Y(0) = \frac{B}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}\kappa_2\right) + 4|c|}} \tag{361}$$

$$= \frac{B}{\sqrt{2\|X_0\|^2\cdot\frac{\|X_0\|^2}{2|c|} + 4|c|}} \tag{362}$$

$$= \frac{B\sqrt{|c|}}{\sqrt{\|X_0\|^4 + 4c^2}} \tag{363}$$

$$= Y_0 \tag{364}$$

and

$$Y'(0) = \frac{\sqrt{\kappa_1}\left[A\left\|X_0\right\|^2 \sinh\left(\kappa_2\sqrt{\kappa_1}\right) - B\left\|X_0\right\|^2 \cosh\left(\kappa_2\sqrt{\kappa_1}\right) + 2A\left|c\right|\right]}{2^{\frac{3}{2}}\left(\left\|X_0\right\|^2 \sinh\left(\kappa_2\sqrt{\kappa_1}\right) + 2\left|c\right|\right)^{\frac{3}{2}}} \tag{365}$$

$$= \frac{\sqrt{\kappa_1}\left[A\left\|X_0\right\|^2 \cdot \frac{\left\|X_0\right\|^2}{2|c|} - B\left\|X_0\right\|^2 \cdot \frac{\sqrt{\left\|X_0\right\|^4+4c^2}}{2|c|} + 2A\left|c\right|\right]}{2^{\frac{3}{2}}\left(\left\|X_0\right\|^2 \cdot \frac{\left\|X_0\right\|^2}{2|c|} + 2\left|c\right|\right)^{\frac{3}{2}}} \tag{366}$$

$$= \frac{\sqrt{\kappa_1}\left[A\left\|X_0\right\|^4 - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2} + 4A\left|c\right|^2\right]}{2|c|\left(\frac{\left\|X_0\right\|^4}{|c|} + 4\left|c\right|\right)^{\frac{3}{2}}} \tag{367}$$

$$= \frac{\sqrt{\kappa_1|c|}\left[A\left\|X_0\right\|^4 - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2} + 4A\left|c\right|^2\right]}{2|c|^{3/2}\left(\frac{\left\|X_0\right\|^4}{|c|} + 4\left|c\right|\right)^{\frac{3}{2}}} \tag{368}$$

$$= \frac{\sqrt{\kappa_1|c|}\left[A\left\|X_0\right\|^4 - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2} + 4A\left|c\right|^2\right]}{2\left(\left\|X_0\right\|^4 + 4c^2\right)^{\frac{3}{2}}} \tag{369}$$

$$= \frac{\sqrt{|c|}\sqrt{\left\|X_0\right\|^4+4c^2}\left[A\left(\left\|X_0\right\|^4 + 4c^2\right) - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2}\right]}{2\left(\left\|X_0\right\|^4 + 4c^2\right)^{\frac{3}{2}}} \tag{370}$$

$$= \frac{\sqrt{|c|}\left[A\left(\left\|X_0\right\|^4 + 4c^2\right) - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2}\right]}{2\left(\left\|X_0\right\|^4 + 4c^2\right)} \tag{371}$$

$$= \mathrm{Tr}(W^* - Y_0X_0)X_0^T \tag{372}$$

So to determine A and B, we solve

$$Y_0 = \frac{B\sqrt{|c|}}{\sqrt{\left\|X_0\right\|^4+4c^2}} \tag{373}$$

$$\mathrm{Tr}(W^* - Y_0X_0)X_0^T = \frac{\sqrt{|c|}\left[A\left(\left\|X_0\right\|^4 + 4c^2\right) - B\left\|X_0\right\|^2 \sqrt{\left\|X_0\right\|^4+4c^2}\right]}{2\left(\left\|X_0\right\|^4 + 4c^2\right)} \tag{374}$$

which gives

$$A = \frac{\left\|X_0\right\|^2}{\sqrt{|c|}}Y_0 + \frac{2\,\mathrm{Tr}(W^* - Y_0X_0)X_0^T}{\sqrt{|c|}} \tag{375}$$

$$B = \frac{\sqrt{\left\|X_0\right\|^4+4c^2}}{\sqrt{|c|}}Y_0 \tag{376}$$

We thus finally arrive at our particular solution for $Y(t)$:

$$Y(t) = \left(\frac{\left\|X_0\right\|^2}{\sqrt{|c|}}Y_0 + \frac{2\,\mathrm{Tr}(W^* - Y_0X_0)X_0^T}{\sqrt{|c|}}\right)\frac{\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\left\|X_0\right\|^2 \sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}}$$

$$+\frac{\sqrt{\|X_0\|^4+4c^2}}{\sqrt{|c|}}\cdot\frac{\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)+4|c|}}Y_0 \tag{377}$$

Or, gathering the Y_0 and $\mathrm{Tr}(W^*-Y_0X_0)X_0^T$ terms, we have

$$\begin{aligned}Y(t)=&\frac{1}{\sqrt{2\|X_0\|^2\omega_3+4|c|}}\left[\frac{\|X_0\|^2}{\sqrt{|c|}}\omega_2+\frac{\sqrt{\|X_0\|^4+4c^2}}{\sqrt{|c|}}\gamma_2\right]Y_0\\&+\frac{1}{\sqrt{2\|X_0\|^2\omega_3+4|c|}}\cdot\frac{2\omega_2}{\sqrt{|c|}}\,\mathrm{Tr}(W^*-Y_0X_0)X_0^T\end{aligned} \tag{378}$$

where, as before, $\omega_2=\sinh(\frac{\sqrt{\kappa_1}}{2}t)$, $\gamma_2=\cosh(\frac{\sqrt{\kappa_1}}{2}t)$, and $\omega_3=\sinh(\sqrt{\kappa_1}(t+\kappa_2))$.

Now define

$$p(t):=\frac{1}{\sqrt{2\|X_0\|^2\omega_3+4|c|}}\left[\frac{\|X_0\|^2}{\sqrt{|c|}}\omega_2+\frac{\sqrt{\|X_0\|^4+4c^2}}{\sqrt{|c|}}\gamma_2\right] \tag{379}$$

$$q(t):=\frac{1}{\sqrt{2\|X_0\|^2\omega_3+4|c|}}\cdot\frac{2\omega_2}{\sqrt{|c|}} \tag{380}$$

We then have the closed-form expression for $Y(t)$:

$$Y(t)=p(t)Y_0+q(t)\,\mathrm{Tr}(W^*-Y_0X_0)\,X_0^T \tag{381}$$

An identical calculation for $X(t)$, noting the symmetry of (322–323), yields

$$X(t)=p(t)X_0+q(t)\,\mathrm{Tr}(W^*-Y_0X_0)\,Y_0^T \tag{382}$$

Using general $Y_0\in\mathbb{R}^{n\times r}$ enabled us to take advantage of the symmetry in (322–323) to immediately find a closed-form expression for $X(t)$. Now that we have solutions for both $X(t)$ and $Y(t)$, we can enforce $Y_0=\mathbf{0}_{n\times r}$ to arrive at

$$Y(t)=q(t)\,\mathrm{Tr}(W^*)\,X_0^T \tag{383}$$

$$X(t)=p(t)X_0 \tag{384}$$

which solve (280–281) for our specified initial conditions. Thus, the rank-r minimizer for (268) at time $t\in[0,T]$ is given by

$$Y(t)X(t)=p(t)q(t)\,\mathrm{Tr}\,(W^*)\,X_0^TX_0 \tag{385}$$

It remains to investigate how the rank-r minimizer in (385) behaves for large training times. Since our closed-form expressions for $X(t)$ and $Y(t)$ solve (280–281) on $[0,T]$ for any $T>0$, they define a single solution to the same ODE system for all $t\in[0,\infty)$. Thus, we investigate the convergence of the rank-r minimizer by examining $Y(t)X(t)$ in the limit that $t\to\infty$. Recall that, as $t\to\infty$,

$$\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right)\sim\frac{1}{2}e^{\frac{\sqrt{\kappa_1}}{2}t}, \tag{386}$$

$$\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)\sim\frac{1}{2}e^{\frac{\sqrt{\kappa_1}}{2}t}, \tag{387}$$

and

$$\sinh(\sqrt{\kappa_1}(t+\kappa_2))\sim\frac{1}{2}e^{\sqrt{\kappa_1}(t+\kappa_2)} \tag{388}$$

This gives us

$$\lim_{t\to\infty}\frac{\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)+4|c|}}=\lim_{t\to\infty}\frac{\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right)+4|c|}} \tag{389}$$

$$= \lim_{t\to\infty} \frac{\frac{1}{2}e^{\frac{\sqrt{\kappa_1}}{2}t}}{\sqrt{2\|X_0\|^2 \cdot \frac{e^{\sqrt{\kappa_1}(t+\kappa_2)}}{2} + 4|c|}} \tag{390}$$

$$= \lim_{t\to\infty} \frac{e^{\frac{\sqrt{\kappa_1}}{2}t}}{2\sqrt{\|X_0\|^2 \cdot e^{\sqrt{\kappa_1}(t+\kappa_2)} + 4|c|}} \cdot \frac{e^{-\frac{\sqrt{\kappa_1}}{2}t}}{e^{-\frac{\sqrt{\kappa_1}}{2}t}} \tag{391}$$

$$= \lim_{t\to\infty} \frac{1}{2\sqrt{\|X_0\|^2 \cdot e^{\kappa_2\sqrt{\kappa_1}} + 4|c|e^{-\sqrt{\kappa_1}t}}} \tag{392}$$

$$= \frac{1}{2\,\|X_0\|\, e^{\frac{\kappa_2\sqrt{\kappa_1}}{2}}} \tag{393}$$

Note that

$$e^{\kappa_2\sqrt{\kappa_1}} = e^{\operatorname{arsinh}\left(\frac{\|X_0\|^2}{2|c|}\right)} \tag{394}$$

$$= e^{\ln\left(\frac{\|X_0\|^2}{2|c|} + \sqrt{1+\left(\frac{\|X_0\|^2}{2|c|}\right)^2}\right)} \tag{395}$$

$$= \frac{\|X_0\|^2}{2|c|} + \sqrt{1 + \left(\frac{\|X_0\|^2}{2|c|}\right)^2} \tag{396}$$

$$= \frac{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}{2|c|} \tag{397}$$

So (393) becomes

$$\frac{1}{2\,\|X_0\|} \cdot \sqrt{\frac{2|c|}{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} = \frac{\sqrt{|c|}}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} \tag{398}$$

So we have

$$\lim_{t\to\infty} \frac{\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}} = \lim_{t\to\infty} \frac{\cosh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}} \tag{399}$$

$$= \frac{\sqrt{|c|}}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} \tag{400}$$

We can then quickly calculate

$$\lim_{t\to\infty} p(t) = \lim_{t\to\infty} \frac{1}{\sqrt{2\|X_0\|^2\omega_3 + 4|c|}} \left[\frac{\|X_0\|^2}{\sqrt{|c|}}\omega_2 + \frac{\Lambda}{\sqrt{|c|}}\gamma_2\right] \tag{401}$$

$$= \lim_{t\to\infty} \frac{\|X_0\|^2}{\sqrt{|c|}} \cdot \frac{\omega_2}{\sqrt{2\|X_0\|^2\omega_3 + 4|c|}} + \frac{\Lambda}{\sqrt{|c|}} \cdot \frac{\gamma_2}{\sqrt{2\|X_0\|^2\omega_3 + 4|c|}} \tag{402}$$

$$= \frac{\|X_0\|^2}{\sqrt{|c|}} \cdot \frac{\sqrt{|c|}}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \Lambda}} + \frac{\Lambda}{\sqrt{|c|}} \cdot \frac{\sqrt{|c|}}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \Lambda}} \tag{403}$$

$$= \frac{\|X_0\|^2 + \Lambda}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \Lambda}} \tag{404}$$

$$= \frac{\sqrt{\|X_0\|^2 + \Lambda}}{\sqrt{2}\,\|X_0\|} \tag{405}$$

where $\Lambda := \sqrt{\|X_0\|^4 + 4c^2}$. We also have

$$\lim_{t\to\infty} q(t) = \lim_{t\to\infty} \frac{2}{\sqrt{|c|}} \cdot \frac{\sinh\left(\frac{\sqrt{\kappa_1}}{2}t\right)}{\sqrt{2\|X_0\|^2\sinh\left(\sqrt{\kappa_1}(t+\kappa_2)\right) + 4|c|}} \tag{406}$$

$$= \frac{2}{\sqrt{|c|}} \cdot \frac{\sqrt{|c|}}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} \tag{407}$$

$$= \frac{2}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} \tag{408}$$

Now calculate the post-training rank-r minimizer for (268), which converges to

$$\lim_{t\to\infty} Y(t)X(t) = \lim_{t\to\infty} p(t)q(t)\,\mathrm{Tr}\,(W^*)\,X_0^T X_0 \tag{409}$$

$$= \frac{\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}}{\sqrt{2}\,\|X_0\|} \cdot \frac{2}{\sqrt{2}\,\|X_0\|\sqrt{\|X_0\|^2 + \sqrt{\|X_0\|^4 + 4c^2}}} \cdot \mathrm{Tr}(W^*)X_0^T X_0 \tag{410}$$

$$= \frac{\mathrm{Tr}(W^*)}{\|X_0\|^2} X_0^T X_0 \tag{411}$$

For a quick sanity check, calculate the trace of our final rank-r minimizer:

$$\mathrm{Tr}\left(\frac{\mathrm{Tr}(W^*)}{\|X_0\|^2} X_0^T X_0\right) = \frac{\mathrm{Tr}(W^*)}{\|X_0\|^2}\,\mathrm{Tr}\left(X_0^T X_0\right) = \frac{\mathrm{Tr}(W^*)}{\|X_0\|^2}\,\|X_0\|^2 = \mathrm{Tr}(W^*) \tag{412}$$

which gives $\lim_{t\to\infty} \frac{1}{2}\,\mathrm{Tr}^2\left(W^* - Y(t)X(t)\right) = 0$ as expected.

E PROOF OF LEMMA D.3

Lemma E.1 (Lipschitz Smoothness for Trace Squared Objective)**.** *Define* $g : \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times n} \to \mathbb{R}$ *via*

$$g(B, A) := \frac{1}{2}\operatorname{Tr}^2(W^* - BA) \tag{413}$$

for constant $W^* \in \mathbb{R}^{n\times n}$*. Then the objective gradient* ∇g *is Lipschitz smooth in our training domain. Namely, there exists* $L_{R'} > 0$ *such that, for any* $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$*, we have*

$$\|\nabla g(B_1, A_1) - \nabla g(B_2, A_2)\| \le L_{R'} \left\| \Big(B_1, A_1\Big) - \Big(B_2, A_2\Big) \right\| \tag{414}$$

Proof:

Let $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$ *be given. Refer to the partial gradients for* $g(B, A)$ *in (269–270) of Appendix D. We have*

$$\|\nabla g(B_1, A_1) - \nabla g(B_2, A_2)\| = \left\| \big(\Delta_B \,,\, \Delta_A\big) \right\| \tag{415}$$

where $\Delta_B := \nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2)$ *and* $\Delta_A := \nabla_A g(B_1, A_1) - \nabla_A g(B_2, A_2)$*.*

The partial gradients with respect to B *simplify to*

$$\nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2) = -\operatorname{Tr}(W^* - B_1A_1)A_1^T + \operatorname{Tr}(W^* - B_2A_2)A_2^T \tag{416}$$

$$= \operatorname{Tr}(W^* - B_2A_2)A_2^T - \operatorname{Tr}(W^* - B_1A_1)A_2^T + \operatorname{Tr}(W^* - B_1A_1)A_2^T - \operatorname{Tr}(W^* - B_1A_1)A_1^T \tag{417}$$

$$= \Big[\operatorname{Tr}(W^* - B_2A_2) - \operatorname{Tr}(W^* - B_1A_1)\Big]A_2^T + \operatorname{Tr}(W^* - B_1A_1)\left(A_2^T - A_1^T\right) \tag{418}$$

$$= \Big[\operatorname{Tr}(B_1A_1 - B_2A_2)\Big]A_2^T + \operatorname{Tr}(W^* - B_1A_1)\left(A_2^T - A_1^T\right) \tag{419}$$

Recall that $|\operatorname{Tr}(BA)| \le \|B\|\,\|A\|$ *(Cauchy-Schwarz for the Frobenius inner product). Taking the norm, we find*

$$\|\nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2)\| = \left\| \Big[\operatorname{Tr}(B_1A_1 - B_2A_2)\Big]A_2^T + \operatorname{Tr}(W^* - B_1A_1)\left(A_2^T - A_1^T\right) \right\| \tag{420}$$

$$\le \left\| \Big[\operatorname{Tr}(B_1A_1 - B_2A_2)\Big]A_2^T \right\| + \left\| \operatorname{Tr}(W^* - B_1A_1)\left(A_2^T - A_1^T\right) \right\| \tag{421}$$

$$= |\operatorname{Tr}(B_1A_1 - B_2A_2)|\,\left\|A_2^T\right\| + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{422}$$

$$= |\operatorname{Tr}(B_1A_1 - B_1A_2 + B_1A_2 - B_2A_2)|\,\left\|A_2^T\right\| + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{423}$$

$$= \left|\operatorname{Tr}\Big(B_1(A_1 - A_2) + (B_1 - B_2)A_2\Big)\right|\,\left\|A_2^T\right\| + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{424}$$

$$= \left|\operatorname{Tr}\Big(B_1(A_1 - A_2)\Big) + \operatorname{Tr}\Big((B_1 - B_2)A_2\Big)\right|\,\left\|A_2^T\right\| + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{425}$$

$$\le \Big(\|B_1\|\,\|A_1 - A_2\| + \|B_1 - B_2\|\,\|A_2\|\Big)\left\|A_2^T\right\| + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{426}$$

$$\le \Big(R'\,\|A_1 - A_2\| + \|B_1 - B_2\|\,R'\Big)R' + |\operatorname{Tr}(W^* - B_1A_1)|\,\left\|A_2^T - A_1^T\right\| \tag{427}$$

$$\le \Big(\|A_1 - A_2\| + \|B_1 - B_2\|\Big)R'^2 + \Big(|\operatorname{Tr}(W^*)| + |\operatorname{Tr}(B_1A_1)|\Big)\left\|A_2^T - A_1^T\right\| \tag{428}$$

$$\le \Big(\|A_1 - A_2\| + \|B_1 - B_2\|\Big)R'^2 + \Big(|\operatorname{Tr}(W^*)| + \|B_1\|\,\|A_1\|\Big)\left\|A_2^T - A_1^T\right\| \tag{429}$$

$$\le \Big(\|A_1 - A_2\| + \|B_1 - B_2\|\Big)R'^2 + \Big(|\operatorname{Tr}(W^*)| + R'^2\Big)\left\|A_2^T - A_1^T\right\| \tag{430}$$

$$= \|B_1 - B_2\|\,R'^2 + \Big(|\operatorname{Tr}(W^*)| + 2R'^2\Big)\,\|A_2 - A_1\| \tag{431}$$

$$\leq \|B_1 - B_2\| \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big) + \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big) \|A_2 - A_1\| \tag{432}$$

$$= \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big) \Big(\|B_1 - B_2\| + \|A_1 - A_2\| \Big) \tag{433}$$

$$\leq \sqrt{2} \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big) \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big)^{1/2} \tag{434}$$

where the last line holds by Cauchy-Schwartz on $\mathbb{R}^2$*. So we have*

$$\|\nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2)\| \leq \sqrt{2} \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big) \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big)^{1/2} \tag{435}$$

Denote $\frac{1}{\sqrt{2}} L_{R'} := \sqrt{2} \Big(|\operatorname{Tr}(W^*)| + 2R'^2 \Big)$*. Squaring* (435)*, we get*

$$\|\nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2)\|^2 \leq \frac{1}{2} L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{436}$$

An identical calculation for $\|\nabla_A g(B_1, A_1) - \nabla_A g(B_2, A_2)\|$ *gives*

$$\|\nabla_A g(B_1, A_1) - \nabla_A g(B_2, A_2)\|^2 \leq \frac{1}{2} L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{437}$$

Together, (436–437) give

$$\|\nabla g(B_1, A_1) - \nabla g(B_2, A_2)\|^2 = \|\nabla_B g(B_1, A_1) - \nabla_B g(B_2, A_2)\|^2 + \|\nabla_A g(B_1, A_1) - \nabla_A g(B_2, A_2)\|^2 \tag{438}$$

$$\leq L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{439}$$

$$= L_{R'}^2 \Big\| \Big(B_1, A_1 \Big) - \Big(B_2, A_2 \Big) \Big\|^2 \tag{440}$$

or simply

$$\|\nabla g(B_1, A_1) - \nabla g(B_2, A_2)\| \leq L_{R'} \Big\| \Big(B_1, A_1 \Big) - \Big(B_2, A_2 \Big) \Big\| \tag{441}$$

for all $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$ *Thus,* g *is* $L_{R'}$ *Lipschitz smooth on* $\mathcal{D}_{R'}$*.*

F Solution to the Initial-Value Problem in (315–318)

We wish to solve the initial-value problem

$$a''(t) = \frac{\Big(a'(t)\Big)^2}{a(t)} + 4a^3(t) - 4ca^2(t) \tag{442}$$

$$a(0) = \operatorname{Tr}\left(W^*\right) := c \tag{443}$$

$$a'(0) = -\operatorname{Tr}\left(W^*\right)\|X_0\|^2 := -c\,\|X_0\|^2 \tag{444}$$

for $t \in [0, T]$, where $T > 0$ is given in Appendix D.

Begin by letting $u(t) = a'(t)$. This gives $u'(t) = a''(t)$. By the Chain Rule, we find

$$a''(t) = \frac{d^2a}{dt^2} = \frac{du}{da}\frac{da}{dt} = \frac{du}{da} \cdot u \tag{445}$$

Substituting for $a'(t)$ and $a''(t)$, (442) becomes

$$u\frac{du}{da} = \frac{u^2}{a} + 4a^3 - 4ca^2 \tag{446}$$

Now let $y = u^2$. Differentiate both sides with respect to a. By the Chain Rule, we have

$$\frac{dy}{da} = 2u\frac{du}{da} \tag{447}$$

$$\frac{1}{2}\frac{dy}{da} = u\frac{du}{da} \tag{448}$$

Then (446) becomes

$$\frac{1}{2}\frac{dy}{da} = \frac{y}{a} + 4a^3 - 4ca^2 \tag{449}$$

$$\frac{dy}{da} = \frac{2y}{a} + 8a^3 - 8ca^2 \tag{450}$$

$$\frac{dy}{da} - \frac{2y}{a} = 8a^3 - 8ca^2 \tag{451}$$

By Remark (D.6) in Appendix D, we have that $a(t) \neq 0$ during training. Since we are concerned only with the trajectory of $a(t)$ prior to convergence, we can divide both sides of (451) by a^2 to get

$$\frac{1}{a^2}\frac{dy}{da} - \frac{2}{a^3}y = 8a - 8c \tag{452}$$

$$\frac{d}{da}\left[\frac{y}{a^2}\right] = 8a - 8c \tag{453}$$

$$\int \frac{d}{da}\left[\frac{y}{a^2}\right] da = \int 8a - 8c\,da \tag{454}$$

$$\frac{y}{a^2} = 4a^2 - 8ca + \kappa_1 \tag{455}$$

$$y = a^2(4a^2 - 8ca + \kappa_1) \tag{456}$$

$$u^2 = a^2(4a^2 - 8ca + \kappa_1) \tag{457}$$

$$u = \pm a\sqrt{4a^2 - 8ca + \kappa_1} \tag{458}$$

$$a' = \pm a\sqrt{4a^2 - 8ca + \kappa_1} \tag{459}$$

Note that the quantity on the right-hand side remains real:

Remark F.1 (Radicand is Non-negative)**.** *From (459), we have*

$$a' = \pm a\sqrt{4a^2 - 8ca + \kappa_1} \tag{460}$$

$$\pm\frac{a'}{a} = \sqrt{4a^2 - 8ca + \kappa_1} \tag{461}$$

$$\left(\frac{a'}{a}\right)^2 = 4a^2 - 8ca + \kappa_1 \tag{462}$$

Thus, $4a^2 - 8ca + \kappa_1 \geq 0$.

From (293) in Appendix D, we have

$$a'(t) = -a(t)\left(\|X\|^2 + \|Y\|^2\right) \tag{463}$$

Since $\left(\|X\|^2 + \|Y\|^2\right) \geq 0$, we have that $a'(t)$ and $a(t)$ are opposite signed. Thus, (459) becomes

$$a' = -a\sqrt{4a^2 - 8ca + \kappa_1} \tag{464}$$

We utilize the initial conditions at this point to solve for κ_1:

$$a'(0) = -a(0)\sqrt{4a^2(0) - 8ca(0) + \kappa_1} \tag{465}$$

$$-\operatorname{Tr}\left(W^*\right)\|X_0\|^2 = -\operatorname{Tr}\left(W^*\right)\sqrt{4\operatorname{Tr}^2\left(W^*\right) - 8c\operatorname{Tr}\left(W^*\right) + \kappa_1} \tag{466}$$

$$\|X_0\|^2 = \sqrt{4c^2 - 8c^2 + \kappa_1} \tag{467}$$

$$\|X_0\|^4 = \kappa_1 - 4c^2 \tag{468}$$

$$\kappa_1 = \|X_0\|^4 + 4c^2 \tag{469}$$

So we get $\kappa_1 = \|X_0\|^4 + 4c^2$. Note that $\kappa_1 > 0$.

Return to (464). We have

$$\frac{da}{dt} = -a\sqrt{4a^2 - 8ca + \kappa_1} \tag{470}$$

$$-\int \frac{da}{a\sqrt{4a^2 - 8ca + \kappa_1}} = \int dt \tag{471}$$

$$-\int \frac{da}{a\sqrt{4a^2 - 8ca + \kappa_1}} = t + \kappa_2 \tag{472}$$

Calculate the integral on left-hand side above. Begin by completing the square in the denominator:

$$\int \frac{da}{a\sqrt{4a^2 - 8ca + \kappa_1}} = \int \frac{da}{a\sqrt{4(a-c)^2 + \kappa_1 - 4c^2}} \tag{473}$$

$$= \frac{1}{2}\int \frac{da}{a\sqrt{(a-c)^2 + \frac{\kappa_1}{4} - c^2}} \tag{474}$$

$$= \frac{1}{2}\int \frac{-a\,da}{-a^2\sqrt{(a-c)^2 + \frac{\kappa_1}{4} - c^2}} \tag{475}$$

Making the substitution:

$$v = \frac{1}{a} \tag{476}$$

$$dv = -\frac{1}{a^2}da, \tag{477}$$

we find

$$\frac{1}{2}\int \frac{-a\,da}{-a^2\sqrt{(a-c)^2 + \frac{\kappa_1}{4} - c^2}} = -\frac{1}{2}\int \frac{dv}{v\sqrt{\left(\frac{1}{v} - c\right)^2 + \frac{\kappa_1}{4} - c^2}} \tag{478}$$

$$= -\frac{1}{2}\int \pm\frac{dv}{\sqrt{(1-cv)^2 + v^2\left(\frac{\kappa_1}{4} - c^2\right)}} \tag{479}$$

We insert the $\pm$ symbol into our integrand above to avoid presuming on the sign of v. To pull v into the square root in the denominator above, we have to write $v = \pm\sqrt{v^2}$ since we have

$$v = +\sqrt{v^2} \qquad v \geq 0 \tag{480}$$

$$v = -\sqrt{v^2} \qquad v < 0 \tag{481}$$

We can then continue our calculation:

$$-\frac{1}{2}\int \pm\frac{dv}{\sqrt{(1-cv)^2 + v^2\left(\frac{\kappa_1}{4} - c^2\right)}} = -\frac{1}{2}\int \pm\frac{dv}{\sqrt{1 - 2cv + c^2v^2 - c^2v^2 + \frac{v^2\kappa_1}{4}}} \tag{482}$$

$$= -\int \pm\frac{dv}{\sqrt{4 - 8cv + v^2\kappa_1}} \tag{483}$$

Complete the square in the denominator once again to find

$$-\int \pm\frac{dv}{\sqrt{4 - 8cv + v^2\kappa_1}} = -\int \pm\frac{dv}{\sqrt{\kappa_1\left(v - \frac{4c}{\kappa_1}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} \tag{484}$$

$$= -\int \pm\frac{dv}{\sqrt{\left(\sqrt{\kappa_1}v - \frac{4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} \tag{485}$$

$$= -\int \pm\frac{dv}{\sqrt{\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} \tag{486}$$

Now make the substitution

$$w = \frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}} \tag{487}$$

$$dw = \sqrt{\frac{\kappa_1}{4 - \frac{16c^2}{\kappa_1}}}dv \tag{488}$$

$$dv = \frac{\sqrt{4 - \frac{16c^2}{\kappa_1}}}{\sqrt{\kappa_1}}dw \tag{489}$$

Remark F.2 (Another Non-negative Radicand)**.** *Note that*

$$4 - \frac{16c^2}{\kappa_1} = 4\left(1 - \frac{4c^2}{\kappa_1}\right) = 4\left(\frac{\kappa_1 - 4c^2}{\kappa_1}\right) = \frac{4}{\kappa_1}(\kappa_1 - 4c^2) \tag{490}$$

Since $\kappa_1 > 0$ for our initialization scheme, the left factor is positive. On the right,

$$\kappa_1 - 4c^2 = \|X_0\|^4 + 4c^2 - 4c^2 = \|X_0\|^4 > 0 \tag{491}$$

Therefore, $4 - \frac{16c^2}{\kappa_1} > 0$, and w is real.

Changing variables to w and carefully tracking the sign of the integrand based on the sign of v , we then have

$$-\int \pm\frac{dv}{\sqrt{\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} = -\int \pm\frac{dw}{\sqrt{\left(w\sqrt{4 - \frac{16c^2}{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} \cdot \frac{\sqrt{4 - \frac{16c^2}{\kappa_1}}}{\sqrt{\kappa_1}} \tag{492}$$

$$= -\int \pm\frac{dw}{\sqrt{w^2\left(4 - \frac{16c^2}{\kappa_1}\right) + 4 - \frac{16c^2}{\kappa_1}}} \cdot \frac{\sqrt{4 - \frac{16c^2}{\kappa_1}}}{\sqrt{\kappa_1}} \tag{493}$$

$$= -\frac{1}{\sqrt{\kappa_1}} \int \pm \frac{dw}{\sqrt{w^2+1}} \tag{494}$$

$$\tag{495}$$

Thus, we have arrived at

$$-\int \pm \frac{dv}{\sqrt{\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} = \begin{cases} -\frac{1}{\sqrt{\kappa_1}} \int + \frac{dw}{\sqrt{w^2+1}} & v \geq 0 \\ -\frac{1}{\sqrt{\kappa_1}} \int - \frac{dw}{\sqrt{w^2+1}} & v < 0 \end{cases} \tag{496}$$

Taking the anti-derivatives on the right, this becomes

$$-\int \pm \frac{dv}{\sqrt{\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} = \begin{cases} -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}(w) & v \geq 0 \\ -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}(-w) & v < 0 \end{cases} \tag{497}$$

Back-substitution then yields

$$-\int \pm \frac{dv}{\sqrt{\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}}\right)^2 + 4 - \frac{16c^2}{\kappa_1}}} = \begin{cases} -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & v \geq 0 \\ -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(-\frac{\kappa_1 v - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & v < 0 \end{cases} \tag{498}$$

Note that since $v = \dfrac{1}{a}$, the signs of a and v will be the same. Recalling the original integral in (473), we finally have

$$\int \frac{da}{a\sqrt{4a^2 - 8ca + \kappa_1}} = \begin{cases} -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & a \geq 0 \\ -\frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(-\frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & a < 0 \end{cases} \tag{499}$$

Eq. (472) then yields

$$t + \kappa_2 = \begin{cases} \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & a \geq 0 \\ \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(-\frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) & a < 0 \end{cases} \tag{500}$$

or simply

$$t + \kappa_2 = \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(a) \frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) \tag{501}$$

Note that since $a(t)$ is continuous and nonzero and that $a(0) = c$, we will have $\operatorname{sgn}(a) = \operatorname{sgn}(c)$ on $[0, T]$. Thus we have

$$t + \kappa_2 = \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{a} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) \tag{502}$$

At this point, we solve for κ_2:

$$\kappa_2 = \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{a(0)} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) \tag{503}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{c} - 4c}{\sqrt{\kappa_1}\sqrt{4 - \frac{16c^2}{\kappa_1}}}\right) \tag{504}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{c} - 4c}{\sqrt{4\kappa_1 - 16c^2}}\right) \tag{505}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{c} - 4c}{\sqrt{4\left\|X_0\right\|^4 + 16c^2 - 16c^2}}\right) \tag{506}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\frac{\kappa_1}{c} - 4c}{2\left\|X_0\right\|^2}\right) \tag{507}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\kappa_1 - 4c^2}{2c\left\|X_0\right\|^2}\right) \tag{508}$$

$$= \frac{1}{\sqrt{\kappa_1}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\left\|X_0\right\|^4 + 4c^2 - 4c^2}{2c\left\|X_0\right\|^2}\right) \tag{509}$$

$$= \frac{1}{\sqrt{\left\|X_0\right\|^4 + 4c^2}} \operatorname{arsinh}\left(\operatorname{sgn}(c) \frac{\left\|X_0\right\|^2}{2c}\right) \tag{510}$$

$$\kappa_2 = \frac{1}{\sqrt{\left\|X_0\right\|^4 + 4c^2}} \operatorname{arsinh}\left(\frac{\left\|X_0\right\|^2}{2|c|}\right) \tag{511}$$

With κ_1 and κ_2 defined, we can find a closed-form solution for $a(t)$. Solve (502) for a to obtain

$$a(t) = \frac{\left\|X_0\right\|^4 + 4c^2}{2\operatorname{sgn}(c)\left\|X_0\right\|^2 \sinh\left(\sqrt{\kappa_1}(t + \kappa_2)\right) + 4c} \tag{512}$$

$$a(t) = \operatorname{sgn}(c) \frac{\left\|X_0\right\|^4 + 4c^2}{2\left\|X_0\right\|^2 \sinh\left(\sqrt{\kappa_1}(t + \kappa_2)\right) + 4|c|} \tag{513}$$

where

$$c := \operatorname{Tr}(W^*). \tag{514}$$

G Learning Dynamics for Trace-Squared Loss (Full-rank)

We wish to analyze the learning dynamics of the full-rank optimizer $\Delta W \in \mathbb{R}^{n\times n}$ produced by applying full-rank GD to the problem

$$\min_{\Delta W\in\mathbb{R}^{n\times n}} \frac{1}{2}\operatorname{Tr}^2(W^* - \Delta W) \tag{515}$$

The gradient for the loss function $f(\Delta W)$ above is given by

$$\nabla f(\Delta W) = -\operatorname{Tr}\left(W^* - \Delta W\right) I_n \tag{516}$$

where $I_n \in \mathbb{R}^{n\times n}$ denotes the identity matrix. We initialize $U(0) = Y_0X_0$ for consistency with the LoRA case. The gradient flow ODE describing the learning dynamics of the trace-squared objective in (515) under full-rank gradient descent is given by

$$\frac{dU(t)}{dt} = -\nabla f\big(U(t)\big) = \operatorname{Tr}\big(W^* - U(t)\big)\, I_n, \tag{517}$$

$$U(0) = Y_0X_0, \tag{518}$$

where the solution $U : \mathbb{R}_0^+ \to \mathbb{R}^{n\times n}$ denotes the continuous-time analogue of the full-rank GD iterates. Throughout the calculation below, we allow a general initialization $Y_0 \in \mathbb{R}^{n\times r}$ for completeness. However, to facilitate direct comparison with the low-rank learning dynamics for the same objective (see Appendix D), we specialize to the choice $Y_0 = \mathbf{0}_{n\times r}$ at the conclusion of our analysis.

Taking the trace of both sides of (517–518), we have

$$\frac{d\operatorname{Tr}\Big(U(t)\Big)}{dt} = \operatorname{Tr}\Big(W^* - U(t)\Big)\operatorname{Tr}\left(I_n\right) \tag{519}$$

$$\frac{d\operatorname{Tr}\Big(U(t)\Big)}{dt} = n\operatorname{Tr}\Big(W^* - U(t)\Big) \tag{520}$$

$$\frac{d\operatorname{Tr}\Big(U(t)\Big)}{dt} = n\operatorname{Tr}\left(W^*\right) - n\operatorname{Tr}\Big(U(t)\Big) \tag{521}$$

$$\operatorname{Tr}\Big(U(0)\Big) = \operatorname{Tr}(Y_0X_0) \tag{522}$$

which is solved by

$$\operatorname{Tr}\Big(U(t)\Big) = \operatorname{Tr}(W^*) - e^{-nt}\operatorname{Tr}\left(W^* - Y_0X_0\right) \tag{523}$$

Plugging our closed-form expression for $\operatorname{Tr}\Big(U(t)\Big)$ into (517), we arrive at

$$\frac{dU(t)}{dt} = \operatorname{Tr}\Big(W^* - U(t)\Big) I_n \tag{524}$$

$$\frac{dU(t)}{dt} = \Big[\operatorname{Tr}\left(W^*\right) - \operatorname{Tr}\Big(U(t)\Big)\Big]\, I_n \tag{525}$$

$$\frac{dU(t)}{dt} = \left[\operatorname{Tr}\left(W^*\right) - \operatorname{Tr}(W^*) + e^{-nt}\operatorname{Tr}\left(W^* - Y_0X_0\right)\right] I_n \tag{526}$$

$$\frac{dU(t)}{dt} = e^{-nt}\operatorname{Tr}\left(W^* - Y_0X_0\right) I_n \tag{527}$$

which, for our initial condition in (518), is solved by

$$U(t) = \frac{1 - e^{-nt}}{n}\operatorname{Tr}\left(W^* - Y_0X_0\right) I_n + Y_0X_0 \tag{528}$$

For a quick sanity check, note that

$$U(0) = Y_0X_0 \tag{529}$$

$$\lim_{t\to\infty} \operatorname{Tr}\Big(U(t)\Big) = \operatorname{Tr}\left(\frac{1}{n}\operatorname{Tr}\left(W^* - Y_0X_0\right) I_n + Y_0X_0\right) \tag{530}$$

$$= \frac{1}{n}\operatorname{Tr}\left(W^* - Y_0X_0\right)\operatorname{Tr}\left(I_n\right) + \operatorname{Tr}(Y_0X_0) \tag{531}$$

$$= \operatorname{Tr}\left(W^* - Y_0X_0\right) + \operatorname{Tr}(Y_0X_0) \tag{532}$$

$$= \operatorname{Tr}(W^*) \tag{533}$$

which gives the result we would expect for the minimization problem in (515).

Taking the limit of $U(t)$ as $t \to \infty$, the gradient flow in (517–518) converges to the following global minimizer of (515):

$$\lim_{t\to\infty} U(t) = \frac{1}{n}\operatorname{Tr}(W^* - Y_0X_0)\, I_n + Y_0X_0 \tag{534}$$

In particular, for the same initialization $Y_0 = \mathbf{0}_{n\times r}$ used in our low-rank analysis (Appendix D), the full-rank flow converges to

$$\lim_{t\to\infty} U(t) = \frac{1}{n}\operatorname{Tr}(W^*)\, I_n. \tag{535}$$

To demonstrate uniqueness of our solution $U(t)$ to the IVP in (517–518), suppose there exists an additional solution $U^*(t)$ which satisfies (517–518). We then have

$$\frac{dU^*(t)}{dt} = \operatorname{Tr}\Big(W^* - U^*(t)\Big) I_n \tag{536}$$

$$U^*(0) = Y_0X_0 \tag{537}$$

We can calculate

$$\frac{dU(t)}{dt} - \frac{dU^*(t)}{dt} = \frac{d}{dt}\Big(U(t) - U^*(t)\Big) = \operatorname{Tr}\Big(W^* - U(t)\Big) I_n - \operatorname{Tr}\Big(W^* - U^*(t)\Big) I_n \tag{538}$$

$$= \operatorname{Tr}\Big(U^*(t) - U(t)\Big) I_n \tag{539}$$

$$= -\operatorname{Tr}\Big(U(t) - U^*(t)\Big) I_n \tag{540}$$

or simply

$$\frac{d}{dt}\Big(U(t) - U^*(t)\Big) = -\operatorname{Tr}\Big(U(t) - U^*(t)\Big) I_n \tag{541}$$

Note that $U(0) - U^*(0) = Y_0X_0 - Y_0X_0 = \mathbf{0}_{n\times n}$. Now define $S : \mathbb{R}_0^+ \to \mathbb{R}^{n\times n}$ via

$$S(t) = U(t) - U^*(t) \tag{542}$$

$$S(0) = Y_0X_0 - Y_0X_0 = \mathbf{0}_{n\times n} \tag{543}$$

Then (541) gives

$$\frac{d}{dt}\Big(S(t)\Big) = -\operatorname{Tr}\Big(S(t)\Big) I_n \tag{544}$$

Take the trace of both sides of (544) to find

$$\frac{d}{dt}\operatorname{Tr}\Big(S(t)\Big) = -\operatorname{Tr}\Big(S(t)\Big)\operatorname{Tr}\left(I_n\right) \tag{545}$$

$$\frac{d}{dt}\operatorname{Tr}\Big(S(t)\Big) = -n\operatorname{Tr}\Big(S(t)\Big) \tag{546}$$

which is solved with the initial condition $\operatorname{Tr}\Big(S(0)\Big) = 0$ by

$$\operatorname{Tr}\Big(S(t)\Big) = 0 \tag{547}$$

Since the trace of $S(t)$ has only the trivial solution, (544) becomes

$$\frac{d}{dt}\Big(S(t)\Big) = \mathbf{0}_{n\times n} \tag{548}$$

This implies that $S(t)$ is a constant equal to its initial condition. Thus,

$$S(t) = \mathbf{0}_{n\times n} \tag{549}$$

for all $t \in [0, \infty)$. This implies that

$$U(t) = U^*(t) \tag{550}$$

and our solution $U(t)$ to the IVP in (517–518) must be unique.

H Low-Rank Approximation Error for the Trace-Squared Objective Relative to the Full-Rank Minimizer

In this work, we analyze both the full-rank and LoRA parameterizations of the trace-squared fine-tuning problem. Using gradient flow dynamics, we characterize the convergence behavior of classical full-rank gradient descent as well as low-rank–adapted gradient descent toward minimizers of their respective objectives:

$$\min_{\Delta W \in \mathbb{R}^{n\times n}} \frac{1}{2}\operatorname{Tr}^2(W^* - \Delta W) \tag{551}$$

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \frac{1}{2}\operatorname{Tr}^2(W^* - BA). \tag{552}$$

We find that application of classical GD to the full-rank problem converges to the following minimizer of (551):

$$\lim_{t\to\infty} U(t) = \frac{1}{n}\operatorname{Tr}(W^*)\, I_n \in \operatorname*{arg\,min}_{\Delta W\in\mathbb{R}^{n\times n}} \frac{1}{2}\operatorname{Tr}^2(W^* - \Delta W) \tag{553}$$

while application of low-rank adapted GD to the low-rank problem converges to the following minimizer of (552):

$$\lim_{t\to\infty} Y(t)X(t) = \frac{\operatorname{Tr}(W^*)}{\|X_0\|^2} X_0^T X_0 \in \operatorname*{arg\,min}_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \frac{1}{2}\operatorname{Tr}^2(W^* - BA) \tag{554}$$

Recall that, for both the low-rank and full-rank variants of our problem, we adopt the standard LoRA initialization scheme Hu et al. (2022), namely $U(0) = Y(0)X(0) = Y_0X_0$, where $Y_0 \in \mathbb{R}^{n\times r}$ is initialized as the zero matrix and $X_0 \in \mathbb{R}^{r\times n}$ has entries drawn i.i.d. from a centered Gaussian distribution with variance σ^2. Namely,

$$x_{ij}(0) \sim \mathcal{N}(0, \sigma^2). \tag{555}$$

where x_{ij} are the individual elements of X.

Recall from Appendices D and G that both the full-rank and low-rank gradient flows converge to global minimizers of the trace-squared objective, achieving zero final loss. We thus examine only the approximation error between the resulting full-rank and low-rank minimizers. In particular, we calculate the final relative error:

$$\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|^2}{\|U(t)\|^2} \tag{556}$$

The norm in the denominator is simply

$$\lim_{t\to\infty} \|U(t)\|^2 = \left\| \frac{1}{n}\operatorname{Tr}(W^*)\, I_n \right\|^2 \tag{557}$$

$$= \frac{|\operatorname{Tr}(W^*)|^2}{n^2} \|I_n\|^2 \tag{558}$$

$$= \frac{|\operatorname{Tr}(W^*)|^2}{n^2} \cdot n \tag{559}$$

$$= \frac{|\operatorname{Tr}(W^*)|^2}{n} \tag{560}$$

For the numerator, we find

$$\lim_{t\to\infty} \|Y(t)X(t) - U(t)\|^2 = \left\| \frac{\operatorname{Tr}(W^*)}{\|X_0\|^2} X_0^T X_0 - \frac{1}{n}\operatorname{Tr}(W^*)\, I_n \right\|^2 \tag{561}$$

$$= |\operatorname{Tr}(W^*)|^2 \left\| \frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right\|^2 \tag{562}$$

Calculate the Frobenius norm above:

$$\left\| \frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right\|^2 = \operatorname{Tr}\left(\left[\frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right]^T \left[\frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right] \right) \tag{563}$$

$$= \operatorname{Tr}\left(\left[\frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right] \left[\frac{X_0^T X_0}{\|X_0\|^2} - \frac{1}{n} I_n \right] \right) \tag{564}$$

$$= \operatorname{Tr}\left(\frac{X_0^T X_0 X_0^T X_0}{\|X_0\|^4} - \frac{2}{n} \frac{X_0^T X_0}{\|X_0\|^2} + \frac{1}{n^2} I_n \right) \tag{565}$$

$$= \operatorname{Tr}\left(\frac{X_0^T X_0 X_0^T X_0}{\|X_0\|^4} \right) - \frac{2}{n} \operatorname{Tr}\left(\frac{X_0^T X_0}{\|X_0\|^2} \right) + \frac{1}{n^2} \operatorname{Tr}(I_n) \tag{566}$$

$$= \frac{1}{\|X_0\|^4} \operatorname{Tr}\left(X_0^T X_0 X_0^T X_0 \right) - \frac{2}{n \|X_0\|^2} \operatorname{Tr}\left(X_0^T X_0 \right) + \frac{1}{n^2} \cdot n \tag{567}$$

$$= \frac{\left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - \frac{2 \|X_0\|^2}{n \|X_0\|^2} + \frac{1}{n} \tag{568}$$

$$= \frac{\left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - \frac{1}{n} \tag{569}$$

We then have

$$\lim_{t \to \infty} \|Y(t)X(t) - U(t)\|^2 = |\operatorname{Tr}(W^*)|^2 \left(\frac{\left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - \frac{1}{n} \right) \tag{570}$$

The square of the relative error between the low-rank and full-rank minimizers for arbitrary $X_0 \neq \mathbf{0}_{r \times n}$ is then

$$\lim_{t \to \infty} \frac{\|Y(t)X(t) - U(t)\|^2}{\|U(t)\|^2} = |\operatorname{Tr}(W^*)|^2 \left(\frac{\left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - \frac{1}{n} \right) \cdot \frac{n}{|\operatorname{Tr}(W^*)|^2} \tag{571}$$

$$= \frac{n \left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - 1 \tag{572}$$

Note that the expression above is non-negative since $\sqrt{n} \left\| X_0^T X_0 \right\| \geq \|X_0\|^2$ by Cauchy-Schwarz applied to the eigenvalues of $X_0^T X_0$. We then have the final relative error

$$\lim_{t \to \infty} \frac{\|Y(t)X(t) - U(t)\|}{\|U(t)\|} = \sqrt{\frac{n \left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} - 1} \tag{573}$$

$$= \frac{\sqrt{n \left\| X_0^T X_0 \right\|^2 - \|X_0\|^4}}{\|X_0\|^2} \tag{574}$$

To reiterate, the expression above for the relative error between the low-rank and full-rank gradient flow minimizers is valid for arbitrary $X_0 \neq \mathbf{0}_{r \times n}$. However, we can study the *expected* approximation error as a function of r and n in the context of the standard LoRA initialization scheme in (555).

Initializing $X_0 \in \mathbb{R}^{n \times r}$ as in (555), begin by calculating the expected value of the ratio

$$\frac{\left\| X_0^T X_0 \right\|^2}{\|X_0\|^4} \tag{575}$$

Write

$$X_0 = rZ \tag{576}$$

where $r = \|X_0\|$ and $Z = \dfrac{X_0}{\|X_0\|}$. Then (575) becomes

$$\frac{\left\|X_0^T X_0\right\|^2}{\|X_0\|^4} = \frac{\left\|r^2 Z^T Z\right\|^2}{\|r\|^4} = \frac{r^4 \left\|Z^T Z\right\|^2}{r^4} = \left\|Z^T Z\right\|^2 = \sum_{a=1}^{r}\sum_{b=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{n} z_{ia}\, z_{ib}\, z_{ja}\, z_{jb} \tag{577}$$

To calculate the expectation of the expression above, we then need the expectation of the individual scalar entries of Z. Let $\mathbf{z} \in \mathbb{R}^{nr}$ denote the vector of stacked entries of Z, and let $\mathbf{x_0} \in \mathbb{R}^{nr}$ denote the vector of stacked entries of X_0. Use $\|\cdot\|_2$ to represent the vector Euclidean norm. Then

$$\mathbf{z} = \frac{\mathbf{x_0}}{\|\mathbf{x_0}\|_2} \tag{578}$$

We have that

$$\mathbf{x_0} \sim \mathcal{N}\left(0, \sigma^2 I_{nr}\right) \tag{579}$$

By the properties of any scalar random variable ξ that

$$\operatorname{Var}(\sigma\xi) = \sigma^2 \operatorname{Var}(\xi) \tag{580}$$

$$\mathbb{E}\left[\sigma\xi\right] = \sigma\mathbb{E}\left[\xi\right] \tag{581}$$

we can write

$$\mathbf{x_0} = \sigma\mathbf{g} \tag{582}$$

for $\mathbf{g} \in \mathbb{R}^{nr}$, where

$$\mathbf{g} \sim \mathcal{N}\left(0, I_{nr}\right) \tag{583}$$

It is known that $\dfrac{\mathbf{g}}{\|\mathbf{g}\|_2}$ is uniformly distributed on the unit sphere in $\mathbb{R}^{nr}$ Vershynin (2018). We then immediately have

$$\mathbf{z} = \frac{\mathbf{x_0}}{\|\mathbf{x_0}\|_2} = \frac{\sigma\mathbf{g}}{\sigma\|\mathbf{g}\|_2} = \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \tag{584}$$

and therefore $\mathbf{z}$ is uniformly distributed on the unit sphere. Now denote $\rho := \|\mathbf{g}\|_2$. We know that $\mathbf{z}$ and ρ are independent Vershynin (2018). Choose arbitrary indices $i, j \in \{1, \ldots, nr\}$. For the corresponding scalar components of $\mathbf{z}$ and $\mathbf{g}$, we have

$$z_i^2 = \frac{g_i^2}{\rho^2} \qquad \text{and} \qquad z_i^2 z_j^2 = \frac{g_i^2 g_j^2}{\rho^4} \tag{585}$$

or

$$\rho^2 z_i^2 = g_i^2 \qquad \text{and} \qquad \rho^4 z_i^2 z_j^2 = g_i^2 g_j^2 \tag{586}$$

Taking advantage of the independence of $\mathbf{z}$ and ρ, we find

$$\mathbb{E}\left[\rho^2\right]\mathbb{E}\left[z_i^2\right] = \mathbb{E}\left[g_i^2\right] \qquad \text{and} \qquad \mathbb{E}\left[\rho^4\right]\mathbb{E}\left[z_i^2 z_j^2\right] = \mathbb{E}\left[g_i^2 g_j^2\right] \tag{587}$$

We can thus calculate

$$\mathbb{E}\left[z_i^2\right] = \frac{\mathbb{E}\left[g_i^2\right]}{\mathbb{E}\left[\rho^2\right]} \qquad \text{and} \qquad \mathbb{E}\left[z_i^2 z_j^2\right] = \frac{\mathbb{E}\left[g_i^2 g_j^2\right]}{\mathbb{E}\left[\rho^4\right]} \tag{588}$$

By standard results for Gaussian normal distributions, we have

$$\mathbb{E}\left[g_i^2\right] = 1 \tag{589}$$

$$\mathbb{E}\left[g_i^2 g_j^2\right] = \begin{cases} 1 & i \neq j \\ 3 & i = j \end{cases} \tag{590}$$

Since $\mathbf{g} \in \mathbb{R}^{nr}$ and $\rho = \|\mathbf{g}\|_2$, we know that

$$\rho^2 = \|\mathbf{g}\|_2^2 = \sum_{i=1}^{nr} g_i^2 \tag{591}$$

Therefore, ρ^2 follows a chi-squared distribution with nr degrees of freedom, or $\rho^2 \sim \chi^2_{nr}$. By standard results for first and second moments of chi-squared distributions, we have

$$\mathbb{E}\left[\rho^2\right] = nr \tag{592}$$

$$\mathbb{E}\left[\rho^4\right] = nr(nr+2) \tag{593}$$

Putting these results together, (588) becomes

$$\mathbb{E}\left[z_i^2\right] = \frac{1}{nr} \qquad \text{and} \qquad \mathbb{E}\left[z_i^2 z_j^2\right] = \begin{cases} \dfrac{1}{nr(nr+2)} & i \neq j \\ \dfrac{3}{nr(nr+2)} & i = j \end{cases} \tag{594}$$

Thus we have the second and fourth moments of the scalar components of $\mathbf{z}$, or, equivalently, the scalar compenents of Z.

We are now ready to compute the expectation of $\dfrac{\left\|X_0^T X_0\right\|^2}{\left\|X_0\right\|^4}$. Taking the expectation of (577) gives

$$\mathbb{E}\left[\frac{\left\|X_0^T X_0\right\|^2}{\left\|X_0\right\|^4}\right] = \sum_{a=1}^{r}\sum_{b=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}\right] \tag{595}$$

The expectation

$$\mathbb{E}[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}]$$

depends only on whether the row indices a, b and the column indices i, j coincide, and hence takes one of four possible values corresponding to the cases where $i = j$ or $i \neq j$ and $a = b$ or $a \neq b$:

A. $a = b$ and $i = j$. In this case, all four factors coincide, and by (594),

$$\mathbb{E}[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}] = \mathbb{E}\left[z_{ia}^4\right] = \frac{3}{nr(nr+2)}.$$

B. $a = b$ and $i \neq j$. In this case,

$$\mathbb{E}[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}] = \mathbb{E}\left[z_{ia}^2 z_{ja}^2\right] = \frac{1}{nr(nr+2)},$$

which is again given by (594).

C. $a \neq b$ and $i = j$. In this case,

$$\mathbb{E}[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}] = \mathbb{E}\left[z_{ia}^2 z_{ib}^2\right] = \frac{1}{nr(nr+2)}.$$

D. $a \neq b$ and $i \neq j$. In this case, the four indices are distinct, and, by symmetry on the unit sphere,

$$\mathbb{E}[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}] = 0.$$

See Appendix I for an explanation of this case.

Note that, for the sum in (595), we will have nr terms fall under case A (r choices for a and n choices for i). We will have $nr(n-1)$ terms satisfy case B (r choices for a and $n(n-1)$ choices for i). For case C, we will have $nr(r-1)$ terms (n choices for i and $r(r-1)$ choices for a). Terms for case D contribute nothing to the expectation value and do not need to be counted. We finally calculate our sum:

$$\sum_{a=1}^{r}\sum_{b=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[z_{ia}\, z_{ib}\, z_{ja}\, z_{jb}\right] = \frac{3nr}{nr(nr+2)} + \frac{nr(n-1)}{nr(nr+2)} + \frac{nr(r-1)}{nr(nr+2)} \tag{596}$$

$$= \frac{3}{(nr+2)} + \frac{(n-1)}{(nr+2)} + \frac{(r-1)}{(nr+2)} \tag{597}$$

$$= \frac{n+r+1}{nr+2} \tag{598}$$

So we arrive at

$$\mathbb{E}\left[\frac{\left\|X_0^T X_0\right\|^2}{\|X_0\|^4}\right] = \frac{n+r+1}{nr+2} \tag{599}$$

The expectation of the *square* of the final relative error between the rank-r and full-rank solutions is then

$$\mathbb{E}\left[\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|^2}{\|U(t)\|^2}\right] = \mathbb{E}\left[\frac{n\left\|X_0^T X_0\right\|^2}{\|X_0\|^4} - 1\right] \tag{600}$$

$$= n\mathbb{E}\left[\frac{\left\|X_0^T X_0\right\|^2}{\|X_0\|^4}\right] - 1 \tag{601}$$

$$= n \cdot \frac{n+r+1}{nr+2} - 1 \tag{602}$$

$$= \frac{n^2+n-2}{nr+2} \tag{603}$$

From our expectation for the square of the final relative error, we can use Jensen's inequality to find an upper bound on the expectation of the final relative error itself. Define

$$R^2 := \lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|^2}{\|U(t)\|^2}. \tag{604}$$

Note that $R^2 \geq 0$ and $\mathbb{E}[R^2]$ is finite by (603). Recall that the square root function $\phi(x) = \sqrt{x}$ is concave on $[0, \infty)$. By Jensen's inequality Dekking et al. (2005), we have

$$\mathbb{E}\left[\sqrt{R^2}\right] \leq \sqrt{\mathbb{E}\left[R^2\right]} \tag{605}$$

Combining this with (603), we arrive at

$$\mathbb{E}\left[\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|}{\|U(t)\|}\right] \leq \sqrt{\frac{n^2+n-2}{nr+2}} \tag{606}$$

I VANISHING FOURTH-ORDER MOMENT IN APPENDIX H

Consider the random vector $\mathbf{z} \in \mathbb{R}^{nr}$ which is uniformly distributed on the unit sphere in $\mathbb{R}^{nr}$. Let

$$S^{nr-1} = \{\mathbf{v} \in \mathbb{R}^{nr} : \|\mathbf{v}\|_2 = 1\}$$

denote the unit sphere, and let $|S^{nr-1}|$ denote its surface area. Throughout this section, integrals over S^{nr-1} are taken with respect to its surface measure dS.

The uniform distribution on S^{nr-1} is then given by the probability measure

$$d\mathbb{P}(\mathbf{z}) = \frac{1}{|S^{nr-1}|}\, dS(\mathbf{z}),$$

which is normalized since

$$\frac{1}{|S^{nr-1}|} \int_{S^{nr-1}} dS(\mathbf{z}) = 1.$$

Let z_i, z_j, z_k, z_l denote any four distinct scalar entries of the random vector $\mathbf{z}$. We wish to show that

$$\mathbb{E}\left[z_i z_j z_k z_l\right] = 0 \tag{607}$$

We can calculate (607) directly by integrating over S^{nr-1}. We have

$$\mathbb{E}\left[z_i z_j z_k z_l\right] = \int_{S^{nr-1}} z_i z_j z_k z_l \, d\mathbb{P}(\mathbf{z}) \tag{608}$$

$$= \frac{1}{|S^{nr-1}|} \int_{S^{nr-1}} z_i z_j z_k z_l \, dS(\mathbf{z}) \tag{609}$$

Note that the unit sphere S^{nr-1} is invariant under coordinate sign flips. In particular, let $T : \mathbb{R}^{nr} \to \mathbb{R}^{nr}$ denote the orthogonal transformation that flips the sign of the ith coordinate and leaves all others unchanged. Then $T(S^{nr-1}) = S^{nr-1}$, and the surface measure $dS(\mathbf{z})$ is invariant under T. Under this transformation,

$$z_i z_j z_k z_l \;\mapsto\; (-z_i) z_j z_k z_l = -\, z_i z_j z_k z_l.$$

Therefore,

$$\int_{S^{nr-1}} z_i z_j z_k z_l \, dS(\mathbf{z}) = -\int_{S^{nr-1}} z_i z_j z_k z_l \, dS(\mathbf{z}),$$

which implies that the integral vanishes.

J Learning Dynamics for Low-rank Matrix Approximation

We wish to analyze the learning dynamics of the low-rank optimizer $BA \in \mathbb{R}^{n\times m}$ produced by applying LoRA to the finetuning problem

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times m}}} \frac{1}{2}\left\|W^* - BA\right\|^2 \tag{610}$$

where $r << \min(n,m)$. The partial gradients for our objective $h(B,A) = \frac{1}{2}\left\|W^* - BA\right\|^2$ are given by

$$\nabla_A h(B,A) = -B^T(W^* - BA) \tag{611}$$

$$\nabla_B h(B,A) = -(W^* - BA)A^T \tag{612}$$

Assumption J.1 (Bounded Domain)**.** *We optimize (610) over the subspace $\mathcal{D}_{R'} \subseteq \Theta$, where $R' > 0$ is some finite number. In other words, we assume that there exists some $R' > 0$ such that the norms on both B and A remain bounded above by R' during training.*

Note that, since the objective in (610) is bounded below by 0, that Assumption J.1 holds for this problem on finite finetuning horizons for sufficiently small stepsize as outlined in Remark 2.2. This assumption above ensures that the objective gradient remains bounded during training:

Remark J.2 (Boundedness of Gradient for Squared Frobenius Loss)**.** *The objective gradient $\nabla h(B,A)$ remains bounded above during training. That is, for all $B, A \in \mathcal{D}_{R'}$, we have*

$$\begin{aligned}
\left\|\nabla h(B,A)\right\|^2 &= \left\|\nabla_B h(B,A)\right\|^2 + \left\|\nabla_A h(B,A)\right\|^2 && (613)\\
&= \left\|(W^* - BA)A^T\right\|^2 + \left\|B^T(W^* - BA)\right\|^2 && (614)\\
&\leq \left\|W^* - BA\right\|^2 \left\|A\right\|^2 + \left\|W^* - BA\right\|^2 \left\|B\right\|^2 && (615)\\
&= \left\|W^* - BA\right\|^2 \left(\left\|A\right\|^2 + \left\|B\right\|^2\right) && (616)\\
&\leq 2R'^2 \left\|W^* - BA\right\|^2 && (617)\\
&\leq 2R'^2 \left(\left\|W^*\right\| + \left\|BA\right\|\right)^2 && (618)\\
&\leq 2R'^2 \left(\left\|W^*\right\| + \left\|B\right\| \left\|A\right\|\right)^2 && (619)\\
&\leq 2R'^2 \left(\left\|W^*\right\| + R'^2\right)^2 && (620)
\end{aligned}$$

So $\left\|\nabla h(B,A)\right\| \leq \sqrt{2}R'\left(\left\|W^*\right\| + R'^2\right)$ throughout training.

We then have Lipschitz smoothness during training:

Lemma J.3 (Lipschitz Smoothness for Squared Frobenius Objective)**.** *Our objective gradient ∇h is Lipschitz smooth in our training domain. Namely, there exists $L_{R'} > 0$ such that, for any $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$, we have*

$$\left\|\nabla h(B_1,A_1) - \nabla h(B_2,A_2)\right\| \leq L_{R'}\left\|\Big(B_1,A_1\Big) - \Big(B_2,A_2\Big)\right\| \tag{621}$$

Proof of this lemma can be found in Appendix K.

Having shown that the squared Frobenius objective in (610) satisfies Assumptions B.2 and B.7 from Appendix B, the ODEs describing the learning dynamics of $h(B,A)$ under LoRA are given by (see appendix B)

$$\frac{dY(t)}{dt} = \left(W^* - YX\right)X^T \tag{622}$$

$$\frac{dX(t)}{dt} = Y^T\left(W^* - YX\right) \tag{623}$$

for any $t \in [0,T]$, where $T > 0$ is arbitrary.

Unlike the trace-squared objective in Appendix D, we are not aware of a closed-form solution to the ODE above under conventional element-wise initialization, i.e., when all rm entries of X_0 are drawn independently from a normal distribution. We will show, however, that the solution to (622–623) converges to the theoretical rank-r minimizer Eckart & Young (1936) of (610) when utilizing the spectral initialization schema.

We begin by assuming that W^* has rank strictly greater than r:

Assumption J.4 (Rank of W^*)**.** *Let $k = \mathrm{rank}(W^*)$. We assume that*

$$r < k \leq \min(n, m). \tag{624}$$

Note that violation of Assumption J.4 renders (610) outside the scope of low-rank approximation.

Denote the singular value decomposition (SVD) of W^* by

$$W^* = U\Sigma^* V^T, \tag{625}$$

where $\Sigma^* \in \mathbb{R}^{n\times m}$ is the rectangular diagonal matrix whose main diagonal entries are the singular values of W^* (in non-increasing order). Here, $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are the left and right singular vector matrices of W^*, respectively. Recall that U and V are orthogonal.

For spectral initialization of X_0 and Y_0, define the rotated variables $\tilde{Y} \in \mathbb{R}^{n\times r}$ and $\tilde{X} \in \mathbb{R}^{r\times m}$ by

$$\tilde{Y} = U^T Y, \tag{626}$$

$$\tilde{X} = XV, \tag{627}$$

so that

$$Y = U\tilde{Y}, \tag{628}$$

$$X = \tilde{X}V^T. \tag{629}$$

To ensure $Y_0 X_0 = \mathbf{0}_{n\times m}$, we initialize $\tilde{Y}_0 = \mathbf{0}_{n\times r}$ and draw the r diagonal entries of $\tilde{X}_0$ from a centered Gaussian distribution:

$$\tilde{x}_{ii}(0) \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, r, \tag{630}$$

where $\tilde{x}_{ii}$ denotes the (i, i) entry of $\tilde{X}$. The off-diagonal elements of $\tilde{X}$ are initialized at zero.

With the transformations in (625) and (628–629) in mind, our ODEs in (622–623) become

$$U\frac{d\tilde{Y}(t)}{dt} = \left(U\Sigma^* V^T - U\tilde{Y}\tilde{X}V^T\right) V\tilde{X}^T \tag{631}$$

$$\frac{d\tilde{X}(t)}{dt}V^T = \tilde{Y}^T U^T \left(U\Sigma^* V^T - U\tilde{Y}\tilde{X}V^T\right) \tag{632}$$

Note that $U^T U$ and $V^T V$ both give the identity matrix since U and V are orthogonal. Left-multiplying U^T on both sides of (631) and right-multiplying V on both sides of (632), our ODEs become

$$\frac{d\tilde{Y}(t)}{dt} = \left(\Sigma^* - \tilde{Y}\tilde{X}\right)\tilde{X}^T \tag{633}$$

$$\frac{d\tilde{X}(t)}{dt} = \tilde{Y}^T\left(\Sigma^* - \tilde{Y}\tilde{X}\right) \tag{634}$$

As a consequence of our initialization scheme for $\tilde{X}_0$ and $\tilde{Y}_0$, all matrices in our transformed ODEs above remain diagonal for all $t \in [0, T]$, implying that the scalar dynamics for the ith diagonal element of $\tilde{Y}$ and $\tilde{X}$ decouple to

$$\frac{d\tilde{y}_{ii}}{dt} = \left(s_i^* - \tilde{y}_{ii}\tilde{x}_{ii}\right)\tilde{x}_{ii} \tag{635}$$

$$\frac{d\tilde{x}_{ii}}{dt} = \left(s_i^* - \tilde{y}_{ii}\tilde{x}_{ii}\right)\tilde{y}_{ii} \tag{636}$$

$$\tilde{y}_{ii}(0) = 0 \tag{637}$$
$$\tilde{x}_{ii}(0) \sim \mathcal{N}(0, \sigma^2) \tag{638}$$

where s_i^* denotes the ith largest singular value of W^*. Since $\tilde{Y}$ and $\tilde{X}$ each contain r diagonal elements, (642–638) give the learning dynamics for the first r diagonals of $\tilde{Y}(t)\tilde{X}(t)$. Note that the dimensions of $\tilde{Y}$ and $\tilde{X}$ ensure that the remaining diagonal values of $\tilde{Y}(t)\tilde{X}(t)$ remain zero for all t.

Assumption J.5 (Nonzero Initialization for $\tilde{x}_{ii}$)**.** *We assume $\tilde{x}_{ii}(0) \neq 0$ for all $i \in \{1, \ldots, r\}$, which holds almost surely under the Gaussian initialization in* (638)*.*

Note that violation of Assumption J.5 above initializes $\left(\tilde{x}_{ii}(t), \tilde{y}_{ii}(t)\right)$ at the saddle point $(0, 0)$ of the dynamical system in (642–638). In this case, the ith diagonal of $\tilde{Y}(t)\tilde{X}(t)$ will fail to converge to s_i^*, and our LoRA gradient flow will not converge to the theoretical rank-r minimizer of (610).

Since (642–638) hold for all $i \in \{1, \ldots, r\}$, it suffices to solve the system for a single generic index i. For convenience, drop the indices in our notation and denote

$$y := \tilde{y}_{ii} \tag{639}$$
$$x := \tilde{x}_{ii} \tag{640}$$
$$s^* := s_i^* \tag{641}$$

We then wish to solve the IVP

$$\frac{dy}{dt} = (s^* - yx)\, x \tag{642}$$
$$\frac{dx}{dt} = (s^* - yx)\, y \tag{643}$$
$$y(0) = 0 \tag{644}$$
$$x(0) \sim \mathcal{N}(0, \sigma^2) \tag{645}$$

Conveniently, we already implicitly have the solution to the IVP above in Appendix D. There we show that the solution to the matrix-valued IVP:

$$\frac{dY(t)}{dt} = \mathrm{Tr}(W^* - YX)X^T \tag{646}$$
$$\frac{dX(t)}{dt} = \mathrm{Tr}(W^* - YX)Y^T \tag{647}$$
$$Y(0) = \mathbf{0}_{n \times r} \tag{648}$$
$$X(0) = X_0 \tag{649}$$

is given by

$$Y(t) = q(t)\, \mathrm{Tr}(W^*)\, X_0^T \tag{650}$$
$$X(t) = p(t) X_0, \tag{651}$$

where $p(t)$ and $q(t)$ are given in Appendix D.

Recall that the above matrix-valued IVP is solved in Appendix D for any $X_0 \in \mathbb{R}^{m \times r}$ such that $\|X_0\| \neq 0$, as well as any W^* such that $\mathrm{Tr}(W^*) \neq 0$. While in practice LoRA uses $r << \min(n, m)$ for parameter efficiency, we make no assumption in Appendix D on the dimensions of Y, X, and W^* when finding the solution of (646–649). Observe that when $n = m = r = 1$, the trace operator is trivial, and $\mathrm{Tr}(W^* - YX) = W^* - yx$. Importantly, Appendix D then presents the resulting *scalar* IVP

$$\frac{dy(t)}{dt} = (w^* - yx)x \tag{652}$$
$$\frac{dx(t)}{dt} = (w^* - yx)y \tag{653}$$
$$y(0) = 0 \tag{654}$$
$$x(0) = x_0 \tag{655}$$

along with the solution

$$y(t) = q_s(t) w^* x_0 \tag{656}$$
$$x(t) = p_s(t) x_0 \tag{657}$$

for arbitrary nonzero w^* and x_0. Since the first r singular values of W^* are implicitly nonzero based on Assumption J.4, we can set $w^* = s^*$ to arrive at the solution of (642–645):

$$y(t) = q_s(t) s^* x_0 \tag{658}$$
$$x(t) = p_s(t) x_0 \tag{659}$$

Note that, for the IVP in (642–645) under Assumption J.5, x_0 is a nonzero random variable drawn from $\mathcal{N}(0, \sigma^2)$. The scalar equivalents of $q(t)$ and $p(t)$, which we denote $q_s(t)$ and $p_s(t)$, are given by

$$p_s(t) := \frac{1}{\sqrt{2x_0^2 \sinh\big(\sqrt{\xi_1}(t+\xi_2)\big) + 4s^*}} \left[\frac{x_0^2}{\sqrt{s^*}} \sinh\left(\frac{\sqrt{\xi_1}}{2} t\right) + \frac{\sqrt{x_0^4 + 4s^{*2}}}{\sqrt{s^*}} \cosh\left(\frac{\sqrt{\xi_1}}{2} t\right) \right] \tag{660}$$

$$q_s(t) := \frac{1}{\sqrt{2x_0^2 \sinh\big(\sqrt{\xi_1}(t+\xi_2)\big) + 4s^*}} \cdot \frac{2}{\sqrt{s^*}} \sinh\left(\frac{\sqrt{\xi_1}}{2} t\right) \tag{661}$$

where ξ_1 and ξ_2 are equal to

$$\xi_1 = {x_0}^4 + 4{s^*}^2 \tag{662}$$
$$\xi_2 = \frac{1}{\sqrt{{x_0}^4 + 4{s^*}^2}} \operatorname{arsinh}\left(\frac{{x_0}^2}{2s^*}\right) \tag{663}$$

Readers of Appendix D should note we drop any absolute value bars above on s^* since W^*'s singular values are nonnegative. We thus have that the ith diagonal element of $\tilde{Y}(t)\tilde{X}(t)$ at time $t \in [0, T]$ is given by

$$y(t)x(t) = p_s(t) q_s(t) s^* x_0^2 \tag{664}$$

Since our closed-form expressions for $y(t)$ and $x(t)$ solve (635–638) on $[0, T]$ for any $T > 0$, they define a single solution to the same ODE system for all $t \in [0, \infty)$. Thus, we can study the convergence of $y(t)x(t)$ by taking their limit as $t \to \infty$. Recall that, as $t \to \infty$,

$$\sinh\left(\frac{\sqrt{\xi_1}}{2} t\right) \sim \frac{1}{2} e^{\frac{\sqrt{\xi_1}}{2} t}, \tag{665}$$
$$\cosh\left(\frac{\sqrt{\xi_1}}{2} t\right) \sim \frac{1}{2} e^{\frac{\sqrt{\xi_1}}{2} t}, \tag{666}$$

and

$$\sinh\Big(\sqrt{\xi_1}(t+\xi_2)\Big) \sim \frac{1}{2} e^{\sqrt{\xi_1}(t+\xi_2)} \tag{667}$$

This gives us

$$\lim_{t\to\infty} \frac{\sinh\left(\frac{\sqrt{\xi_1}}{2} t\right)}{\sqrt{2x_0^2 \sinh\big(\sqrt{\xi_1}(t+\xi_2)\big) + 4s^*}} = \lim_{t\to\infty} \frac{\cosh\left(\frac{\sqrt{\xi_1}}{2} t\right)}{\sqrt{2x_0^2 \sinh\big(\sqrt{\xi_1}(t+\xi_2)\big) + 4s^*}} \tag{668}$$

$$= \lim_{t\to\infty} \frac{\frac{1}{2} e^{\frac{\sqrt{\xi_1}}{2} t}}{\sqrt{2x_0^2 \cdot \frac{e^{\sqrt{\xi_1}(t+\xi_2)}}{2} + 4s^*}} \tag{669}$$

$$= \lim_{t\to\infty} \frac{e^{\frac{\sqrt{\xi_1}}{2} t}}{2\sqrt{x_0^2 \cdot e^{\sqrt{\xi_1}(t+\xi_2)} + 4s^*}} \cdot \frac{e^{-\frac{\sqrt{\xi_1}}{2} t}}{e^{-\frac{\sqrt{\xi_1}}{2} t}} \tag{670}$$

$$= \lim_{t\to\infty} \frac{1}{2\sqrt{x_0^2 \cdot e^{\xi_2\sqrt{\xi_1}} + 4s^* e^{-\sqrt{\xi_1}t}}} \tag{671}$$

$$= \frac{1}{2|x_0| e^{\frac{\xi_2\sqrt{\xi_1}}{2}}} \tag{672}$$

Note that

$$e^{\xi_2\sqrt{\xi_1}} = e^{\operatorname{arsinh}\left(\frac{x_0^2}{2s^*}\right)} \tag{673}$$

$$= e^{\ln\left(\frac{x_0^2}{2s^*} + \sqrt{1+\left(\frac{x_0^2}{2s^*}\right)^2}\right)} \tag{674}$$

$$= \frac{x_0^2}{2s^*} + \sqrt{1+\left(\frac{x_0^2}{2s^*}\right)^2} \tag{675}$$

$$= \frac{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}{2s^*} \tag{676}$$

So (672) becomes

$$\frac{1}{2|x_0|} \cdot \sqrt{\frac{2s^*}{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} = \frac{\sqrt{s^*}}{\sqrt{2}|x_0|\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} \tag{677}$$

We have

$$\lim_{t\to\infty} \frac{\sinh\left(\frac{\sqrt{\xi_1}}{2}t\right)}{\sqrt{2x_0^2\sinh\left(\sqrt{\xi_1}(t+\xi_2)\right) + 4s^*}} = \lim_{t\to\infty} \frac{\cosh\left(\frac{\sqrt{\xi_1}}{2}t\right)}{\sqrt{2x_0^2\sinh\left(\sqrt{\xi_1}(t+\xi_2)\right) + 4s^*}} \tag{678}$$

$$= \frac{\sqrt{s^*}}{\sqrt{2}|x_0|\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} \tag{679}$$

For brevity, define $\omega_2 := \sinh\left(\frac{\sqrt{\xi_1}}{2}t\right)$, $\gamma_2 := \cosh\left(\frac{\sqrt{\xi_1}}{2}t\right)$, and $\omega_3 := \sinh\left(\sqrt{\xi_1}(t+\xi_2)\right)$. Then we have

$$\lim_{t\to\infty} p_s(t) = \lim_{t\to\infty} \frac{1}{\sqrt{2x_0^2\omega_3 + 4s^*}} \left[\frac{x_0^2}{\sqrt{s^*}}\omega_2 + \frac{\sqrt{\xi_1}}{\sqrt{s^*}}\gamma_2\right] \tag{680}$$

$$= \lim_{t\to\infty} \frac{x_0^2}{\sqrt{s^*}} \cdot \frac{\omega_2}{\sqrt{2x_0^2\omega_3 + 4s^*}} + \frac{\sqrt{\xi_1}}{\sqrt{s^*}} \cdot \frac{\gamma_2}{\sqrt{2x_0^2\omega_3 + 4s^*}} \tag{681}$$

$$= \frac{x_0^2}{\sqrt{s^*}} \cdot \frac{\sqrt{s^*}}{\sqrt{2}\,|x_0|\sqrt{x_0^2 + \sqrt{\xi_1}}} + \frac{\sqrt{\xi_1}}{\sqrt{s^*}} \cdot \frac{\sqrt{s^*}}{\sqrt{2}\,|x_0|\sqrt{x_0^2 + \sqrt{\xi_1}}} \tag{682}$$

$$= \frac{x_0^2 + \sqrt{\xi_1}}{\sqrt{2}\,|x_0|\sqrt{x_0^2 + \sqrt{\xi_1}}} \tag{683}$$

$$= \frac{\sqrt{x_0^2 + \sqrt{\xi_1}}}{\sqrt{2}\,|x_0|} \tag{684}$$

We also have

$$\lim_{t\to\infty} q_s(t) = \lim_{t\to\infty} \frac{2}{\sqrt{s^*}} \cdot \frac{\sinh\left(\frac{\sqrt{\xi_1}}{2}t\right)}{\sqrt{2x_0^2\sinh\left(\sqrt{\xi_1}(t+\xi_2)\right) + 4s^*}} \tag{685}$$

$$= \frac{2}{\sqrt{s^*}} \cdot \frac{\sqrt{s^*}}{\sqrt{2}|x_0|\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} \tag{686}$$

$$= \frac{2}{\sqrt{2}|x_0|\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} \tag{687}$$

We finally have all we need to arrive at the convergence of the ith diagonal of $\tilde{Y}(t)\tilde{X}(t)$. Taking the limit of (664) as $t \to \infty$, we find

$$\lim_{t\to\infty} y(t)x(t) = \lim_{t\to\infty} p_s(t)q_s(t)s^* x_0^2 \tag{688}$$

$$= s^* x_0^2 \cdot \frac{\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}}{\sqrt{2}|x_0|} \cdot \frac{2}{\sqrt{2}|x_0|\sqrt{x_0^2 + \sqrt{x_0^4 + 4s^{*2}}}} \tag{689}$$

$$= s^* \tag{690}$$

Return to indexed notation. We thus find that, for $i \in \{1, \ldots, r\}$, the convergence of the ith diagonal of $\tilde{Y}\tilde{X}$ is given by

$$\lim_{t\to\infty} \tilde{y}_{ii}(t)\tilde{x}_{ii}(t) = s_i^* = (\Sigma^*)_{ii} \tag{691}$$

where s_i^* denotes the ith largest singular value of W^*.

Consequently, under the spectral initialization scheme in (637-638), the solution to the ODE system in (633–634) satisfies

$$\lim_{t\to\infty} \tilde{Y}(t)\tilde{X}(t) = \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{692}$$

where $\Sigma_r^* \in \mathbb{R}^{r\times r}$ denotes the diagonal matrix whose entries are the top r singular values of W^* in non-increasing order. Recover the converged optimizer $\lim_{t\to\infty} Y(t)X(t)$ for (610) by recalling the transformations in (628–629). Left multiplying (692) by U and right multiplying by V^T, we arrive at

$$\lim_{t\to\infty} U\tilde{Y}(t)\tilde{X}(t)V^T = U \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T \tag{693}$$

or simply

$$Y^*X^* := \lim_{t\to\infty} Y(t)X(t) = U \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T \tag{694}$$

We thus obtain the final rank-r solution produced by the LoRA gradient flow for the problem in (610). We can calculate the final loss as

$$\frac{1}{2}\|W^* - Y^*X^*\|^2 = \frac{1}{2}\left\|U\Sigma^*V^T - U \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T\right\|^2 \tag{695}$$

$$= \frac{1}{2}\left\|U\left(\Sigma^* - \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) V^T\right\|^2 \tag{696}$$

$$= \frac{1}{2}\left\|\Sigma^* - \begin{bmatrix} \Sigma_r^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right\|^2 \tag{697}$$

$$= \frac{1}{2}\sum_{i=r+1}^{\min(n,m)} s_i^{*2} \tag{698}$$

Note also that our LoRA rank-r solution Y^*X^* is equivalent to

$$Y^*X^* = U_r\Sigma_r^*V_r^T \tag{699}$$

where U_r and V_r (truncations of U and V) are the matrices of left and right singular vectors for the top r singular values of W^*. Our final loss and optimizer in (698–699) coincide with the theoretical minimum loss and optimizer for (610) characterized by the Eckart–Young–Mirsky theorem Eckart & Young (1936). We thus find via our gradient flow analysis that LoRA converges to the optimal rank-r solution of (610) under spectral initialization.

K PROOF OF LEMMA J.3

Lemma K.1 (Lipschitz Smoothness Squared Frobenius Objective)**.** *Define $h : \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m} \to \mathbb{R}$ via*

$$h(B, A) := \frac{1}{2} \|W^* - BA\|^2 \tag{700}$$

for constant $W^* \in \mathbb{R}^{n\times m}$. Then the objective gradient ∇h is Lipschitz smooth in our training domain. Namely, there exists $L_{R'} > 0$ such that, for any $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$, we have

$$\|\nabla h(B_1, A_1) - \nabla h(B_2, A_2)\| \leq L_{R'} \left\| \Big(B_1, A_1\Big) - \Big(B_2, A_2\Big) \right\| \tag{701}$$

Proof:

Let $(B_1, A_1), (B_2, A_2) \in \mathcal{D}_{R'}$ be given. Refer to the partial gradients for $h(B, A)$ in (611–612) of Appendix J. We have

$$\|\nabla h(B_1, A_1) - \nabla h(B_2, A_2)\| = \left\| \begin{pmatrix} \Delta B \,,\, \Delta A \end{pmatrix} \right\| \tag{702}$$

where $\Delta A := \nabla_A h(B_1, A_1) - \nabla_A h(B_2, A_2)$ and $\Delta B := \nabla_B h(B_1, A_1) - \nabla_B h(B_2, A_2)$.

The partial gradients with respect to B simplify to

$$\Delta B = -(W^* - B_1A_1)A_1^T + (W^* - B_2A_2)A_2^T \tag{703}$$

$$= (W^* - B_2A_2)A_2^T - (W^* - B_1A_1)A_2^T + (W^* - B_1A_1)A_2^T - (W^* - B_1A_1)A_1^T \tag{704}$$

$$= \Big((W^* - B_2A_2) - (W^* - B_1A_1)\Big)A_2^T + (W^* - B_1A_1)\left(A_2^T - A_1^T\right) \tag{705}$$

$$= (B_1A_1 - B_2A_2)\, A_2^T + (W^* - B_1A_1)\left(A_2^T - A_1^T\right) \tag{706}$$

Taking the norm, we find

$$\|\Delta B\| = \left\|(B_1A_1 - B_2A_2)\, A_2^T + (W^* - B_1A_1)\left(A_2^T - A_1^T\right)\right\| \tag{707}$$

$$\leq \left\|(B_1A_1 - B_2A_2)\, A_2^T\right\| + \left\|(W^* - B_1A_1)\left(A_2^T - A_1^T\right)\right\| \tag{708}$$

$$\leq \|(B_1A_1 - B_2A_2)\| \left\|A_2^T\right\| + \|(W^* - B_1A_1)\| \left\|\left(A_2^T - A_1^T\right)\right\| \tag{709}$$

$$\leq \|(B_1A_1 - B_2A_2)\|\, R' + \Big(\, \|W^*\| + \|B_1A_1\| \,\Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{710}$$

$$\leq \|(B_1A_1 - B_2A_2)\|\, R' + \Big(\, \|W^*\| + \|B_1\|\, \|A_1\| \,\Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{711}$$

$$\leq \|(B_1A_1 - B_2A_2)\|\, R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{712}$$

$$\leq \|(B_1A_1 - B_1A_2 + B_1A_2 - B_2A_2)\|\, R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{713}$$

$$\leq \|B_1(A_1 - A_2) + (B_1 - B_2)A_2\|\, R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{714}$$

$$\leq \Big(\, \|B_1(A_1 - A_2)\| + \|(B_1 - B_2)A_2\| \,\Big)R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{715}$$

$$\leq \Big(\, \|B_1\|\, \|(A_1 - A_2)\| + \|(B_1 - B_2)\|\, \|A_2\| \,\Big)R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{716}$$

$$\leq \Big(R'\, \|(A_1 - A_2)\| + \|(B_1 - B_2)\|\, R'\Big)R' + \Big(\, \|W^*\| + R'^2 \Big) \left\|\left(A_2^T - A_1^T\right)\right\| \tag{717}$$

$$= \Big(\, \|(A_1 - A_2)\| + \|(B_1 - B_2)\| \,\Big)R'^2 + \Big(\, \|W^*\| + R'^2 \Big)\, \|(A_1 - A_2)\| \tag{718}$$

$$= \|B_1 - B_2\|\, R'^2 + \Big(\, \|W^*\| + 2R'^2 \Big)\, \|(A_1 - A_2)\| \tag{719}$$

$$\leq \|B_1 - B_2\| \Big(\, \|W^*\| + 2R'^2 \Big) + \Big(\, \|W^*\| + 2R'^2 \Big)\, \|A_1 - A_2\| \tag{720}$$

$$= \Big(\|W^*\| + 2R'^2 \Big) \Big(\|B_1 - B_2\| + \|A_1 - A_2\| \Big) \tag{721}$$

$$\leq \sqrt{2} \Big(\|W^*\| + 2R'^2 \Big) \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big)^{1/2} \tag{722}$$

where the last line holds by Cauchy-Schwartz on $\mathbb{R}^2$*. So we have*

$$\|\Delta B\| \leq \sqrt{2} \Big(\|W^*\| + 2R'^2 \Big) \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big)^{1/2} \tag{723}$$

Denote $\frac{1}{\sqrt{2}} L_{R'} := \sqrt{2} \Big(\|W^*\| + 2R'^2 \Big)$*. Squaring both sides of our inequality, we get*

$$\|\Delta B\|^2 \leq \frac{1}{2} L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{724}$$

An identical calculation for $\|\Delta A\|$ *gives*

$$\|\Delta A\|^2 \leq \frac{1}{2} L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{725}$$

Add together (724) *and* (725) *to find*

$$\|\Delta B\|^2 + \|\Delta A\|^2 \leq L_{R'}^2 \Big(\|B_1 - B_2\|^2 + \|A_1 - A_2\|^2 \Big) \tag{726}$$

or, equivalently,

$$\|\nabla h(B_1, A_1) - \nabla h(B_2, A_2)\|^2 \leq L_{R'}^2 \, \|(B_1, A_1) - (B_2, A_2)\|^2 \tag{727}$$

Taking the square root, we finally arrive at

$$\|\nabla h(B_1, A_1) - \nabla h(B_2, A_2)\| \leq L_{R'} \, \|(B_1, A_1) - (B_2, A_2)\| \tag{728}$$

Thus, h *is* $L_{R'}$ *smooth on* $\mathcal{D}_{R'}$*.*

L Proof of Remark 2.2

Consider the continuously differentiable objective function $g : \Theta \to \mathbb{R}$, where $\Theta := \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m}$. Let $T > 0$ be finite. Assume g satisfies Lipschitz smoothness as described in Assumption 2.4. Moreover, suppose g is bounded below, or that there exists some $M \in \mathbb{R}$ such that

$$g(B, A) \geq M \quad \text{for all } (B, A) \in \Theta. \tag{729}$$

We wish to show that there exist $R'_T > 0$ and $\alpha_0 > 0$, both depending only on T, k, $\theta(0) \in \Theta$, and g, such that for every $\alpha < \alpha_0$ the iterates of Algorithm 1 satisfy

$$\max\left(\left\|B^{(\alpha)}_{ik+j}\right\|, \left\|A^{(\alpha)}_{ik+j}\right\|\right) \leq R'_T \tag{730}$$

for all $ik + j \in \mathbb{N}_0$ such that $(ik + j)\alpha \leq T$.

Before demonstrating that the *discrete iterates* themselves satisfy the inequality in (787), we first show that the *gradient flow* remains within some $\mathcal{D}_{\rho_T}$ for all $t \in [0, T]$. That is, there exists $\rho_T > 0$ depending only on T, g, and $\theta(0)$ such that the solution $\theta(t)$ to

$$\frac{d\theta(t)}{dt} = -\nabla g\Big(\theta(t)\Big) \tag{731}$$

$$\theta(0) = \left(B^{(\alpha)}_0, A^{(\alpha)}_0\right) \tag{732}$$

remains within $\mathcal{D}_{\rho_T}$ for all $t \in [0, T]$. Note by Remarks B.4-B.5 that the solution $\theta(t)$ exists given the assumptions we have placed on g.

Begin by taking a derivative of $g\left(\theta(t)\right)$ with respect to time:

$$\frac{d}{dt}\left[g\Big(\theta(t)\Big)\right] = \nabla g^\top \cdot \frac{d}{dt}\theta(t) = \nabla g^\top \cdot -\nabla g = -\left\|\nabla g\right\|^2 \tag{733}$$

Let $T > 0$ be an arbitrary finite number. Integrating on $[0, T]$, we find

$$\int_0^T \left\|\nabla g\right\|^2 \, ds = \int_T^0 -\left\|\nabla g\right\|^2 \, ds = \int_T^0 \frac{d}{ds}\left[g\Big(\theta(s)\Big)\right] \, ds = g\Big(\theta(0)\Big) - g\Big(\theta(T)\Big) \tag{734}$$

Assuming g is bounded below, we have

$$g\Big(\theta(0)\Big) - g\Big(\theta(T)\Big) \leq g\Big(\theta(0)\Big) - \inf_{\theta\in\Theta} g\Big(\theta(t)\Big) \tag{735}$$

For any $t \in [0, T]$ we can use Cauchy-Schwarz on the L_2 inner product to derive

$$\left|\int_0^t \left\|\frac{d\theta(s)}{ds}\right\| \cdot 1 \, ds\right| = \int_0^t \left\|\frac{d\theta(s)}{ds}\right\| \cdot 1 \, ds \tag{736}$$

$$\leq \sqrt{\int_0^t \left\|\frac{d\theta(s)}{ds}\right\|^2 ds} \cdot \sqrt{\int_0^t 1^2 \, ds} \tag{737}$$

$$= \sqrt{\int_0^t \frac{d\theta^\top(s)}{ds} \cdot \frac{d\theta(s)}{ds} \, ds} \cdot \sqrt{t} \tag{738}$$

$$= \sqrt{\int_0^t -\nabla g^\top \cdot \frac{d\theta(s)}{ds} \, ds} \cdot \sqrt{t} \tag{739}$$

$$= \sqrt{\int_0^t -\frac{d}{ds}\left[g\Big(\theta(s)\Big)\right] \, ds} \cdot \sqrt{t} \tag{740}$$

$$= \sqrt{g\Big(\theta(0)\Big) - g\Big(\theta(t)\Big)} \cdot \sqrt{t} \tag{741}$$

$$\leq \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] t} \tag{742}$$

By the triangle inequality then,

$$\left\| \int_0^t \frac{d\theta(s)}{ds} \cdot 1 \, ds \right\| = \|\theta(t) - \theta(0)\| \leq \int_0^t \left\| \frac{d\theta(s)}{ds} \right\| \cdot 1 \, ds \leq \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] t} \tag{743}$$

or simply

$$\|\theta(t) - \theta(0)\| \leq \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] t} \leq \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] T} \tag{744}$$

Applying the reverse triangle inequality to (744) gives

$$|\|\theta(t)\| - \|\theta(0)\|| \leq \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] T} \tag{745}$$

which implies

$$\|\theta(t)\| \leq \|\theta(0)\| + \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] T} \tag{746}$$

Thus we have that there exists ρ_T, namely

$$\rho_T \coloneqq \|\theta(0)\| + \sqrt{\left[g\Big(\theta(0)\Big) - \inf_{\theta \in \Theta} g\Big(\theta(t)\Big)\right] T} \tag{747}$$

such that $\theta(t)$ remains within $\mathcal{D}_{\rho_T}$ for all $t \in [0, T]$.

It remains to show that the *discrete iterates* themselves remain within $\mathcal{D}_{R'_T}$ for some R'_T and sufficiently small step size $\alpha < \alpha_0$, where R'_T and α_0 depend only on T, k, $\theta(0)$, and g. We do so via a strong induction argument. Begin by denoting

$$R'_T \coloneqq 2\rho_T + 2 \tag{748}$$

Base case: $ik + j = 0$

Note that

$$\max\left(\left\|B_0^{(\alpha)}\right\|, \left\|A_0^{(\alpha)}\right\|\right) \leq \left\|\left(B_0^{(\alpha)}, A_0^{(\alpha)}\right)\right\| = \|\theta(0)\| \leq \rho_T \leq R'_T - 1 \tag{749}$$

which establishes the base case.

Strong inductive step: $ik + j > 0$

Assume

$$\max\left(\left\|B_q^{(\alpha)}\right\|, \left\|A_q^{(\alpha)}\right\|\right) \leq R'_T - 1 \tag{750}$$

for all $0 < q \leq ik$ and any i such that $(i+1)k\alpha \leq T$. We will show that (750) holds for $q = ik + j$, where $0 < j \leq k$.

Stage 1 (B updates):

For $j = \{0, \ldots, k-1\}$, we have

$$B_{ik+j+1}^{(\alpha)} = B_{ik+j}^{(\alpha)} - \alpha \nabla_B g\left(B_{ik+j}^{(\alpha)}, A_{ik}^{(\alpha)}\right) \tag{751}$$

By the inductive hypothesis, $\left\|A_{ik}^{(\alpha)}\right\| \leq R_T' - 1$. Denote the upper bound for $\nabla g\left(B, A\right)$ on $\mathcal{D}_{R_T'}$ as $M_{R_T'}$, which exists by Assumption B.7. We make a subsidiary inductive argument to establish, for sufficiently small α, an upper bound on the norms of the B iterates when $0 \leq j < k$, namely that

$$\left\|B_{ik+j}^{(\alpha)}\right\| \leq R_T' - 1 + j\alpha M_{R_T'} \tag{752}$$

Assume that $\alpha \leq \dfrac{1}{kM_{R_T'}}$. The base case corresponding to $j = 0$ holds by the inductive hypothesis in (750). Now assume that $\left\|B_{ik+j}^{(\alpha)}\right\| \leq R_T' - 1 + j\alpha M_{R_T'} \leq R_T'$ for $0 < j < k$. By Remark B.4, we have that

$$\left\|\nabla_B g\left(B_{ik+j}^{(\alpha)}, A_{ik}^{(\alpha)}\right)\right\| \leq M_{R_T'} \tag{753}$$

Then we have

$$\left\|B_{ik+j+1}^{(\alpha)}\right\| \leq \left\|B_{ik+j}^{(\alpha)}\right\| + \alpha \left\|\nabla_B g\left(B_{ik+j}^{(\alpha)}, A_{ik}^{(\alpha)}\right)\right\| \tag{754}$$

$$\leq R_T' - 1 + j\alpha M_{R_T'} + \alpha M_{R_T'} \tag{755}$$

$$= R_T' - 1 + (j+1)\alpha M_{R_T'} \tag{756}$$

which proves the upper bound in (752) for step size $\alpha \leq \alpha'$, where

$$\alpha' \coloneqq \frac{1}{kM_{R_T'}}. \tag{757}$$

In particular, note that we have

$$\left\|B_{ik+j}^{(\alpha)}\right\| \leq R_T' \tag{758}$$

for all $0 \leq j \leq k$.

Stage 2 (A updates):

For $j = \{0 \dots k-1\}$, we have

$$A_{ik+j+1}^{(\alpha)} = A_{ik+j}^{(\alpha)} - \alpha \nabla_A g\left(\left[\lambda B_{ik}^{(\alpha)} + (1-\lambda) B_{(i+1)k}^{(\alpha)}\right], A_{ik+j}^{(\alpha)}\right) \tag{759}$$

Using (758), we have that

$$\left\|\lambda B_{ik}^{(\alpha)} + (1-\lambda) B_{(i+1)k}^{(\alpha)}\right\| \leq \lambda \left\|B_{ik}^{(\alpha)}\right\| + (1-\lambda)\left\|B_{(i+1)k}^{(\alpha)}\right\| \tag{760}$$

$$\leq \lambda R_T' + (1-\lambda) R_T' \tag{761}$$

$$= R_T' \tag{762}$$

We again make a subsidiary inductive argument to establish, for sufficiently small α, an upper bound on the norms of the A iterates when $0 \leq j < k$, namely that

$$\left\|A_{ik+j}^{(\alpha)}\right\| \leq R_T' - 1 + j\alpha M_{R_T'} \tag{763}$$

Assume that $\alpha \leq \dfrac{1}{kM_{R_T'}}$. The base case corresponding to $j = 0$ holds by the inductive hypothesis in (750). Now assume that $\left\|A_{ik+j}^{(\alpha)}\right\| \leq R_T' - 1 + j\alpha M_{R_T'} \leq R_T'$ for $0 < j < k$. By Remark B.4, we have that

$$\left\|\nabla_A g\left(\left[\lambda B_{ik}^{(\alpha)} + (1-\lambda) B_{(i+1)k}^{(\alpha)}\right], A_{ik+j}^{(\alpha)}\right)\right\| \leq M_{R_T'} \tag{764}$$

Then we have

$$\left\|A_{ik+j+1}^{(\alpha)}\right\| \leq \left\|A_{ik+j}^{(\alpha)}\right\| + \alpha \left\|\nabla_A g\left(\left[\lambda B_{ik}^{(\alpha)} + (1-\lambda) B_{(i+1)k}^{(\alpha)}\right], A_{ik+j}^{(\alpha)}\right)\right\| \tag{765}$$

$$\leq R_T' - 1 + j\alpha M_{R_T'} + \alpha M_{R_T'} \tag{766}$$

$$= R'_T - 1 + (j+1)\alpha M_{R'_T} \tag{767}$$

which proves the upper bound in (763) for step size $\alpha < \alpha'$, where

$$\alpha' \coloneqq \frac{1}{kM_{R'_T}}. \tag{768}$$

In particular, note that we have

$$\left\| A^{(\alpha)}_{ik+j} \right\| \leq R'_T \tag{769}$$

for all $0 \leq j \leq k$.

We thus have

$$\max\left(\left\| B^{(\alpha)}_q \right\|, \left\| A^{(\alpha)}_q \right\| \right) \leq R'_T \tag{770}$$

for all $0 \leq q \leq ik + j$, where $0 < j \leq k$. However, to complete the inductive proof, we still need to show that

$$\max\left(\left\| B^{(\alpha)}_q \right\|, \left\| A^{(\alpha)}_q \right\| \right) \leq R'_T - 1 \tag{771}$$

With Assumption B.2 satisfied on $\mathcal{D}_{R'_T}$, the derivation from Appendix B applies to g, and Eq. 215 gives

$$\left\| \theta\left(q\alpha\right) - \left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \right\| \leq G\alpha \tag{772}$$

for all $0 \leq q \leq ik + j$ such that $(i+1)k\alpha \leq T$, where $0 < j \leq k$. Note that from Appendix B we have

$$G \coloneqq \left[L_{\rho_T} M_{\rho_T} + (j+k) M_{R'_T} L_{R'_T} \right] \cdot \frac{e^{2q\alpha L_{R'_T}} - 1}{2L_{R'_T}} \tag{773}$$

where

$$L_{\rho_T} \equiv \text{ Lipschitz constant of } \nabla g\left(\theta(t)\right) \text{ on } \mathcal{D}_{\rho_T} \tag{774}$$

$$M_{\rho_T} \equiv \text{ Upper bound of } \left\| \nabla g\left(\theta(t)\right) \right\| \text{ on } \mathcal{D}_{\rho_T} \tag{775}$$

$$L_{R'_T} \equiv \text{ Lipschitz constant of } \nabla g\left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \text{ on } \mathcal{D}_{R'_T} \tag{776}$$

$$M_{R'_T} \equiv \text{ Upper bound of } \left\| \nabla g\left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \right\| \text{ on } \mathcal{D}_{R'_T} \tag{777}$$

To avoid G's dependence on the stepsize α, note that $q\alpha \leq T$ for our analysis, and we have $G \leq G_T$, where

$$G_T \coloneqq \left[L_{\rho_T} M_{\rho_T} + (j+k) M_{R'_T} L_{R'_T} \right] \cdot \frac{e^{2TL_{R'_T}} - 1}{2L_{R'_T}} \tag{778}$$

Assume that $\alpha \leq \dfrac{\rho_T + 1}{G_T}$. We then have that

$$\max\left(\left\| B^{(\alpha)}_q \right\|, \left\| A^{(\alpha)}_q \right\| \right) \leq \left\| \left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \right\| \tag{779}$$

$$= \left\| - \left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \right\| \tag{780}$$

$$= \left\| \theta(q\alpha) - \left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) - \theta(q\alpha) \right\| \tag{781}$$

$$\leq \left\| \theta(q\alpha) - \left(B^{(\alpha)}_q, A^{(\alpha)}_q \right) \right\| + \left\| \theta(q\alpha) \right\| \tag{782}$$

$$\leq G_T\alpha + \rho_T \tag{783}$$

$$\leq 2\rho_T + 1 \tag{784}$$

$$= R'_T - 1 \tag{785}$$

which finishes the proof by induction. Denote

$$\alpha_0 \coloneqq \min\left\{\frac{\rho_T + 1}{G_T}, \frac{1}{kM_{R'_T}}\right\} \tag{786}$$

We have thus shown that there exists α_0 depending only on g, $\theta(0)$, T, and k such that, whenever $\alpha < \alpha_0$, we have

$$\max\left(\left\|B^{(\alpha)}_{ik+j}\right\|, \left\|A^{(\alpha)}_{ik+j}\right\|\right) \leq R'_T = 2\rho_T + 2 \tag{787}$$

for all $ik + j \in \mathbb{N}_0$ such that $(ik + j)\alpha \leq T$. Thus, R'_T exists and itself depends only on $\theta(0)$, T, and g.

M Experiments on Error Decay Rate for Varying Stepsize α

M.1 Trace-Squared Objective

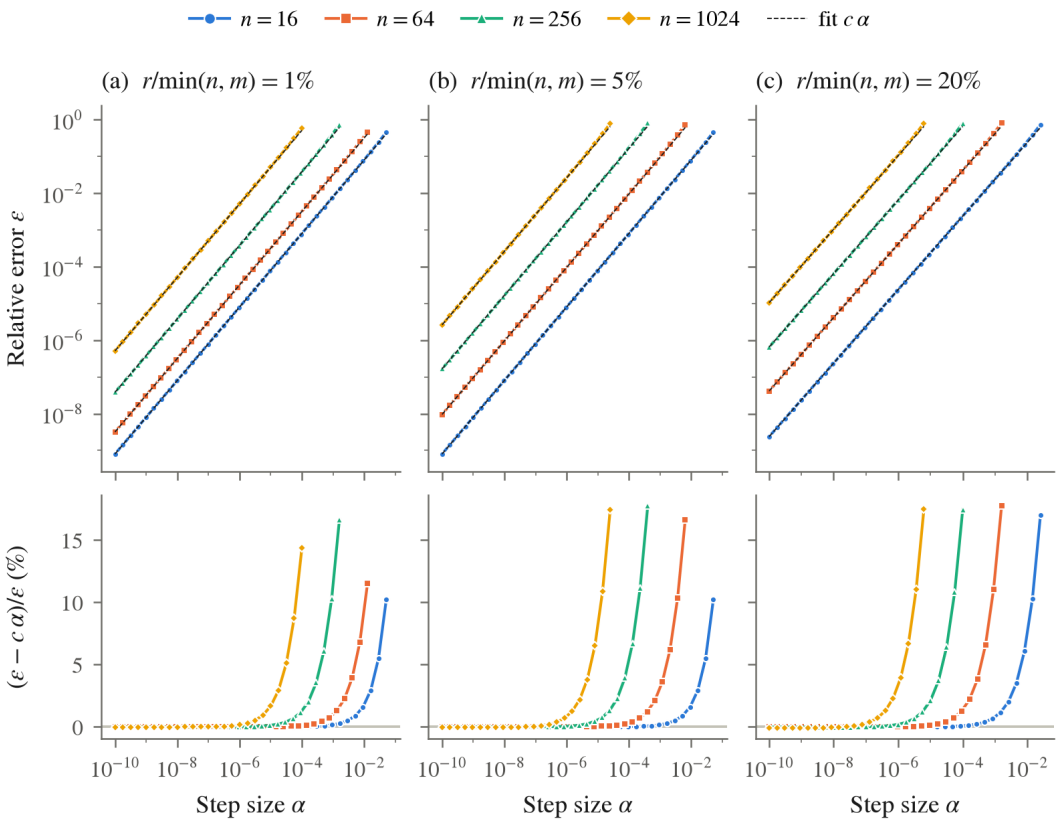

Figure 1: (Top) Relative global error between the numerical solution (trace-squared finetuned via Algorithm 1) and exact solution from the gradient flow ODEs in Eqs. (10–11) for various learning rates α. (Bottom) The relative residuals between the observed error and the theoretically predicted worst-case linear fit in α. The data plotted are taken from the mean over 500 runs with differently seeded initial conditions X_0 and weight matrix W^*. The values $k = \lambda = 1$ are used for all runs, corresponding to simultaneous LoRA parameter updates in Algorithm 1. Note the logarithmic scales on both axes.

In Theorem 2.5, the ODEs describing the exact trajectory for the LoRA iterates produced by Algorithm 1 in the limit that $\alpha \to 0$ are given by

$$\frac{dY(t)}{dt} = -\nabla_Y g\Big(Y(t), X(t)\Big) \tag{788}$$

$$\frac{dX(t)}{dt} = -\nabla_X g\Big(Y(t), X(t)\Big) \tag{789}$$

$$Y(0) = B_0, \tag{790}$$

$$X(0) = A_0 \tag{791}$$

Furthermore, the maximum (i.e. global) error at any point during finetuning between the exact trajectory given above and the discrete numerical trajectory of LoRA parameters produced by Algorithm 1 for $\alpha > 0$ decays *at least* on the order of $O(\alpha)$ for fixed k according to Theorem 2.5. That is, the global relative error is theoretically predicted to decay *at least* as fast as linearly with decreasing finetuning rate α.

To investigate whether the error truly decays linearly in α or at a faster rate $\big($ such as $O(\alpha^2)$ $\big)$, we studied this problem on the trace-squared objective given in (22):

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \frac{1}{2}\operatorname{Tr}^2(W^* - BA) \tag{792}$$

For the objective function above, Eqs. (809–812) become

$$\frac{dY(t)}{dt} = \operatorname{Tr}(W^* - YX)X^T \tag{793}$$

$$\frac{dX(t)}{dt} = \operatorname{Tr}(W^* - YX)Y^T \tag{794}$$

$$Y(0) = Y_0 \tag{795}$$

$$X(0) = X_0 \tag{796}$$

where we initialize $Y_0 = \mathbf{0}_{n\times r}$ per standard LoRA procedure Hu et al. (2022). For the system of ODEs above, we find a closed-form solution in Appendix D. We can then calculate the global relative error between the exact gradient flow solution and approximate numerical solution generated by Algorithm 1 for finite $\alpha > 0$ as the Frobenius norm difference at analogous points along the finetuning trajectory:

$$\text{Relative error } \epsilon = \max_{t\in[0,T]} \frac{\|YX(t)_{\text{exact}} - YX(t)_{\text{numerical}}\|}{\|YX(t)_{\text{exact}}\|} \tag{797}$$

According to Theorem 2.5, the relative error ϵ should decay linearly in decreasing stepsize α *or faster*.

Fig. 1 shows the results of our numerical experiment carried out to study the error decay rate in α for the trace-squared problem on varying matrix dimensions n and r. The figure shows that, for α sufficiently small, the relative error decays linearly as predicted in Theorem 2.5, suggesting a strict adherence to the $O(\alpha)$ bound in this region. However, the residual plots in Fig. 1 show residuals above ten percent between the observed error and predicted linear fit $c\alpha$ for larger values of α, which we study in Fig. 2 on the next page:

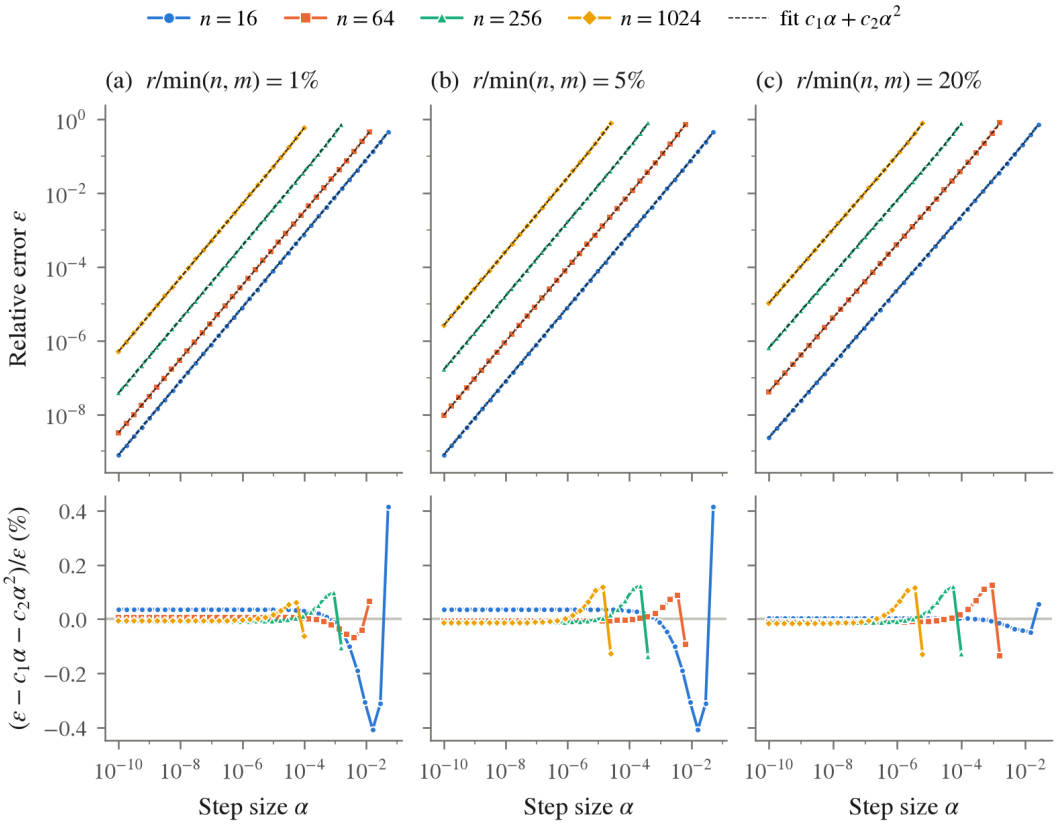

Figure 2: (Top) Relative global error between the numerical solution (trace-squared finetuned via Algorithm 1) and exact solution from the gradient flow ODEs in Eqs. (10–11) for various learning rates α. (Bottom) The relative residuals between the observed error and our hypothesized fit when including terms of $O(\alpha^2)$ in the decay rate. The data plotted are taken from the mean over 500 runs with differently seeded initial conditions X_0 and weight matrix W^*. The values $k = \lambda = 1$ are used for all runs, corresponding to simultaneous LoRA parameter updates in Algorithm 1. Note the logarithmic scales on both axes.

When we include second-order terms in the statistical fit for the relative error data, the residuals between the predicted and actual relative error drop below 0.4 percent, suggesting a second-order correction with noticeable effect at larger stepsizes. However, from the fitted model parameters, the second-order term accounts for at most 18 percent of the error in the region of stable α. We thus conclude that, for the trace-squared problem, the iterates produced by Algorithm 1 tightly follow the $O(\alpha)$ error bound given in Theorem 2.5.

M.2 Low-Rank Approximation

We studied the error decay rate in α claim from Theorem 2.5 on the L_2 norm loss function from the low-rank approximation problem in Appendix J as well:

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times m}}} \frac{1}{2}\|W^* - BA\|^2 \tag{798}$$

For the objective function above, the gradient flow ODEs giving the exact trajectory of the iterates produced by Algorithm 1 in the limit that $\alpha \to 0$ are given by

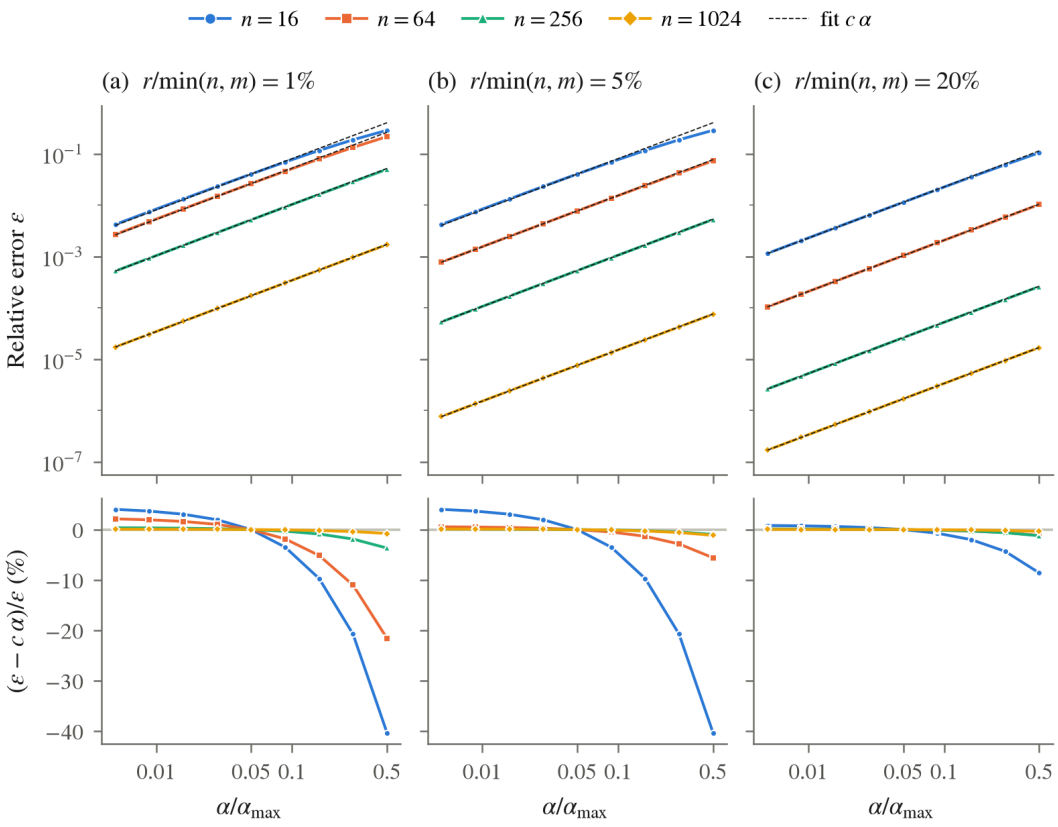

Figure 3: (Top) Relative global error between the numerical solution (low-rank approximation finetuned via Algorithm 1) and exact solution from the gradient flow ODEs in Eqs. (10–11) for various learning rates α. (Bottom) The relative residuals between the observed error and the theoretically predicted worst-case linear fit in α. The data plotted are taken from the mean over 500 runs with differently seeded initial conditions X_0 and weight matrix W^*. The values $k = \lambda = 1$ are used for all runs, corresponding to simultaneous LoRA parameter updates in Algorithm 1. Note the logarithmic scales on both axes.

$$\frac{dY(t)}{dt} = (W^* - YX)\,X^T \tag{799}$$

$$\frac{dX(t)}{dt} = Y^T\,(W^* - YX) \tag{800}$$

$$Y(0) = Y_0 \tag{801}$$

$$X(0) = X_0 \tag{802}$$

In Appendix J, we show that the ODEs above have a closed-form solution when we utilize a spectral initialization scheme which allows us to exactly solve for the trajectory of the singular values of the low-rank minimizer during finetuning. As we do for the trace-squared problem, we calculate the global relative error between the exact gradient flow solution and approximate numerical solution generated by Algorithm 1 for various stepsizes α and plot the results alongside a linear fit in Fig. 3. The plots show that the error tightly follows the linear error decay rate in α for $n \geq 256$. At larger stepsizes, we begin to observe residuals below -20 percent for $n \leq 64$, suggesting that the error grows sublinearly in α in this regime. In regimes with larger α and smaller n for this problem, the theory gives an upper bound on the error decay rate rather than a strict rule.

N Experiments on Tightness of Theoretical Rank-dependent Approximation Error Bound for Tr^2 Objective

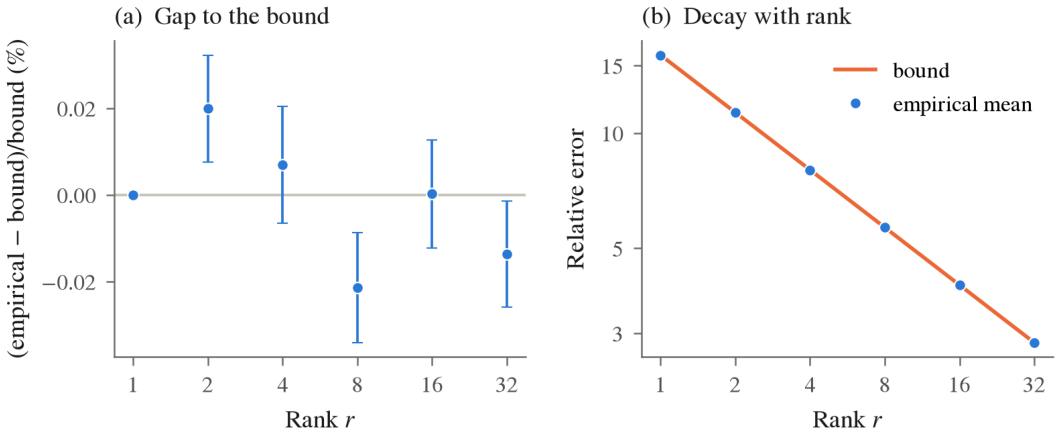

Figure 4: (Left) Residuals for the observed relative approximation error between the low-rank and full-rank solutions finetuned via gradient descent and the theoretical rank-dependent bound given in (35). The data in both plots is calculated as the mean over 1000 different seeds for $n = 256$ and r varying as indicated on the horizontal axis. Note that the error bars shown are $\pm$ 1 SEM. (Right) The observed relative approximation error between the low-rank and full-rank minimizers for six different r values is plotted along with the theoretical rank-dependent bound on the approximation error. Note that all low-rank minimizers are finetuned via Algorithm 1 with parameters $k = \lambda = 1$ (simultaneous updates).

Recall the trace-squared low-rank minimization problem studied in Eq. (22):

$$\min_{\substack{B\in\mathbb{R}^{n\times r}\\ A\in\mathbb{R}^{r\times n}}} \frac{1}{2}\,\mathrm{Tr}^2(W^* - BA) \tag{803}$$

along with its full-rank equivalent:

$$\min_{\Delta W\in\mathbb{R}^{n\times n}} \frac{1}{2}\,\mathrm{Tr}^2(W^* - \Delta W) \tag{804}$$

For a given initialization, we denote the converged low-rank minimizer

$$\lim_{t\to\infty} Y(t)X(t) \tag{805}$$

as well as the converged full-rank minimizer

$$\lim_{t\to\infty} U(t) \tag{806}$$

We give the relative *approximation error* between the converged low-rank and full-rank minimizers as

$$\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|}{\|U(t)\|} \tag{807}$$

For our initialization scheme where $Y(0) = \mathbf{0}_{n\times r}$, and the entries of $X(0)$ are drawn i.i.d. from a centered Gaussian with arbitrary variance σ^2, Theorem 2.8 gives a n, r-dependent theoretical upper bound on the expectation of (807) as

$$\mathbb{E}\left[\lim_{t\to\infty} \frac{\|Y(t)X(t) - U(t)\|}{\|U(t)\|}\right] \le \sqrt{\frac{n^2+n-2}{nr+2}} \tag{808}$$

Fig. 4 shows that the bound is tight, with normalized residuals between the theoretical upper bound and the mean observed approximation error across 1000 initialization seeds capped at $\pm$ 0.4 percent.

O EXPERIMENTS ON VARYING ERROR DECAY RATE'S DEPENDENCE ON k

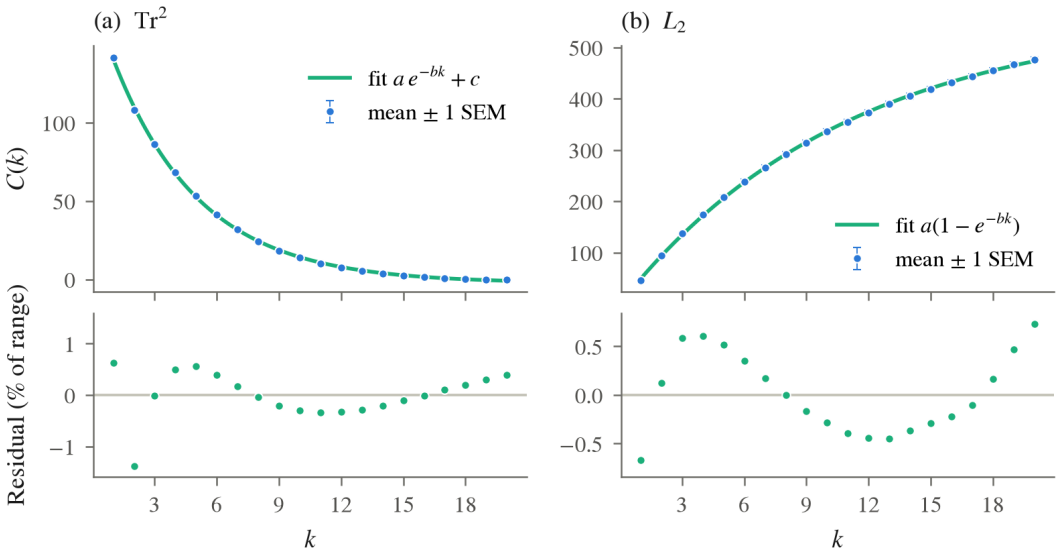

Figure 5: (a) Dependence of the first order $O(\alpha)$ error fit constant on inner step parameter k for the trace-squared objective. This plot shows how the slope $C(k)$ of the linear fit $C(k)\alpha$ of the error varies with k. (b) The same plot for the L_2 objective in the low-rank approximation function. Solid curves are the best-fitting candidate forms, $ae^{-bk} + c$ in (a) and $a(1 - e^{-bk})$ in (b); bottom panels show their residuals as a percentage of the range of $C(k)$. Note that values plotted are means over 500 runs of finetuning via Algorithm 1 with $\lambda = r = 1$ and $n = m = 256$; error bars (± 1 SEM) are smaller than the markers. For each value of k studied, the slope $C(k)$ is measured over six values of α, log-spaced over five decades below $\alpha_{\max}(k)/5$, where $\alpha_{\max}(k)$ is the empirical stability limit.

In Theorem 2.5, the ODEs describing the exact trajectory for the LoRA iterates produced by Algorithm 1 in the limit that $\alpha \to 0$ are given by

$$\frac{dY(t)}{dt} = -\nabla_Y g\Big(Y(t), X(t)\Big) \tag{809}$$

$$\frac{dX(t)}{dt} = -\nabla_X g\Big(Y(t), X(t)\Big) \tag{810}$$

$$Y(0) = B_0, \tag{811}$$

$$X(0) = A_0 \tag{812}$$

Furthermore, according to Theorem 2.5, the maximum (i.e. global) error at any point during finetuning between the exact trajectory given above and the discrete numerical trajectory of LoRA parameters produced by Algorithm 1 for $\alpha > 0$ is *at most* of order $O(k\alpha)$. That is, the global relative error decays at least linearly with decreasing α, and the slope of this linear bound grows at most linearly in the parameter k. In other words, k controls the worst-case slope of the relative error decay in α: for smaller k, the error decay in α is bounded by a line with a shallow slope, and for larger k, by a line with a steeper slope. Note that these are worst-case upper bounds on the relative error; the observed slope $C(k)$ may grow more slowly with k, or even decrease. The actual findings in Fig. 5 show that, for the two objective functions studied in this work, the slope $C(k)$ grows at most *sublinearly* with k. For the trace-squared objective, the slope of the error fit actually *decreases* with increasing k, exponentially toward zero. For the low-rank approximation problem, the slope of the error fit increases with k, but with increments that shrink approximately exponentially, consistent with $C(k)$ saturating at a positive constant. In both cases, we find that the actual slope $C(k)$ does *not* tightly follow the worst-case linear-in-k upper bound predicted by Theorem 2.5; the bound holds but is not tight in k for these objectives.