Understanding the Learning Dynamics of LoRA: A Gradient Flow Perspective on Low-Rank Adaptation in Matrix Factorization

Ziqing Xu[†] **Hancheng Min**[†] **Lachlan Ewen MacDonald**[†] **Jinqi Luo**[†]

Salma Tarmoun[†] **Enrique Mallada**[*] **René Vidal**[†]

[†]University of Pennsylvania [*]Johns Hopkins University

Abstract

Despite the empirical success of Low-Rank Adaptation (LoRA) in fine-tuning pre-trained models, there is little theoretical understanding of how first-order methods with carefully crafted initialization adapt models to new tasks. In this work, we take the first step towards bridging this gap by theoretically analyzing the learning dynamics of LoRA for matrix factorization (MF) under gradient flow (GF), emphasizing the crucial role of initialization. For small initialization, we theoretically show that GF converges to a neighborhood of the optimal solution, with smaller initialization leading to lower final error. Our analysis shows that the final error is affected by the misalignment between the singular spaces of the pre-trained model and the target matrix, and reducing the initialization scale improves alignment. To address this misalignment, we propose a spectral initialization for LoRA in MF and theoretically prove that GF with small spectral initialization converges to the fine-tuning task with arbitrary precision. Numerical experiments from MF and image classification validate our findings.

1 INTRODUCTION

Low-Rank Adaptation (Hu et al., 2022) (LoRA) has proven to be a highly effective and parameter-efficient method for fine-tuning pre-trained models, showing significant empirical success in both natural language processing (Meng et al., 2020; Liang et al., 2022; Zhang et al., 2023; Yang et al., 2024) and computer vision tasks (Zhai et al., 2022; Filatov and Kindulov, 2023). This method can be broadly described as follows: given a pre-trained model $f(x; W_1, \dots, W_L)$ parameterized by $W_1, \dots, W_L$, LoRA modifies it to $f(x; W_1 + B_1A_1, \dots, W_L + B_LA_L)$, where $W_i \in \mathbb{R}^{n_{i+1} \times n_i}$, $B_i \in \mathbb{R}^{n_{i+1} \times r_i}$, $A_i \in \mathbb{R}^{r_i \times n_i}$, and $r_i \ll \min(n_i, n_{i+1})$. In the fine-tuning stage, A_i and B_i are the trainable parameters, while W_i remains fixed, with B_i initialized as zero matrices and A_i initialized randomly.

Despite LoRA's empirical success, its theoretical underpinnings remain poorly understood. A key question is why pre-trained models can be efficiently fine-tuned with LoRA using gradient-based methods, despite the non-convex objective. Moreover, given LoRA's fine-tuning nature and specific initialization, it is essential to explore how the pre-trained model and initialization affect its learning dynamics. Existing theoretical work primarily focuses on LoRA's expressiveness (Zeng and Lee, 2023) or characterizing the optimization landscape and generalization in the Neural Tangent Kernel (NTK) regime (Malladi et al., 2023; Jang et al., 2024). Other studies (Hayou et al., 2024a,b) suggest different learning rate scales for A_i and B_i, and initializing A_i to zero while initializing B_i randomly yields better performance on average compared to the reverse. However, to the best of our knowledge, no prior work has provided a thorough theoretical analysis of LoRA's learning dynamics with explicit convergence rates or guarantees on the accuracy of the solution.

Contributions. In this paper, we study LoRA for fine-tuning a matrix factorization (MF) task via gradient flow (GF). Our analysis represents an initial step towards understanding the learning dynamics of LoRA. Our contributions are as follows:

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1. We first theoretically analyze the learning dynamics of LoRA under GF with LoRA-based small initialization, focusing on the case where the difference between the target matrices of the pre-training and fine-tuning tasks is rank-one, and assuming the pre-training MF task is perfectly solved by the pre-trained weights. Our analysis reveals two key phases: An *alignment* phase, where GF orients the singular vectors of the LoRA weights to correct the misalignment between the model and the fine-tuning task, with smaller initialization scale implying greater alignment. A *local convergence* phase, where the loss decreases linearly for a finite time until reaching a threshold determined by the initialization scale, with smaller initialization scale implying smaller final loss.

2. Motivated by the dependence of the final loss on model misalignment, we propose a spectral initialization that incorporates information from both the fine-tuning task and the pre-trained model. We theoretically prove that GF with small spectral initialization can converge to the target matrix with arbitrary precision for general MF fine-tuning tasks.

3. We validate our theoretical findings through extensive experiments on MF and several image classification tasks. In both settings, we observe that smaller scales of LoRA-based initialization lead to better alignment and lower final training error under GD. Additionally, in certain computer vision tasks, we observe improved test performance as the initialization scale decreases. While our focus is on LoRA's optimization process, these results suggest that the initialization scale may also impact generalization performance. Finally, GF with small spectral initialization in MF converges to the target matrix with arbitrary precision.

1.1 Related Work

Theory of LoRA. Zeng and Lee (2023) analyze the expressiveness of LoRA, and show that under certain assumptions, LoRA can approximate any deep linear network, multi-layer feed-forward network and transformer network. However, it is unclear whether LoRA, optimized via gradient-based algorithms, can learn these weights efficiently. Another line of work (Malladi et al., 2023; Jang et al., 2024) studies LoRA in the NTK regime. Specifically, Malladi et al. (2023) characterize the conditions under which one can study LoRA in the NTK regimes, and Jang et al. (2024) show that when the LoRA rank is $r_i \gtrsim \sqrt{N}$ where N is the number of samples, the optimization landscape of LoRA has no spurious local minima, and GD can find $\mathcal{O}(\sqrt{N})$-rank solutions that generalize well. Additional research (Hayou et al., 2024a,b) experimentally explores the effects of learning rates and initialization, recommending different learning rates for A_i and B_i, and showing that initializing B_i as zero matrices and A_i randomly improves performance compared to the reverse. However, none of these studies provide explicit convergence rates or consider the influence of pre-trained models in the optimization process, leaving gaps in the understanding of LoRA's learning dynamics.

Learning Dynamics of Low-Rank MF with Small Initialization. Our analysis of LoRA builds upon techniques developed for low-rank MF with small initialization, which falls outside the NTK regime. Specifically, Stöger and Soltanolkotabi (2021); Jin et al. (2023); Soltanolkotabi et al. (2023) analyze GD under small initialization, demonstrating the convergence and learning dynamics of GD. However, applying these techniques to LoRA in the context of MF introduces several key differences, such as distinct learning dynamics and initialization methods. We refer readers to §2.1 for a detailed discussion of these differences and the additional challenges they present.

Spectral Initialization for LoRA. Several spectral initialization methods for LoRA (Bałazy et al., 2024; Lin et al., 2024; Meng et al., 2024; Wang et al., 2024) have been proposed, based on the singular value decomposition (SVD) of pre-trained weights. Specifically, Bałazy et al. (2024); Lin et al. (2024); Meng et al. (2024) suggest initializing the left (and right) singular spaces of LoRA weights B_i (and A_i) using the top-r_i singular spaces of the pre-trained weights, ensuring that during training, B_iA_i aligns with the principal components of W_i. Conversely, Wang et al. (2024) initialize B_i and A_i using the bottom-r_i singular spaces of the pre-trained weights. Despite their differences, both approaches report better performance than standard LoRA. However, these methods overlook the fine-tuning task. In Appendix G, we provide examples where previous methods fail to fine-tune pre-trained models for MF, underscoring the importance of considering the fine-tuning data when designing spectral initialization for LoRA. In this work, we propose a spectral initialization that leverages both the pre-trained weights and the fine-tuning target matrix, demonstrating superior performance compared to standard LoRA-based initialization (see §4 for details).

1.2 Notation

We use lower case letters a to denote a scalar, and capital letters A and $A^\top$ to denote a matrix and its transpose. We use $\|A\|_F$ and $\|A\|$ to denote the Frobenius and spectral norms of A, and $A[i, j]$ to denote its

(i,j)-th element. We use $\boldsymbol{a}$ to denote a vector. For a function $f(t)$, we use $\dot{f}(t) := \frac{d}{dt}f(t)$ to denote its time derivative.

2 PRELIMINARIES

In this section, we first introduce the problem of applying LoRA to fine-tune MF tasks. Then we outline the key assumptions that underlie our analysis, addressing their implications. Furthermore, we highlight the critical distinctions between LoRA and traditional MF, and discuss the associated challenges in the analysis.

We consider applying LoRA to the classic MF. Specifically, we assume that we have a solution (W_1, W_2) to a pre-training task of factorizing a target matrix Y_{pre}:

$$(W_2, W_1) \in \arg\min_{U,V} \frac{1}{2}\|Y_{\text{pre}} - UV\|_F^2 \tag{1}$$

where $W_2 \in \mathbb{R}^{m\times h}, W_1 \in \mathbb{R}^{h\times n}$ are the *pre-trained weights*. Problem 1 covers low-rank MF (Chi et al., 2019; Koren et al., 2009) ($h \le \min(m,n)$), and over-parametrized MF (Li et al., 2018; Tarmoun et al., 2021; Min et al., 2021; Geyer et al., 2020; Eftekhari, 2022; Xu et al., 2023) ($h > \min(m,n)$).

We are interested in solving a fine-tuning task using LoRA by minimizing the following objective

$$\min_{A_1,A_2,B_1,B_2} \underbrace{\frac{1}{2}\|Y_{\text{ft}} - (W_2 + B_2A_2)(W_1 + B_1A_1)\|_F^2}_{:=L(A_1,A_2,B_1,B_2)} \tag{2}$$

where $Y_{\text{ft}} \in \mathbb{R}^{m\times n}$ is the target data matrix for this fine-tuning task, $A_1 \in \mathbb{R}^{r\times n}, B_1 \in \mathbb{R}^{h\times r}, A_2 \in \mathbb{R}^{r\times h}, B_2 \in \mathbb{R}^{m\times r}$ are *LoRA weight matrices*, and r is the *LoRA rank*. In practice, the chosen *LoRA rank* is typically much smaller than the dimensions of the pre-trained weights, satisfying $r \ll \min(m,h,n)$.

Gradient Flow as Training Algorithm. We consider solving Problem 2 using GF

$$\dot{A}_i(t) = -\frac{\partial L(t)}{\partial A_i(t)},\ \dot{B}_i(t) = -\frac{\partial L(t)}{\partial B_i(t)},\ i = 1,2 \tag{3}$$

where we use t to denote the time index in the gradient flow, and use $L(t)$ as a shorthand for $L(A_1(t), A_2(t), B_1(t), B_2(t))$. Throughout the paper, we will use A_i, B_i to denote $A_i(t), B_i(t)$ when no confusion is caused by dropping t.

LoRA-based Initialization. We initialize $A_1(0), A_2(0), B_1(0), B_2(0)$ entry-wise i.i.d. as follows: $A_1(0)[i,j], A_2(0)[i,j] \sim \mathcal{N}(0,\alpha^2), B_1(0)[i,j] = B_2(0)[i,j] = 0$. We are particularly interested in the training dynamics when α is small, or equivalently when A_1 and A_2 are initialized near zero matrices. The learning dynamics of GF/GD with weights initialized close to zero have been studied in the context of MF (Stöger and Soltanolkotabi, 2021; Jin et al., 2023; Soltanolkotabi et al., 2023) and two-layer neural networks (Bui Thi Mai and Lampert, 2021; Min et al., 2024). Throughout this paper, we use the term *small initialization* to refer to LoRA-based initialization with a small α, unless otherwise specified.

We make the following assumptions in the paper:

Assumption 2.1. *We assume that W_2, W_1 factorize Y_{pre} perfectly, i.e., $Y_{\text{pre}} = W_2W_1$. Moreover, we assume that W_2, W_1 satisfy: $W_2 = U_Y\Sigma_{W_2}G^\top, W_1 = G\Sigma_{W_1}V_Y^\top$, where G is an orthogonal matrix.*

Assumption 2.2. *Consider the SVD of $Y_{\text{pre}} = U_Y\Sigma_{\text{pre}}V_Y^\top$. Then, we assume the SVD of $\Delta Y := Y_{\text{ft}} - Y_{\text{pre}}$ has the same left and right singular matrices as Y_{pre}, i.e., $\Delta Y = U_Y\Sigma_{\Delta Y}V_Y^\top$.*

Assumption 2.1 ensures that the pre-training task is solved perfectly, and that the left and right singular matrices of W_1 and W_2, respectively, are perfectly aligned with the singular matrices of Y_{pre}. This assumption is satisfied if the pre-trained task is solved using GF with either spectral initialization (Li et al., 2019; Luo et al., 2019; Lu and Li, 2020) or balanced initialization (Bui Thi Mai and Lampert, 2021; Min et al., 2024).

Assumption 2.2 ensures certain similarities between pre-trained and fine-tuning tasks. Empirical evidence from several studies (Yosinski et al., 2014; Peters et al., 2019; Matsoukas et al., 2022) demonstrates that fine-tuning yields optimal results when there exists a substantial correlation between the pre-training and fine-tuning tasks. In the context of MF, we formally characterize this similarity as a $\text{rank}(\Delta Y)$ modification of the pre-training target matrix to the fine-tuning target matrix. Specifically, this modification occurs along the subspace spanned by a subset of singular vectors of the pre-trained target matrix, Y_{pre}, and the magnitude of this modification along the subspaces is quantified by the parameter $\Sigma_{\Delta Y}$.

2.1 Challenges in the Analysis of Learning Dynamics of Problem 2 Optimized via GF

In this section, we discuss the difference between Problem 2 and MF, and how this difference leads to additional challenges in analyzing the learning dynamics of Problem 2 optimized via GF.

Initialization of LoRA Weights Near Saddle Points. In LoRA, the weights A_1 and A_2 are randomly initialized from a Gaussian distribution, while

B_1 and B_2 are initialized as zero matrices. This contrasts with MF, where both factor matrices are typically initialized randomly. Upon inspection, it is evident that when both A_i and B_i are zero matrices, they correspond to a saddle point in the objective. Thus, the LoRA weights are initialized near these saddle points when the initialization scale α is small. It remains unclear whether GF will converge to a global optimum or become trapped in a local minimum or saddle point for Problem 2.

Complex Learning Dynamics. In low-rank MF, the objective is to approximate a target matrix by the product of two low-rank matrices. However, in LoRA, the model adopts a distinct architecture due to the influence of the pre-trained weights, as shown below:

$$\begin{aligned}&(W_2+B_2A_2)(W_1+B_1A_1)\\=&W_2B_1A_1+B_2A_2W_1+B_2A_2B_1A_1\end{aligned}\tag{4}$$

As one can see, the products of the low-rank matrices are influenced by the pre-trained weights W_2 and W_1, resulting in terms $W_2B_1A_1+B_2A_2W_1$. Moreover, the fine-tuning model includes a fourth-order term, $B_2A_2B_1A_1$. These adjustments create dependencies between the low-rank factors and the fixed pre-trained weights and introduce higher-order terms of the LoRA weights, which complicates the learning dynamics and poses new challenges for analyzing the optimization process.

The aforementioned challenges indicate that Problem 2 is more complex than classical MF. In the following sections, we address these challenges and derive convergence results for LoRA.

3 LEARNING DYNAMICS UNDER SMALL INITIALIZATION

In this section, we present our main theorem on the learning dynamics of Problem 2 trained via GF for the case where $\text{rank}(\Delta Y) = 1$. In addition, we provide an overview of the proof strategy by formally characterizing the *alignment* and *local convergence* phases, and examine how key parameters—such as pre-trained weights, initialization scale, and the target matrix—affect the learning behavior.

3.1 Main Theorem

In this section, we first introduce some notation that will be used later. Then, we present our main results in the context of $\text{rank}(\Delta Y) = 1$.

Notation. Let $Z_1 = \begin{pmatrix} B_1 \\ A_1^\top \end{pmatrix}$ and $Z_2 = \begin{pmatrix} B_2 \\ A_2^\top \end{pmatrix}$ denote the concatenation of the LoRA weights. We use $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and $\boldsymbol{U}_{\boldsymbol{Z_2}}^{\boldsymbol{S}}$ to represent the top left singular vectors of Z_1 and Z_2, respectively. When $\text{rank}(\Delta Y) = 1$, let $\Delta Y = \sigma \boldsymbol{u}\boldsymbol{v}^\top$ where $\boldsymbol{u}$ and $\boldsymbol{v}$ are a pair of singular vectors of Y_{pre} (Assumption 2.2). From Assumption 2.1, we can infer that u and v are left and right singular vectors of W_2 and W_1, respectively. Let (σ_{W_1}, g, v) and (σ_{W_2}, u, g) be the pairs of singular values and their corresponding singular vectors for W_1 and W_2. Finally, we define $\delta_w = \frac{\sigma_{W_1}}{\sigma_{W_2}}$.

Theorem 3.1. *Assume $\delta_w \neq 1$; without loss of generality take $\delta_w < 1$. Then, for any LoRA rank r, there exists constants $c_1, c_2 = \text{polylog}(\frac{1}{|1-\delta_w|}, \sigma_{W_2}, \sigma_{W_1})$, and $c_3, \alpha^* = \text{polylog}(|1-\delta_w|, \sigma_{W_2}, \sigma_{W_1})$ such that for any $0 < \alpha \leq \alpha^*$, after time $T = \frac{c_1 \log\left(\frac{1}{\alpha}\right)}{\sigma\sigma_{W_2}} + \frac{c_2 \log\left(\frac{L(0)}{\alpha}\right)}{\sigma\sigma_{W_2}}$, we have $L(T) \leq 2\alpha^{c_3}$.*

The proof of Theorem can be found in Appendix B. In Theorem 3.1, we assume without loss of generality that $\delta_w < 1$. This assumption introduces a dependency of the training time T on σ_{W_2}. Symmetrically, if one assumes $\delta_w > 1$, the training time would instead be given by:

$$T = \frac{c_1 \log\left(\frac{1}{\alpha}\right)}{\sigma\sigma_{W_1}} + \frac{c_2 \log\left(\frac{L(0)}{\alpha}\right)}{\sigma\sigma_{W_1}}. \tag{5}$$

For consistency and ease of presentation, we will assume $\delta_w < 1$ throughout the paper.

Theorem 3.1 states that after training for time T, the training loss $L(T)$ decreases to a value depending on the initialization scale α. Notice that $L(T)$ can be made arbitrarily small by selecting a sufficiently small α, and the dependence of T on α is logarithmic. Thus, reducing α leads to only a mild increase in the required training time to achieve a low training loss. Furthermore, in Figure 1, we numerically validate that, for different initialization scales α, the final loss to which GF converges decreases as α decreases.

Interpretation of Training Time T. Recall in §1, the learning dynamics can be decomposed into two phases: alignment and local convergence. As will become clear in §3.2 and §3.3, the sum decomposition of the training time T in Theorem 3.1 in fact derives from this two-phase structure:

$$T = \underbrace{\frac{c_1 \log\left(\frac{1}{\alpha}\right)}{\sigma\sigma_{W_2}}}_{\text{Phase I: alignment}} + \underbrace{\frac{c_2 \log\left(\frac{L(0)}{\alpha}\right)}{\sigma\sigma_{W_2}}}_{\text{Phase II: local convergence}}. \tag{6}$$

As one can observe, the training time required for each phase scales inversely with $\sigma\sigma_{W_2}$. In the subsequent sections (§3.2 and §3.3), we demonstrate that this is because, during the *alignment* phase, $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)$ aligns

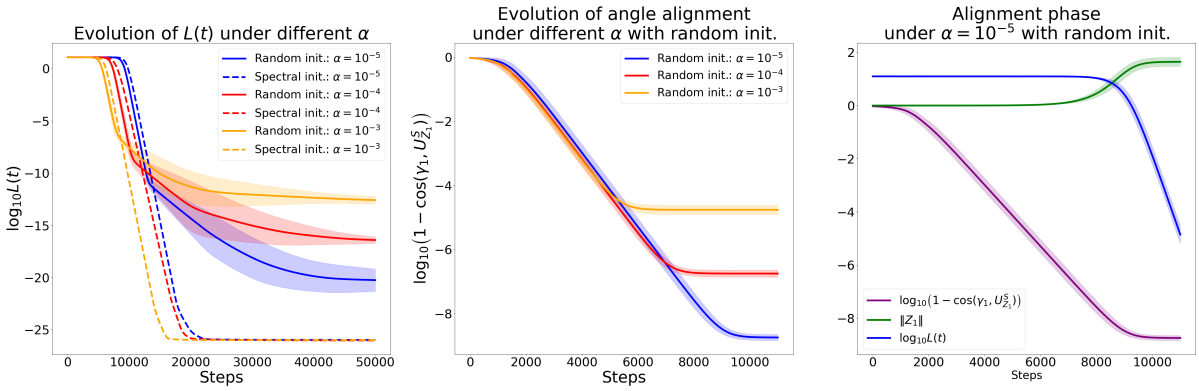

Figure 1: We simulate Problem 2 in the context of $\delta_w < 1$ using both small initialization (see §2) and small spectral initialization (see §4). Each simulation is repeated thirty times, with shaded regions representing one standard deviation above and below the mean (see §5.1 for details). The left panel shows the evolution of the loss for different initialization scales α with small and spectral initialization. The middle panel tracks the alignment quality between $\boldsymbol{U}^{\boldsymbol{S}}_{\boldsymbol{Z_1}}$ and $\boldsymbol{\gamma_1}$, measured by $\log_{10}(1-\cos(\gamma_1, \boldsymbol{U}^{\boldsymbol{S}}_{\boldsymbol{Z_1}}(t)))$, where smaller values indicate better alignment. The right panel focuses on small initialization with $\alpha = 10^{-5}$, illustrating how the reconstruction loss, alignment between $\boldsymbol{U}^{\boldsymbol{S}}_{\boldsymbol{Z_1}}$ and $\boldsymbol{\gamma_1}$, and $\|Z_1\|$ evolve during the *alignment* phase.

with $\boldsymbol{\gamma_1} = \begin{pmatrix} \boldsymbol{g} \\ \boldsymbol{v} \end{pmatrix}$, and the speed of this alignment is governed by $\sigma\sigma_{W_2}$. Moreover, smaller initialization leads to longer *alignment* phase and better quality of the alignment as illustrated in the middle panel in Figure 1. In the *local convergence* phase, we demonstrate that Problem 2 satisfies a local Polyak-Łojasiewicz (PL) condition, with the PL constant being proportional to $\sigma\sigma_{W_2}$. Consequently, the training time for each phase is inversely proportional to $\sigma\sigma_{W_2}$.

In the following sections, we present a proof sketch for Theorem 3.1 by formally characterizing the *alignment* phase and *local convergence* phase.

3.2 Proof Strategy of Alignment Phase

In this section, we argue that during the initial phase of training, GF implicitly performs spectral initialization by aligning $\boldsymbol{U}^{\boldsymbol{S}}_{\boldsymbol{Z_1}}(t)$ with $\boldsymbol{\gamma_1}$. Simultaneously, $\|Z_1\|$ continues to grow, and the difference between σ_{W_1} and σ_{W_2} leads to a progressive increase in the imbalance between $\|Z_1\|$ and $\|Z_2\|$. Moreover, the initialization scale affects the quality of alignment, with smaller initialization leading to better alignment.

.

We now present the main intuition behind the alignment phase. Our starting point is the following observation that the learning dynamics can be decomposed as follows

$$\begin{aligned} \dot{Z}_1 &= \left(\sigma\sigma_{W_2} H_1 + \begin{pmatrix} 0 & D_1 \\ D_1^\top & 0 \end{pmatrix}\right) Z_1 \,, \\ \dot{Z}_2 &= \left(\sigma\sigma_{W_1} H_2 + \begin{pmatrix} 0 & D_2 \\ D_2^\top & 0 \end{pmatrix}\right) Z_2 \,, \end{aligned} \tag{7}$$

where H_1, H_2, D_1, D_2 are defined as follows

$$\begin{aligned} &H_1 = \begin{pmatrix} 0 & \boldsymbol{g}\boldsymbol{v}^\top \\ \boldsymbol{v}\boldsymbol{g}^\top & 0 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 0 & \boldsymbol{u}\boldsymbol{g}^\top \\ \boldsymbol{g}\boldsymbol{u}^\top & 0 \end{pmatrix} \\ &F = W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1, \quad E = \Delta Y - F \\ &D_1 = A_2^\top B_2^\top E - W_2^\top F, \quad D_2 = E A_1^\top B_1^\top - F W_1^\top \,. \end{aligned} \tag{8}$$

At initialization, the magnitude of $\sigma\sigma_{W_2} H_1, \sigma\sigma_{W_1} H_2$ are significantly larger than those of D_1, D_2, since $\|D_1\|_F, \|D_2\|_F \sim \mathcal{O}(\alpha^2)$. Therefore, when α is sufficiently small, the learning dynamics shown in (7) can be approximated as follows

$$\dot{Z}_1 \approx \sigma\sigma_{W_2} H_1 Z_1, \quad \dot{Z}_2 \approx \sigma\sigma_{W_1} H_2 Z_2 \,. \tag{9}$$

Our analysis of the alignment phase centers on interpreting (7) as a *perturbation* of the approximate learning dynamics in (9). To build intuition, we summarize the simplified dynamics (9) in the following claim, deferring the perturbation analysis to Appendix C.

Claim 3.1. *Consider the following ODE system:* $\dot{Z}_1 = \sigma\sigma_{W_2} H_1 Z_1$, $\dot{Z}_2 = \sigma\sigma_{W_1} H_2 Z_2$. *Taking* Z_1 *as an example, one can first show that the eigenvalue decomposition of* H_1 *is* $H_1 = \boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top - \bar{\boldsymbol{\gamma}}_{\boldsymbol{1}}\bar{\boldsymbol{\gamma}}_{\boldsymbol{1}}^\top$, *where* $\bar{\boldsymbol{\gamma}}_{\boldsymbol{1}} = \begin{pmatrix} \boldsymbol{g} \\ -\boldsymbol{v} \end{pmatrix}$. *Based on this, the following conditions hold:*

1. *Closed-form Solution for Z_1:*

$$Z_1(t) = \exp(\sigma\sigma_{W_2}t)\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1(0) + \exp(-\sigma\sigma_{W_2}t)\bar{\boldsymbol{\gamma}}_{\mathbf{1}}\bar{\boldsymbol{\gamma}}_{\mathbf{1}}^\top Z_1(0)\,. \tag{10}$$

2. *Growth of $\|\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1\|$:* $\frac{d}{dt}\log\|\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1\|^2 = 2\sigma\sigma_{W_2}$.

3. *Imbalance Between $\|\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1\|$ and $\|\boldsymbol{\gamma_2}\boldsymbol{\gamma_2}^\top Z_2\|$:*

$$\frac{d}{dt}\log\left(\frac{\|\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1(t)\|^2}{\|\boldsymbol{\gamma_2}\boldsymbol{\gamma_2}^\top Z_2(t)\|^2}\right) = 2\sigma(\sigma_{W_2}-\sigma_{W_1})\,, \tag{11}$$

where $\boldsymbol{\gamma_2} = \begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{g}\end{pmatrix}$.

Claim 3.1 demonstrates that the simplified dynamics of Z_1 and Z_2 lead to the alignment of $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ with $\boldsymbol{\gamma_1}$, the growth of the spectrum of Z_1 projected onto the space spanned by $\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top$, and the increasing imbalance between $\|Z_1\|$ and $\|Z_2\|$. Moreover, as training progresses, $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ will be sufficiently aligned with $\boldsymbol{\gamma_1}$, leading to $\|Z_1\| \approx \|\boldsymbol{\gamma_1}\boldsymbol{\gamma_1}^\top Z_1\|$ (respectively Z_2).

LoRA Weights Escape Saddle Points. In §2.1, we discussed that at initialization, the LoRA weights are close to a saddle point where both A_i and B_i, for $i = 1$ or 2, are zero matrices. Claim 3.1 provides insights into how the LoRA weights gradually escape this saddle point during the *alignment* phase, through the growth of $\|Z_1\|$. Consequently, GF will not be trapped in this saddle point, which ensures convergence. The right panel in Figure 1 numerically demonstrates that $\|Z_1\|$ moves away from zero at the end of the *alignment* phase, following the actual dynamics of LoRA.

The simplified dynamics summarized in Claim 3.1 accurately approximate the true LoRA dynamics only up to a finite time T_1, due to the growth of perturbation terms D_1 and D_2 (see again Appendix C). The time T_1 is the end of the alignment phase, at which point the alignment and imbalance of the true dynamics are characterized by the following claim.

Claim 3.2. *Under the same setting in Theorem 3.1, there exists constants c_4, c_5 such that at T_1, one has*

1. *Good Alignment:* $\cos^2\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1), \boldsymbol{\gamma_1}\right) \geq 1 - c_4\alpha^{\frac{1+\delta_w}{3+\delta_w}}$.

2. *Sufficient Imbalance:* $\frac{\|Z_1(T_1)\|_2^2}{\|Z_2(T_1)\|_2^2} \geq c_5\alpha^{-\frac{4(1-\delta_w)}{(5-\delta_w)}}$.

The Proof of Claim 3.2 can be found in Appendix C.

3.3 Proof Strategy of Local Convergence Phase

At the end of the *alignment* phase, we established sufficient alignment between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1)$ and γ_1, as well as a significant imbalance between the norms of $\|Z_1(T_1)\|$ and $\|Z_2(T_1)\|$ (see Claim 3.2). However, it remains unclear whether these properties persist into the *local convergence* phase, or how the quality of alignment and imbalance affect the convergence results.

In this section, we first demonstrate how the large imbalance between $\|Z_1\|$ and $\|Z_2\|$ leads to simplified learning dynamics for gradient flow (GF). Specifically, with $\|Z_1\| \gg \|Z_2\|$, the objective function can be approximated as:

$$\begin{aligned} L(t) &= \frac{1}{2}\|\Delta Y - W_2A_1B_1 - A_2B_2W_1 - A_2B_2A_1B_1\|_F^2 \\ &\approx \frac{1}{2}\|\Delta Y - W_2A_1B_1\|_F^2. \end{aligned} \tag{12}$$

This approximation reduces the objective to a form similar to MF, where the factors are A_1 and B_1. Consequently, we can leverage the techniques developed in Stöger and Soltanolkotabi (2021); Jin et al. (2023); Soltanolkotabi et al. (2023) to derive the linear convergence of GF in this setting.

The following theorem formally characterizes the persistence of the alignment and the imbalance throughout the *local convergence* phase, and establishes the linear convergence rate of GF.

Theorem 3.2 (Local convergence)**.** *Under the same setting as Theorem 3.1, let $T_2 = \frac{c_2\log\left(\frac{\sqrt{2L(0)}}{\alpha}\right)}{\sigma\sigma_{W_2}}$. Then, there exists constants c_6, c_7 such that for $\forall t \in [T_1, T_1 + T_2]$, the following holds*

1. *Good Alignment of $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)$ with $\boldsymbol{\gamma_1}$:*

$$\cos^2\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t), \boldsymbol{\gamma_1}\right) \geq \cos^{2c_6}\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1), \boldsymbol{\gamma_1}\right). \tag{13}$$

2. *Imbalance Between $\|Z_1(t)\|$ and $\|Z_2(t)\|$ Persists:*

$$\frac{\|Z_1(t)\|}{\|Z_2(t)\|} \geq \left(\frac{\|Z_1(T_1)\|}{\|Z_2(T_1)\|}\right)^{c_7}. \tag{14}$$

3. *Loss Converges Linearly:*

$$\begin{aligned} &L(t) \\ \leq& 2\exp\left(-\frac{\left(1-\frac{\|Z_2(T_1)\|}{\|Z_1(T_2)\|}\right)(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{16}\right)L(T_1) \\ &+ c_8\left(1-\cos^2\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1), \boldsymbol{\gamma_1}\right)\right). \end{aligned} \tag{15}$$

Moreover, by substituting the alignment and imbalance results from Claim 3.2 and assuming $\alpha \leq \alpha^*$, we can simplify convergence rate in (15) *as follows:*

$$L(t) \leq 2\exp\left(\frac{-(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right)L(T_1) + 2\alpha^{c_3}. \tag{16}$$

The proof of Theorem 3.2 can be found in Appendix D. Theorem 3.2 establishes that the alignment and imbalance achieved during the *alignment* phase persist throughout the *local convergence* phase, and shows how these properties influence the rate of convergence. Specifically, (15) demonstrates that the upper bound on $L(t)$ consists of two components: the first part converges to zero at a linear rate, while the second part is a constant that depends on the quality of the alignment between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1)$ and $\boldsymbol{\gamma_1}$ at the end of the *alignment* phase, with better alignment leading to smaller final loss. By applying the result from Claim 3.2, we conclude that the rate of convergence is determined by the pre-trained weights and the difference between the target matrices of the pre-trained and fine-tuning tasks, i.e., $\sigma\sigma_{W_2}$. Moreover, this rate is independent of the initialization scale α.

In §3, we analyzed GF with small initialization for $\operatorname{rank}(\Delta Y)=1$, showing that smaller initialization improves alignment between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and $\boldsymbol{\gamma_1}$, leading to a lower final loss. This raises the question: if $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ is perfectly aligned with $\boldsymbol{\gamma_1}$ at initialization, can GF achieve zero loss? The following section confirms this by carefully designing spectral initialization for LoRA in MF.

4 LEARNING DYNAMICS UNDER SPECTRAL INITIALIZATION

In this section, we introduce the spectral initialization for LoRA applied to Problem 2 when LoRA rank $r \geq \operatorname{rank}(\Delta Y) > 1$. Then, we present the main theorem on its learning dynamics, and compare them to the rank-one case analyzed in §3.

Spectral Initialization. We initialize B_1, B_2 as zero matrices. For A_1, A_2, we initialize them using the singular vector matrices of $\Delta Y, W_2$. In Assumptions 2.2 and 2.1, we assume the full SVD of $\Delta Y, W_2$ are $U_Y \Sigma_{\Delta Y} V_Y^\top, U_Y \Sigma_{W_2} G^\top$ respectively. Let U_Y^S, V_Y^S be the singular matrices of ΔY corresponding to the non-zero singular values. Then, we define G^S to be the left singular vectors of W_2 corresponding to the subcolumns U_Y^S of U_Y. Based on the definition above, we initialize A_1, A_2 as follows: $A_1 = G_{11}\Sigma_{A_1} G_{12}^\top, A_2 = G_{21}\Sigma_{A_2} G_{22}^\top$, where $\Sigma_{A_1}, \Sigma_{A_2}$ are diagonal matrices and initialized entry-wise i.i.d. as $\Sigma_{A_1}[i,i], \Sigma_{A_2}[i,i] \sim \mathcal{N}(0, \alpha^2)$, and $G_{11}, G_{21} \in \mathbb{R}^{r\times r}$ are arbitrary orthogonal matrices. The matrices $G_{12} = [V_Y^S, V_{Y,\perp}^S], G_{22} = [G^S, G_\perp^S]$ are constructed from the orthogonal matrices $V_{Y,\perp}^S \in \mathbb{R}^{r\times(n-r)}, G_\perp^S \in \mathbb{R}^{r\times(h-r)}$ where $V_Y^S \perp V_{Y,\perp}^S, G^S \perp G_\perp^S$.

Remark 4.1. *Bałazy et al. (2024); Lin et al. (2024); Meng et al. (2024); Wang et al. (2024) propose spectral initialization for LoRA based solely on the pre-trained weights. In contrast, our spectral initialization depends on both the pre-trained weights and the fine-tuning target matrix. In Appendix G, we provide examples where methods built purely on pre-trained weights fail to fine-tune pre-trained models for MF, highlighting the importance of incorporating the fine-tuning target matrix when designing spectral initialization for LoRA.*

Learning Dynamics of LoRA Under Spectral Initialization. Under spectral initialization, the learning dynamics of GF can be decoupled into several scalar dynamics. Specifically, let $\bar{B}_1 = G^\top B_1 G_{11}^\top, \bar{A}_1 = G_{11}^\top A_1 G_{12}, \bar{B}_2 = U_Y^\top B_2 G_{21}^\top, \bar{A}_2 = G_{21}^\top A_2 G_{22}$, then one can write the learning dynamics of $\bar{A}_1, \bar{B}_1$ as follows

$$\begin{aligned}\frac{d}{dt}\bar{B}_1 &= (\Sigma_{W_2} + \bar{B}_2\bar{A}_2)(\Sigma_{\Delta Y} - \bar{F})\bar{A}_1^\top,\\ \frac{d}{dt}\bar{A}_1 &= \bar{B}_1^\top(\Sigma_{W_2} + \bar{B}_2\bar{A}_2)^\top(\Sigma_{\Delta Y} - \bar{F}).\end{aligned}$$

where $\bar{F} := \Sigma_{W_2}\bar{B}_1\bar{A}_1 + \bar{B}_2\bar{A}_2\Sigma_{W_1} + \bar{B}_2\bar{A}_2\bar{B}_1\bar{A}_1$. At initialization, $\bar{B}_i, \bar{A}_i, i = 1, 2$ are diagonal matrices due to spectral initialization, then they remain diagonal during the training based on (17). Thus, the learning dynamics of LoRA weights decouple into the learning dynamics of the singular values of LoRA weights. For ease of presentation in the following theorem, we will use $\sigma_{A_1}^{(i)}$ to represent the ith diagonal element of $\bar{A}_1$ (similar definition for $\sigma_{A_2}^{(i)}, \sigma_{B_1}^{(i)}, \sigma_{B_2}^{(i)}, \sigma_{\Delta Y}^{(i)}, \sigma_{W_2}^{(i)}, \sigma_{W_1}^{(i)}, \sigma_{\bar{F}}^{(i)}$).

Theorem 4.1. *Let* $\delta_w^{(i)} = \frac{\sigma_{W_2}^{(i)}}{\sigma_{W_1}^{(i)}}, \ell^{(i)}(t) = \frac{1}{2}\big(\sigma_{\Delta Y}^{(i)} - \sigma_{\bar{F}}^{(i)}\big)^2, z_1^{(i)} = (\sigma_{A_1}^{(i)})^2 + (\sigma_{B_1}^{(i)})^2$ *(respectively* $z_2^{(i)}$*). In the case where* $\delta_w^{(i)} \neq 1$*, we assume* $\delta_w^{(i)} < 1$ *WLOG, then the learning dynamics has two phases which can be separated by* $T_1^{(i)} = \frac{2}{(3+\delta_w^{(i)})\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}}\log\big(\frac{\sigma_{\Delta Y}^{(i)}}{16 z_1^{(i)}(0)}\big)$

1. *Growth of Norm and Imbalance:* $\forall t \leq T_1^{(i)}$

$$\begin{aligned}\frac{d}{dt}\log z_1^{(i)} &\geq \frac{3+\delta_w^{(i)}}{2}\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)},\\ \frac{d}{dt}\log\left(\frac{z_1^{(i)}}{z_2^{(i)}}\right) &\geq \frac{3(1-\delta_w^{(i)})}{2}\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}.\end{aligned} \tag{17}$$

2. *Local Convergence: for* $\forall t \geq T_1^{(i)}$*, the loss converges linearly*

$$\ell^{(i)}(t) \leq \exp\left(-\frac{(1-\delta_w^{(i)})\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}(t-T_1)}{8}\right)\ell^{(i)}(T_1).$$

The proof of Theorem 4.1 can be found in Appendix F.

Comparison with Small Initialization. Theorem 4.1 shows that the dynamics of each singular value

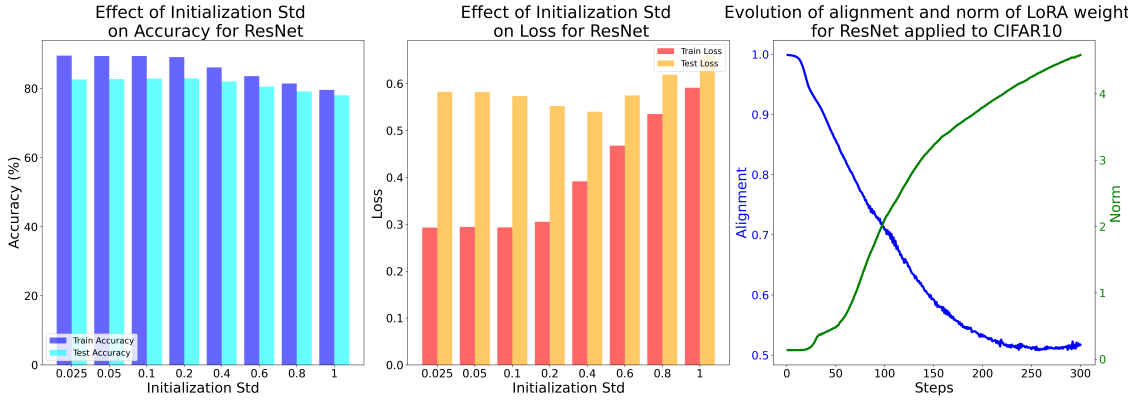

Figure 2: The left and middle panels report the loss and accuracy evaluated on the training and evaluation datasets for ResNet on the CIFAR-10 dataset. The right panel shows the evolution of the alignment between the singular matrices of the LoRA weights and the target directions (with smaller values indicating better alignment), as well as the norm of the LoRA weights. We repeat the simulation thirty times, with the shaded regions representing one standard deviation above and below the mean.

go through two distinct phases: a *growth of norm and imbalance* phase, where the norm of the singular values increases from zero, and one pair of singular values of the LoRA weights becomes dominant depending on the singular values of the pre-trained weights along the same direction, i.e., $\delta_w^{(i)}$. A *local convergence* phase, where the singular values of the model converge to the singular values of the target matrix, with the loss decreasing at a linear rate. The phenomena described above can also be observed in Theorem 3.1. Due to the spectral initialization of the LoRA weights, the model is perfectly aligned with the singular spaces of ΔY from the beginning. Thus, unlike Theorem 3.1, Theorem 4.1 does not require an *alignment* phase. A key difference of Theorem 4.1 is that the loss provably converges to arbitrary precision, whereas in Theorem 3.1, the final loss is constrained by the initialization scale (see Figure 1 for numerical validation).

Comparison with Incremental Learning in Low-rank MF. Jin et al. (2023); Pesme and Flammarion (2023) theoretically establish the incremental learning phenomenon in MF/matrix sensing problems. They show that GD/GF with small initialization gradually learns solutions with increasing ranks, with the order of learned singular values determined by their magnitudes. In Theorem 4.1, we characterize the duration of the *growth of norm and imbalance* phase and the convergence rate in the *local convergence* phase, both determined by the product of the singular values of ΔY and W_2 (or W_1, depending on $\delta_w^{(i)}$). Specifically, larger values of $\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}$ (or $\sigma_{\Delta Y}^{(i)}\sigma_{W_1}^{(i)}$, depending on $\delta_w^{(i)}$) lead to a shorter *growth of norm and imbalance* phase and faster convergence in the *local convergence* phase, indicating that the singular values of the target matrix along this direction are learned faster.

5 EXPERIMENTAL RESULTS

In this section, we run experiments on LoRA applied to MF and image classification problems to validate the theoretical findings in §3. Specifically, we are interested in the following questions: First, does decreasing the initialization scale lead to a smaller final training error? Second, does the alignment phenomenon occur in the early stages of training? In both experiments, we provide affirmative answers to these questions.

5.1 Experiments for MF

In this section, we solve Problem 2 with $\text{rank}(\Delta Y)=1$, LoRA rank $r=4$, and the factor matrix sizes are $W_2 \in \mathbb{R}^{10\times 100}$ and $W_1 \in \mathbb{R}^{100\times 10}$. We optimize the objective using GD with a small step size of 10^{-4}. We generate the pre-trained and fine-tuning target matrices as follows: $Y_{\text{pre}} \in \mathbb{R}^{10\times 10}$, where $Y_{\text{pre}}[i,j] \sim \mathcal{N}(0,1)$, and $Y_{\text{ft}} = Y_{\text{pre}} + 5u_1 v_1^\top$, where u_1 and v_1 are the top singular vectors of Y_{pre}. For the pre-trained weights, we generate them as follows: $W_2 = 1.05 \times U_Y \Sigma_{\text{pre}}^{1/2} G^\top$, and $W_1 = \frac{1}{1.05} G \Sigma_{\text{pre}}^{1/2} V_Y^\top$, where $G \in \mathbb{R}^{100\times 100}$ is an orthogonal matrix. We initialize the LoRA weights using small initialization and spectral initialization, with different initialization scales $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$.

Figure 1 numerically validates our theoretical findings from Theorem 3.1, Theorem 4.1 and Claim 3.2. The right panel shows that when the initialization scale is small, the final loss converges to a value that depends on initialization scale, with smaller scales leading to lower final loss, as predicted by Theorem 3.1.

Moreover, when using spectral initialization, GF converges to machine-precision loss regardless of the initialization scale supporting Theorem 4.1. The middle panel illustrates that smaller initialization results in a longer *alignment* phase and better alignment, consistent with Claim 3.2. Finally, the left panel demonstrates that during the *alignment* phase, $\boldsymbol{U}_{\boldsymbol{Z}_1}^{\boldsymbol{S}}$ first aligns with γ_1, followed by an increase in the norm of the LoRA weights. By the end of the *alignment* phase, sufficient alignment is achieved, and the LoRA weights move away from saddle points, where A_i and B_i are zero matrices.

5.2 Experiments for Image Classification

In this section, we fine-tune `ResNet-50` (He et al., 2016), pre-trained on ImageNet (Deng et al., 2009), for CIFAR-10 classification by adjusting the final fully-connected layer to 10 classes. LoRA is applied to this layer with initialization standard deviations ranging from 1.0 to 0.025, and training is conducted using the Adam optimizer with a learning rate of 0.01. We measure the effect of initialization scales on LoRA weights by evaluating training and validation losses and accuracies, as shown in Figure 2. Alignment measures how well LoRA's singular vectors align with target directions, with smaller values indicating better alignment. Appendix H details how target directions are determined, and includes more experiments on VGG (Simonyan and Zisserman, 2014) and `ViT-base-patch16` (Dosovitskiy et al., 2021).

The left and middle panel in Figure 2 demonstrate that smaller initialization leads to lower training and evaluation losses for ResNet. The left panel shows that in the first hundred steps of training, the singular vectors of LoRA align with the target directions, and the norm grows. The simulation results for ResNet match our theoretical findings for Problem 2. Moreover, although our theory focuses on training error of Problem 2, the improved test performance as the initialization scale decreases suggests that the initialization scale may also affect generalization performance.

6 Conclusion

This paper studied the learning dynamics of LoRA for MF under GF, highlighting the critical role of initialization. We theoretically derive convergence results for LoRA with both small initialization and spectral initialization. Our analysis reveals different behaviors under these two initializations: GF with small initialization converges to a neighborhood around the target matrix, with smaller initialization scales leading to more accurate convergence, while GF with spectral initialization converges to the target matrix with arbitrary precision. Numerical results from MF and computer vision validate our theoretical findings.

Acknowledgements

The authors acknowledge the support of the Office of Naval Research (grant 503405-78051), the National Science Foundation (grants 2031985 and 2330450), and the Simons Foundation (grant 814201).

References

Bałazy, K., Banaei, M., Aberer, K., and Tabor, J. (2024). Lora-xs: Low-rank adaptation with extremely small number of parameters. *arXiv preprint arXiv:2405.17604*.

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

Chen, Y. and Chi, Y. (2013). Spectral compressed sensing via structured matrix completion. In *International Conference on Machine Learning*, pages 414–422. PMLR.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eftekhari, A. (2022). Over-parametrized matrix factorization in the presence of spurious stationary points. *IEEE Transactions on Signal Processing*, 70:482–496.

Filatov, N. and Kindulov, M. (2023). Low rank adaptation for stable domain adaptation of vision transformers. *Optical Memory and Neural Networks*, 32(Suppl 2):S277–S283.

Geyer, K., Kyrillidis, A., and Kalev, A. (2020). Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In Chiappa, S. and

Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 930–940. PMLR.

Hayou, S., Ghosh, N., and Yu, B. (2024a). The impact of initialization on lora finetuning dynamics. *arXiv preprint arXiv:2406.08447*.

Hayou, S., Ghosh, N., and Yu, B. (2024b). Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jang, U., Lee, J. D., and Ryu, E. K. (2024). Lora training in the ntk regime has no spurious local minima. *arXiv preprint arXiv:2402.11867*.

Jin, J., Li, Z., Lyu, K., Du, S. S., and Lee, J. D. (2023). Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pages 15200–15238. PMLR.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Li, Y., Ma, C., Chen, Y., and Chi, Y. (2019). Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505. PMLR.

Li, Y., Ma, T., and Zhang, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, C., Li, L., Li, D., Zou, J., Xue, W., and Guo, Y. (2024). Nora: Nested low-rank adaptation for efficient fine-tuning large models. *arXiv preprint arXiv:2408.10280*.

Lu, Y. M. and Li, G. (2020). Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541.

Luo, W., Alghamdi, W., and Lu, Y. M. (2019). Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356.

Malladi, S., Wettig, A., Yu, D., Chen, D., and Arora, S. (2023). A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pages 23610–23641. PMLR.

Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., and Smith, K. (2022). What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9225–9234.

Meng, F., Wang, Z., and Zhang, M. (2024). Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072.

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., and Han, J. (2020). Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

Min, H., Mallada, E., and Vidal, R. (2024). Early neuron alignment in two-layer relu networks with small initialization. In *the International Conference on Learning Representations*.

Min, H., Tarmoun, S., Vidal, R., and Mallada, E. (2021). On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pages 7760–7768. PMLR.

Pesme, S. and Flammarion, N. (2023). Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 36:7475–7505.

Peters, M. E., Ruder, S., and Smith, N. A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soltanolkotabi, M., Stöger, D., and Xie, C. (2023). Implicit balancing and regularization: Generalization and convergence guarantees for overparameter-

ized asymmetric matrix sensing. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5140–5142. PMLR.

Stöger, D. and Soltanolkotabi, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843.

Tarmoun, S., Franca, G., Haeffele, B. D., and Vidal, R. (2021). Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pages 10153–10161. PMLR.

Wang, H., Li, Y., Wang, S., Chen, G., and Chen, Y. (2024). Milora: Harnessing minor singular components for parameter-efficient llm finetuning. *arXiv preprint arXiv:2406.09044*.

Xu, Z., Min, H., Tarmoun, S., Mallada, E., and Vidal, R. (2023). Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In *International Conference on Artificial Intelligence and Statistics*, pages 2262–2284. PMLR.

Yang, H., Zi, Y., Qin, H., Zheng, H., and Hu, Y. (2024). Advancing emotional analysis with large language models. *Journal of Computer Science and Software Applications*, 4(3):8–15.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Zeng, Y. and Lee, K. (2023). The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). Scaling vision transformers. *arXiv preprint arXiv:2106.04560*.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [**Yes**/No/Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Supplementary Materials

A Preliminary Lemma

In this section, we present several preliminary lemma which will be used in the following sections.

Lemma A.1. *For matrix A, B, we have*

$$\begin{aligned}\sigma_{\min}^2(A)\|B\|_F^2 \le \|AB\|_F^2 \le \sigma_{\max}^2(A)\|B\|_F^2\\ \sigma_{\min}^2(B)\|A\|_F^2 \le \|AB\|_F^2 \le \sigma_{\max}^2(B)\|A\|_F^2.\end{aligned} \tag{18}$$

Proof.

$$\begin{aligned}\|AB\|_F^2 &= \operatorname{tr}\left(ABB^\top A^\top\right)\\ &= \operatorname{tr}\left(A^\top ABB^\top\right) \qquad \text{use cyclic property of trace}\\ &\le \lambda_{\max}\left(A^\top A\right)\|B\|_F^2 \qquad \text{use trace inequality}\\ &= \sigma_{\max}^2(A)\|B\|_F^2.\end{aligned} \tag{19}$$

For the other way

$$\begin{aligned}\|AB\|_F^2 &= \operatorname{tr}\left(ABB^\top A^\top\right)\\ &= \operatorname{tr}\left(A^\top ABB^\top\right)\\ &\le \lambda_{\max}\left(BB^\top\right)\|A\|_F^2\\ &= \sigma_{\max}^2(B)\|A\|_F^2.\end{aligned} \tag{20}$$

The lower bound is similar. □

Lemma A.2. *For matrix $A \in \mathbb{R}^{m\times r}, B^{r\times n}$, we have*

$$\|AB\| \le \|A\|\cdot\|B\| \le \frac{1}{2}\left(\|A\|^2 + \|B\|^2\right) \le \frac{1}{2}\left(\|A\|_F^2 + \|A\|_F^2\right). \tag{21}$$

The proof of Lemma A.2 follows the basic property of norm and Cauchy-Swartz inequality.

Lemma A.3 (Singular Space Perturbation Bound)**.** *Let M^* and $M = M^* + E$ be two matrices in $\mathbb{R}^{m\times n}$ (WLOG, we assume $m \le n$) whose SVDs are given respectively by*

$$M^* = \begin{pmatrix}U^* & U_\perp^*\end{pmatrix}\begin{pmatrix}\Sigma^* & 0 & 0\\ 0 & \Sigma_\perp^* & 0\end{pmatrix}\begin{pmatrix}(V^*)^\top\\ (V_\perp^*)^\top\end{pmatrix}, \tag{22}$$

$$M = \begin{pmatrix}U & U_\perp\end{pmatrix}\begin{pmatrix}\Sigma & 0 & 0\\ 0 & \Sigma_\perp & 0\end{pmatrix}\begin{pmatrix}(V)^\top\\ (V_\perp)^\top\end{pmatrix}. \tag{23}$$

Here, $\sigma_1 \ge \cdots \ge \sigma_m$ (respectively $\sigma_1^* \ge \cdots \ge \sigma_m^*$), and $\Sigma, \Sigma^* \in \mathbb{R}^{r\times r}$. If $\|E\| < \sigma_r^* - \sigma_{r+1}^*$, then one has

$$\|UU^\top - U^*(U^*)^\top\| \le \frac{\sqrt{2}\max\|E\|}{\sigma_r^* - \sigma_{r+1}^* - \|E\|}. \tag{24}$$

We refer the readers to the proof of this lemma to Theorem2.9 in Chen et al. (2021).

Lemma A.4 (Theorem 7.3.3 from Horn and Johnson (2012)). *Let $A \in \mathbb{R}^{m\times n}$, let $q = \min(m,n)$, let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ be the ordered singular values of A, and define the Hemitian matrix*

$$\mathcal{A} = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}. \tag{25}$$

WLOG, assume $n \geq m$. Then, let the SVD of A is $A = V\Sigma W^*, \Sigma = [\Sigma_m, 0], V = [V_1, V_2], \hat{V}_1 = \frac{1}{\sqrt{2}}V_1, \hat{W} = \frac{1}{\sqrt{2}}W$, and $U = \begin{pmatrix} \hat{V} & -\hat{V} & V_2 \\ \hat{W} & \hat{W} & 0 \end{pmatrix}$. Then, the eigenvalue decomposition of $\mathcal{A}$ is

$$\mathcal{A} = U \begin{pmatrix} \Sigma_m & 0 & 0 \\ 0 & -\Sigma_m & 0 \\ 0 & 0 & 0 \end{pmatrix} U^*. \tag{26}$$

We refer the readers to Horn and Johnson (2012) for detailed proof.

Lemma A.5 (Learning dynamics of LoRA). *When applying GF in (3) to Problem 2, the learning dynamics of LoRA weights satisfy the following ODEs*

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} B_1 \\ A_1^\top \end{pmatrix} &= \begin{pmatrix} 0 & (W_2 + B_2A_2)^\top E \\ E^\top(W_2 + B_2A_2) & 0 \end{pmatrix}\begin{pmatrix} B_1 \\ A_1^\top \end{pmatrix}, \\ \frac{d}{dt}\begin{pmatrix} B_2 \\ A_2^\top \end{pmatrix} &= \begin{pmatrix} 0 & E(W_1 + B_1A_1)^\top \\ (W_1 + B_1A_1)E^\top & 0 \end{pmatrix}\begin{pmatrix} B_1 \\ A_1^\top \end{pmatrix} \end{aligned} \tag{27}$$

Proof. We take A_1, B_1 as an example. Direct calculations based on (3) yields the following

$$\dot{A}_1 = B_1^\top(W_2 + B_2A_2)^\top E, \quad \dot{B}_1 = (W_2 + B_2A_2)^\top EA_1^\top. \tag{28}$$

Then, one can combine the above equations to derive the result. □

Lemma A.6. *Given matrix $A \in \mathbb{R}^{m\times n}$, and use a_i to denotes its i-th column. Let $\gamma \in \mathbb{R}^m$ be any unit vector. If $\cos\left(\gamma, \frac{a_i}{\|ai\|}\right) \geq q$ holds for all $i \in [r]$, then the following holds*

$$\gamma^\top AA^\top\gamma \geq q^2\|A\|_F^2. \tag{29}$$

Proof.

$$\begin{aligned} \gamma^\top AA^\top\gamma \geq &= \gamma^\top \sum_{i=1}^r a_ia_i^\top\gamma \\ &= \sum_{i=1}^r \|a_i\|^2 \cdot \cos^2\left(\gamma, \frac{a_i}{\|a_i\|}\right) \\ &\geq q^2\sum_{i=1}^r \|a_i\|^2 \\ &= q^2\|A\|_F^2. \end{aligned} \tag{30}$$

□

B Proof of Theorem 3.1

Our proof strategy for Theorem 3.1 involves decomposing the learning dynamics into two distinct phases: the *alignment* phase and the *local convergence* phase. Theorem 3.1 builds on Claim 3.2 and Theorem 3.2, with Theorem 3.2 itself relying on Claim 3.2. In this section, we assume Claim 3.2 and Theorem 3.2 hold, and prove Theorem 3.1 based on this. Detailed proofs for Claim 3.2 and Theorem 3.2 are provided separately in Appendix C and Appendix D. We state a detailed version of Theorem 3.1 as follows.

Theorem B.1. *Assuming $\delta_w \neq 1$, for any LoRA rank r, there exists constants $c_1, c_2 = \text{polylog}(\frac{1}{|1-\delta_w|}, \sigma_{W_2}, \sigma_{W_1})$, and $c_3, \alpha^* = \text{polylog}(|1-\delta_w|, \sigma_{W_2}, \sigma_{W_1})$ such that for any $0 < \alpha \leq \alpha^*$, after time $T = \frac{c_1 \log\left(\frac{1}{\alpha}\right)}{\sigma\sigma_{W_2}} + \frac{c_2 \log\left(\frac{L(0)}{\alpha}\right)}{\sigma\sigma_{W_2}}$, we have $L(T) \leq 2\alpha^{c_3}$.*

Proof. Under the assumption that Theorem 3.2 holds, the following holds (See (16))

$$L(t) \leq \exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right) L(T_1) + \alpha^{c_3} \leq \exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right) L(0) + \alpha^{c_3}\,, \tag{31}$$

where the last inequality holds because L under GF is non-increasing. Then, we choose $t = T_2 + T_1$ such that $\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right) L(0) = \alpha^{c_3}$, or equivalently

$$T_2 = \frac{32}{(1-\delta_w)\sigma\sigma_{W_2}} \log\left(\frac{L(0)}{\alpha^{c_3}}\right), \tag{32}$$

then we have $L(T_1 + T_2) \leq 2\alpha^{c_3}$. Finally, we choose c_2 in Theorem 3.1 such that

$$c_2 = \frac{32 c_3 \log\left(\frac{L(0)}{\alpha}\right)}{(1-\delta_w)\log\left(\frac{L(0)}{\alpha}\right)} = \frac{32 c_3}{1-\delta_w}\,, \tag{33}$$

which completes the proof. □

C Proof of Claim 3.2

In this section, we provide the proof of Claim 3.2. Our proof strategy is to divide the *alignment* phase into two distinct subphases: *early alignment* phase and *escape saddle* phase. In the *early alignment*, we will show that $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)$ aligns with γ_1, and the norm of the LoRA weights stay small. In the *escape saddle* phase, we will show that $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)$ continues to align with γ_1 while the norm of the Z_1 move away from zero. Claim 3.2 can be viewed as a consequence of the *escape saddle* phase.

C.1 Proof of *early alignment* phase

In this section, we first provide a characterization of the *early alignment* phase, and its proof.

Theorem C.1 (*early alignment* phase)**.** *Under the same setting as Theorem B.1, we define*

$$\hat{T}_1 = \frac{2}{(5-\delta_w)\sigma\sigma_{W_2}} \log\left(\frac{1}{\alpha z_{\max}^2\left(\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2 + \sqrt{\|W_1\| + \|W_2\|}\right)}\right). \tag{34}$$

There exists constants $\beta_1, \beta_2, \beta_3$ that is independent of the initialization scale α such that the following holds

$$\begin{aligned} \frac{\|Z_1(\hat{T}_1)\|_F^2}{\|Z_2(\hat{T}_1)\|_F^2} &\geq \beta_1 \alpha^{-\frac{2(1-\delta_w)}{5-\delta_w}}\,, \\ \cos^2\left(\gamma_1, \boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(\hat{T}_1)\right) &\geq 1 - \beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}}\,, \\ \|Z_1(\hat{T}_1)\|_F, \|Z_2(\hat{T}_1)\|_F &\leq \beta_3 \alpha\,. \end{aligned} \tag{35}$$

Proof. Our proof for Theorem C.1 is based on the following decomposition of the learning dynamics of LoRA (which is also shown in (7)),

$$\begin{aligned} \dot{Z}_1 &= \left(\sigma\sigma_{W_2} H_1 + \hat{D}_1\right) Z_1\,, \\ \dot{Z}_2 &= \left(\sigma\sigma_{W_1} H_2 + \hat{D}_2\right) Z_2\,, \end{aligned} \tag{36}$$

where H_1, H_2, D_1, D_2 are defined as follows

$$H_1 = \begin{pmatrix} 0 & gv^\top \\ vg^\top & 0 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 0 & ug^\top \\ gu^\top & 0 \end{pmatrix}, \quad F = W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1, \quad E = \Delta Y - F,$$
$$\hat{D}_1 = \begin{pmatrix} 0 & D_1 \\ D_1^\top & 0 \end{pmatrix}, \quad \hat{D}_2 = \begin{pmatrix} 0 & D_2 \\ D_2^\top & 0 \end{pmatrix}, \quad D_1 = A_2^\top B_2^\top E - W_2^\top F, \quad D_2 = EA_1^\top B_1^\top - FW_1^\top . \tag{37}$$

Our analysis is built on the observation that during the early stage of the training, $\|\hat{D}_1\|, \|\hat{D}_2\| \sim \mathcal{O}(\alpha^2)$, which is extremely small compared with the magnitude of $\sigma\sigma_{W_1}H_2, \sigma\sigma_{W_2}H_1$. Claim 3.1 characterizes

We first characterize the learning dynamics of the angle and norm of each column of Z_1 (or equivalently Z_2) as follows

$$\frac{d}{dt}\|w_{1j}\|^2 = 2\langle \dot{w}_{1j}, w_{1j}\rangle = 2\langle w_{1j}, (\sigma\sigma_{W_2}H_1 + \hat{D}_1)w_{1j}\rangle$$
$$\frac{d}{dt}\frac{w_{1j}}{\|w_{1j}\|} = \frac{\dot{w}_{1j}}{\|w_{1j}\|} - \frac{w_{1j}}{\|w_{1j}\|^2}\cdot\frac{\dot{w}_{1j}^\top w_{1j}}{\|w_{1j}\|} = \sigma\sigma_{W_2}\left(I - \frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{H_1w_{1j}}{\|w_{1j}\|} + \left(I - \frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{\hat{D}_1w_{1j}}{\|w_{1j}\|}. \tag{38}$$

Based on the above equation, one can characterize the growth of $\|w_{1j}\|$, $\frac{\|Z_1\|_F}{\|Z_2\|_F}$ and angle alignment between w_{1j} with γ_1, as is shown the following lemma

Lemma C.1. *For any unit vector $\hat{\gamma}_1 \perp \gamma_1$, the following conditions hold*

$$\frac{d}{dt}\log(\|w_{1j}\|^2) \le 2(\sigma\sigma_{W_2} + \|D_1\|), \quad \frac{d}{dt}\log(\|Z_1\|_F^2) \le 2(\sigma\sigma_{W_2} + \|D_1\|)$$
$$\frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) \ge 2((1-\delta_w)\sigma\sigma_{W_2} - \|D_1\| - \|D_2\|),$$
$$\frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right) \ge \sigma\sigma_{W_2} - 2\|D_1\|. \tag{39}$$

We refer the readers to Appendix E.1 for the proof of Lemma C.1. Based on Lemma C.1, one can see that when $\|D_1\|, \|D_2\|$ is sufficiently small, one can control the growth of $\|w_{1j}\|, \|Z_1\|_F$ (and $\|w_{2j}\|, \|Z_2\|_F$ respectively) and $\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}, \cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})$ monotonically increase. The following lemma formally characterize these properties.

Lemma C.2. *We first define the following quantities*

$$w_{\max} = \frac{1}{\alpha}\max_{j\in[r]}\left\{\|w_{1j}\|, \|w_{2j}\|\right\}, \quad z_{\max} = \frac{1}{\alpha}\max\left\{\|Z_1(0)\|_F, \|Z_2(0)\|_F\right\}$$
$$c_{\min} = \min_{j\in[r]}\left\{\frac{\cos\left(\gamma_1, \frac{w_{1j}(0)}{\|w_{1j}(0)\|}\right)}{\cos\left(\hat{\gamma}_1, \frac{w_{1j}(0)}{\|w_{1j}(0)\|}\right)}\right\}$$
$$d_1 = \frac{\|Z_1(0)\|_F}{\|Z_2(0)\|_F}, \quad d_2 = \frac{(m+h-1)}{c_{\min}}. \tag{40}$$

Assume until $T_1(\delta_1, \delta_2)$, the following holds $\|D_1(T_1(\delta_1,\delta_2))\| \le \delta_1, \|D_2(T_1(\delta_1,\delta_2))\| \le \delta_2$. Then, the following holds for all $0 \le t \le T_1(\delta_1, \delta_2)$

$$\|w_{1j}(t)\|^2 \le \alpha^2\exp\Big(2(\sigma\sigma_{W_2} + \delta_1)t\Big)w_{\max}^2,$$
$$\|Z_1(t)\|_F^2 \le \alpha^2\exp\Big(2(\sigma\sigma_{W_2} + \delta_1)t\Big)z_{\max}^2,$$
$$\frac{\|Z_1(t)\|_F^2}{\|Z_2(t)\|_F^2} \ge d_1\exp\Big(2\big((1-\delta_w)\sigma\sigma_{W_2} - \delta_1 - \delta_2\big)t\Big),$$
$$\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) \ge \frac{1}{1 + d_2\exp\big(-2(\sigma\sigma_{W_2} - 2\delta_1)t\big)}. \tag{41}$$

We refer the readers to Appendix E.2 for the proof of Lemma C.2. Since we initialize $A_i = 0, B_i \sim \mathcal{N}(0, \alpha^2)$, therefore the value of $w_{\max}, z_{\max}, c_{\min}, d_1, d_2$ are all independent of α. Moreover, when $T_1(\delta_1, \delta_2)$ is large, one can show there will be sufficient imbalance and alignment at $T_1(\delta_1, \delta_2)$. Next, we will show that when $\delta_1 = \delta_2 = \alpha$, we can characterize a lower bound on $T_1(\delta_1, \delta_2)$ which depends on α, and it further leads to an characterization of the alignment and imbalance between $\|Z_1\|_F, \|Z_2\|_F$ evaluated at $T_1(\delta_1, \delta_2)$ depending on α.

Lemma C.3. *Under the same setting as Theorem C.1, we define $\hat{T}_1 := T_1(\alpha, \alpha)$ as follows*

$$\hat{T}_1 = \frac{2}{(5-\delta_w)\sigma\sigma_{W_2}} \log\left(\frac{1}{\alpha z_{\max}^2(\|E\|+\|W_2\|^2+\|W_1\|\|W_2\|+\sqrt{\|W_2\|})}\right). \tag{42}$$

There exists constants $\beta_1, \beta_2, \beta_3$ that is independent of the initialization scale α such that the following holds

$$\begin{aligned}
\frac{\|Z_1(\hat{T}_1)\|_F^2}{\|Z_2(\hat{T}_1)\|_F^2} &\geq d_1 d_3^{\frac{2(1-\delta_w)}{5-\delta_w}} \alpha^{-\frac{2(1-\delta_w)}{5-\delta_w}}, \\
\cos^2\left(\gamma_1, \frac{w_{1j}(\hat{T}_1)}{\|w_{1j}(\hat{T}_1)\|}\right) &\geq 1 - d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}, \\
\|Z_1(\hat{T}_1)\|_F, \|Z_2(\hat{T}_1)\|_F &\leq \frac{\alpha}{d_3},
\end{aligned} \tag{43}$$

where $d_3 = \frac{1}{z_{\max}^2(\|E\|+\|W_2\|^2+\|W_1\|\|W_2\|+\sqrt{\|W_2\|})}$.

We refer the readers to Appendix E.3 for details. Then, in the remaining of the proof of Theorem C.1, we will show that $\cos^2\left(\gamma_1, \frac{w_{1j}(\hat{T}_1)}{\|w_{1j}(\hat{T}_1)\|}\right) \geq 1 - d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}$ ensures sufficiently alignment between γ_1 and $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(\hat{T}_1)$. □

Our starting point is the observation that

$$\begin{aligned}
\|\gamma_1\gamma_1^\top Z_1(\hat{T}_1)\|_F^2 &= \sum_{j=1}^{2} \|\gamma_1\gamma_1^\top w_{1j}(\hat{T}_1)\|^2 \\
&= \sum_{j=1}^{2} \cos^2\left(\gamma_1, \frac{w_{1j}(\hat{T}_1)}{\|w_{1j}(\hat{T}_1)\|}\right) \|w_{1j}(\hat{T}_1)\|^2 \\
&\geq \left(1 - d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}\right) \|Z_1(\hat{T}_1)\|_F^2 .
\end{aligned} \tag{44}$$

and $\|(I - \gamma_1\gamma_1^\top) Z_1(\hat{T}_1)\|_F^2 \leq d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|Z_1(\hat{T}_1)\|_F^2$.

Then, we can apply Lemma A.3 (singular space perturbation bound) with $M = Z_1(\hat{T}_1), M^* = \gamma_1\gamma_1^\top Z_1(\hat{T}_1)$ and $r = 1$:

$$\begin{aligned}
\|\gamma_1\gamma_1^\top - \boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(\hat{T}_1)\big(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(\hat{T}_1)\big)^\top\| &\leq \frac{\|(I - \gamma_1\gamma_1^\top) Z_1(\hat{T}_1)\|}{\|\gamma_1\gamma_1^\top Z_1(\hat{T}_1)\| - \|(I - \gamma_1\gamma_1^\top) Z_1(\hat{T}_1)\|} \\
&\leq \frac{d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}}{1 - 2 d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}} \\
&\leq 2 d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}},
\end{aligned} \tag{45}$$

where the last inequality holds due to the assumption that $d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}} \leq \frac{1}{2}$.

C.2 Proof of *escape saddle* phase

At the end of the *early alignment* phase, we have shown sufficient alignment between each column of $Z_1(\hat{T}_1)$ (and $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(\hat{T}_1)$) with γ_1, along with a significant imbalance between $\|Z_1(\hat{T}_1)\|_F$ and $\|Z_2(\hat{T}_1)\|_F$. Moreover, the norm

of the LoRA weights remains small at this stage. In the *escape saddle* phase, we show that this alignment and imbalance persist, while the norm of the LoRA weights gradually increases until reaching a certain threshold.

The following theorem characterizes these properties formally.

Theorem C.2 (*escape saddle* phase)**.** *Under the same setting as Theorem B.1, when α is sufficiently small, let*

$$\tilde{T}_1 = \frac{2\log\left(1+\frac{(1-\delta_w^2)\sigma^2}{4\alpha^2 z_{\max}^2}\right)}{(1+\delta_w)\sigma\sigma_{W_2}}\,. \tag{46}$$

Then, there exists a time $T_1 \in [\hat{T}_1, \hat{T}_1+\tilde{T}_1]$ such that the following holds

$$\begin{aligned} g^\top B_1(T_1)A_1(T_1)v &\geq \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}\,,\\ \frac{\|Z_1(T_1)\|_F^2}{\|Z_2(T_1)\|_F^2} &\geq \beta_1\alpha^{-\frac{2(1-\delta_w)}{5-\delta_w}}\,,\\ \cos^2\left(\gamma_1, \boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1)\right) &\geq 1-\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\,. \end{aligned} \tag{47}$$

Proof. In this regime, we first show that when $\|D_1\|, \|D_2\| \leq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$, each column of Z_1 continues to align with γ_1, and $\frac{\|Z_1\|_F}{\|Z_2\|_F}$ continues to grow. This is because in Lemma C.1 and Lemma C.2, we have shown that

$$\begin{aligned} \frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) &\geq 2((1-\delta_w)\sigma\sigma_{W_2} - \|D_1\| - \|D_2\|)\,,\\ \frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right) &\geq \sigma\sigma_{W_2} - 2\|D_1\|\,. \end{aligned} \tag{48}$$

When $\|D_1\|, \|D_2\| \leq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$, we have

$$\begin{aligned} \frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) &\geq 0\,,\\ \frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right) &\geq \sigma\sigma_{W_2} - 2\|D_1\| \geq \delta\sigma\sigma_{W_2} > 0\,. \end{aligned} \tag{49}$$

Then, we characterize speed of growth of $g^\top B_1 A_1 v$.

Lemma C.4. *One can characterize the growing speed of $g^\top B_1 A_1 v$ as follows*

$$\frac{d}{dt}g^\top B_1 A_1 v \geq \left(\sigma\sigma_{W_2} - \|D_1\|\right)\|Z_1\|_F^2 \min_{j\in[r]} \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right). \tag{50}$$

The detailed proof of Lemma C.4 can be found in Appendix E.4

Based on Lemma C.4, we can show that for all $t \geq \hat{T}_1$, we have

$$\begin{aligned} \frac{d}{dt}g^\top B_1 A_1 v &\geq \frac{(1+\delta_w)\sigma\sigma_{W_2}}{2}\|Z_1\|_F^2\left(1-\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\right)\\ &\geq \alpha^2 z_{\max}^2 \exp\left(\frac{(1+\delta_w)\sigma\sigma_{W_2}t}{2}\right)\frac{(1+\delta_w)\sigma\sigma_{W_2}}{2}\left(1-\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\right)\\ \iff g^\top B_1 A_1 v &\geq \alpha^2 z_{\max}^2 \frac{(1+\delta_w)\sigma\sigma_{W_2}}{2}\left(1-\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\right)\frac{\exp\left(\frac{(1+\delta_w)\sigma\sigma_{W_2}t}{2}\right)-1}{\frac{(1+\delta_w)\sigma\sigma_{W_2}}{2}}\,. \end{aligned} \tag{51}$$

Thus, for the time for $g^\top B_1 A_1 v$ to reach $\frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}$ is upper bounded by

$$\tilde{T}_1 = \frac{2\log\left(1+\frac{(1-\delta_w^2)\sigma^2}{4\alpha^2 z_{\max}^2}\right)}{(1+\delta_w)\sigma\sigma_{W_2}}. \tag{52}$$

Notice the above $\tilde{T}_1$ is derived under the assumption that $\|D_1\|, \|D_2\| \leq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ for all $\hat{T}_1 \leq t \leq \hat{T}_1$. However, it is possible for $\|D_1\|, \|D_2\|$ to reach $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ during this time. In the following section, we assume $\|D_1\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ first at time t_1^* where $\hat{T}_1 \leq t_1^* \leq \hat{T}_1$. Then, we show how to derive a lower bound on $g^\top B_1 A_1 v$ based on this condition.

We first provide a lemma that will be used in the remaining of the proof.

Lemma C.5. *Let $\Lambda_1 = \frac{1}{\alpha^2}(B_1(0)^\top B_1(0) - A_1(0)A_1(0)^\top)$. Then, when α is sufficiently small, the following condition holds for all $t \geq 0$*

$$\frac{\|Z_1\|_F^2 - 2\alpha^2 r^2 \|\Lambda_1\|}{2r} \leq \|B_1 A_1\| \leq \frac{1}{2}\|Z_1\|_F^2. \tag{53}$$

Remark C.1. *Notice when LoRA weights are trained via GF, D_1 is constant during the training. Similar argument has been shown in Saxe et al. (2013); Tarmoun et al. (2021). Moreover, since $B_1(0)$ is initialized as a zero matrix, and $A_1(0)$ is initialized entry-wise i.i.d. using $\mathcal{N}(0,\alpha^2)$, D_1 defined as above is determined by a random matrix whose entry are $\mathcal{N}(0,1)$. Thus, when α is chosen to be small, it does not affect the magnitude of D_1.*

We refer the readers to Appendix E.5 for detailed proof of Lemma C.5.

Now, we show an lower bound on $g^\top B_1 A_1 v$ as follows

$$\begin{aligned}
\|D_1\| =& \|A_2^\top B_2^\top E - W_2^\top (W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1)\| \\
\leq& \|W_2^\top W_2 B_1 A_1\| + \|B_2 A_2\|\Big(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\|B_1 A_1\|\Big) \\
=& \|uu^\top W_2^\top W_2 B_1 A_1 vv^\top\| + \|uu^\top W_2^\top W_2 B_1 A_1 v_\perp v_\perp^\top\| + \|u_\perp u_\perp^\top W_2^\top W_2 B_1 A_1\| \\
&+ \|B_2 A_2\|\Big(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\|B_1 A_1\|\Big) \\
=& \|\sigma_{W_2}^2 u g^\top B_1 A_1 vv^\top\| + \|\sigma_{W_2}^2 u g^\top B_1 A_1 v_\perp v_\perp^\top\| + \|u_\perp u_\perp^\top W_2^\top W_2 B_1 A_1\| \\
&+ \|B_2 A_2\|\Big(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\|B_1 A_1\|\Big) \\
\leq& \sigma_{W_2}^2 g^\top B_1 A_1 v + \sigma_{W_2}^2 \|B_1 A_1 v_\perp v_\perp^\top\| + \|W_2\|^2 \|u_\perp u_\perp^\top B_1 A_1\| \\
&+ \|B_2 A_2\|\Big(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\|B_1 A_1\|\Big).
\end{aligned} \tag{54}$$

The first term on the RHS is our target quantity, we will show that the rest of the terms on the RHS is small.

For $\|B_1 A_1 v_\perp v_\perp^\top\|$, we can use the alignment between each column of Z_1 with γ_1 to lower bound it. Specifically, in the *early alignment* phase, we have proved that each column of Z_1 aligns with γ_1, i.e., $\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}) \geq 1 - \beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}}$. Moreover, in the previous discussion, we can see that this alignment persist in the *escape saddle* phase. This property ensures that each column of Z_1 aligns well with γ_1. The following lemma characterizes this property ensures $\|g_\perp g_\perp^\top B_1 A_1\|^2, \|B_1 A_1 v_\perp v_\perp^\top\|^2$ are small.

Lemma C.6. *Under the assumption that $\cos^2(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}) \geq 1 - \beta_2 \alpha^d$, then one can show that*

$$\|g_\perp g_\perp^\top B_1 A_1\|, \|B_1 A_1 v_\perp v_\perp^\top\| \leq \sqrt{\beta_2 \alpha^d}\|Z_1\|_F^2 \tag{55}$$

We refer the readers to Appendix E.6 for the proof. Based on Lemma C.6, we can show that

$$\|B_1A_1v_\perp v_\perp^\top\|^2, \|g_\perp g_\perp^\top B_1A_1\|^2 \le \beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4\,. \tag{56}$$

Based on these, we can further show that

$$\begin{aligned}
\|D_1\| &\le \sigma_{W_2}^2 g^\top B_1A_1v + \sigma_{W_2}^2\|B_1A_1v_\perp v_\perp^\top\| + \|W_2\|^2\|u_\perp u_\perp^\top B_1A_1\| \\
&\quad + \|B_2A_2\|\bigg(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\|B_1A_1\|\bigg) \\
&\le \sigma_{W_2}^2 g^\top B_1A_1v + 2\|W_2\|^2\sqrt{\beta_2}\alpha^{\frac{3+\delta_w}{10-2\delta_w}}\|Z_1\|_F^2 \\
&\quad + \frac{1}{2}\|Z_2\|_F^2\bigg(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\frac{1}{2}\|Z_1\|_F^2\bigg) \\
&\le \sigma_{W_2}^2 g^\top B_1A_1v + 2\|W_2\|^2\sqrt{\beta_2}\alpha^{\frac{3+\delta_w}{10-2\delta_w}}\|Z_1\|_F^2 \\
&\quad + \frac{1}{2\beta_1}\alpha^{\frac{2(1-\delta_w)}{5-\delta_w}}\|Z_1\|_F^2\bigg(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\frac{1}{2}\|Z_1\|_F^2\bigg)
\end{aligned} \tag{57}$$

If one can show that $\|Z_1\|_F^2$ is upper bounded, i.e., $\|Z_1\|_F^2 \le d_4$, then we can show

$$\begin{aligned}
g^\top B_1A_1v &\ge \frac{(1-\delta)\sigma}{2\sigma_{W_2}} - \frac{2\|W_2\|^2\sqrt{\beta_2}\alpha^{\frac{3+\delta_w}{10-2\delta_w}}d_4^2 + \frac{1}{2\beta_1}\alpha^{\frac{2(1-\delta_w)}{5-\delta_w}}d_4^2\bigg(\|E\| + \|W_1\|\|W_2\| + \|W_2\|\frac{1}{2}d_4^2\bigg)}{\sigma_{W_2}^2} \\
&\ge \frac{(1-\delta)\sigma}{4\sigma_{W_2}}\,,
\end{aligned} \tag{58}$$

where the last inequality holds when α is sufficiently small, and the absolute value of all the negative term are less or equal than $\frac{(1-\delta)\sigma}{4\sigma_{W_2}}$.

Finally, we show at t_1^*, $\|Z_1\|$ is upper bounded. By the assumption that $\|D_1\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ before $g^\top B_1A_1v$ reaches $\frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}$. We start with the following observation

$$\begin{aligned}
\|\gamma_1\gamma_1^\top Z_1\|_F^2 &= \sum_{j=1}^2\|\gamma_1\gamma_1^\top w_{1j}\|^2 \\
&= \sum_{j=1}^2\cos^2\bigg(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\bigg)\|w_{1j}\|^2 \\
&\ge \bigg(1 - d_2d_3^{-\frac{3+\delta_w}{5-\delta_w}}\alpha^{\frac{3+\delta_w}{5-\delta_w}}\bigg)\|Z_1\|_F^2\,.
\end{aligned} \tag{59}$$

Thus, if one can show that $\|\gamma_1\gamma_1^\top Z_1\|_F^2$ is upper bounded, then when α is small, it directly implies $\|Z_1\|_F^2$ is bounded.

$$\|\gamma_1\gamma_1^\top Z_1\|_F^2 = g^\top B_1B_1^\top g + v^\top A_1^\top A_1v + 2g^\top B_1A_1v \le g^\top B_1B_1^\top g + v^\top A_1^\top A_1v + \frac{(1-\delta_w)\sigma}{2\sigma_{W_2}}\,. \tag{60}$$

Due to the property that $A_1A_1^\top - B_1^\top B_1 = \alpha^2\Lambda_1$ is small, one can connect $g^\top B_1B_1^\top g + g^\top B_1B_1^\top g$ with $g^\top B_1A_1v$. The following Lemma characterizes this formally.

Lemma C.7.

$$\begin{aligned}
(g^\top B_1B_1^\top g)^2 &\le (v^\top A_1^\top B_1^\top g)^2 + \alpha^2\|D_1\|g^\top B_1B_1^\top g + 2\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4 \\
(v^\top A_1^\top A_1v)^2 &\le (v^\top B^\top A_1^\top g)^2 + \alpha^2\|D_1\|v^\top A_1^\top A_1v + 2\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4\,.
\end{aligned} \tag{61}$$

We refer the readers to Appendix E.7 for detailed proof of Lemma C.7. Based on Lemma C.7, one has

$$\begin{aligned}\left(v^\top A_1^\top A_1 v + g^\top B_1 B_1^\top g\right)^2 &\le 2(v^\top A_1^\top A_1 v)^2 + 2(g^\top B_1 B_1^\top g)^2 \\ &\le 2(v^\top A_1^\top B_1^\top g)^2 + \alpha^2\|D_1\|\left(v^\top A_1^\top A_1 v + g^\top B_1 B_1^\top g\right) + 4\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4 . \end{aligned} \tag{62}$$

And it leads to the following upper bound on $g^\top B_1 B_1^\top g + v^\top A_1^\top A_1 v$

$$\begin{aligned} g^\top B_1 B_1^\top g + v^\top A_1^\top A_1 v &\le \frac{\alpha^2\|D_1\| + \sqrt{\alpha^4\|D_1\|^2 + 8(g^\top B_1 A_1 v)^2 + 16\beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4}}{2} \\ &\le \alpha^2\|D_1\| + \sqrt{2} g^\top B_1 A_1 v + 2\sqrt{\beta}\|Z_1\|_F^2\alpha^{\frac{3+\delta_w}{10-2\delta_w}} \end{aligned} \tag{63}$$

Then, we plug in the above equation to (60).

$$\begin{aligned}\|\gamma_1\gamma_1^\top Z_1\|_F^2 &\le g^\top B_1 B_1^\top g + v^\top A_1^\top A_1 v + \frac{(1-\delta_w)\sigma}{2\sigma_{W_2}} \\ &\le \alpha^2\|D_1\| + 2\sqrt{\beta}\|Z_1\|^2\alpha^{\frac{3+\delta_w}{10-2\delta_w}} + \frac{(2+\sqrt{2})(1-\delta_w)\sigma}{4\sigma_{W_2}} . \end{aligned} \tag{64}$$

Combine this result with (59), we can derive the following upper bound on $\|Z_1\|$

$$\begin{aligned}\|Z_1\| &\le \frac{\alpha^2\|D_1\| + \frac{(2+\sqrt{2})(1-\delta_w)\sigma}{4\sigma_{W_2}}}{1 - d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}}\alpha^{\frac{3+\delta_w}{5-\delta_w}} - 2\sqrt{\beta}\|Z_1\|\alpha^{\frac{3+\delta_w}{10-2\delta_w}}} \\ &\le 2\|D_1\| + \frac{(2+\sqrt{2})(1-\delta_w)\sigma}{2\sigma_{W_2}} := d_4 , \end{aligned} \tag{65}$$

where the last inequality holds when α is sufficiently small.

The second case is when $\|D_2\| = \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ happens first. We argue that if one lets the training goes on, one of the following happens: either $\|D_1\| = \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ or $g^\top B_1 A_1 v = \frac{(1-\delta)\sigma}{4\sigma_{W_2}}$. Let us use time T_1' to denote the time that the above event happens. If one can show that there exists constant d such that $\|Z_1(t)\|_F^2\alpha^d \ge \|Z_2(t)\|_F^2$ holds for all $\hat{T}_1 \le t \le T_1'$, then $\|D_1\| = \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2}$ implies $g^\top B_1 A_1 v = \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}$. Moreover, it is obvious that $T_1' \le \hat{T}_1 + \hat{T}_2$. In the rest of the proof, we will show that one can actually find the constant d which is independent of α.

Our starting point is that

$$\begin{aligned}&\frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) \ge 2((1-\delta_w)\sigma\sigma_{W_2} - \|D_1\| - \|D_2\|) , \\ \Rightarrow &\log\left(\frac{\|Z_1(t)\|_F^2}{\|Z_2(t)\|_F^2}\right) - \log\left(\frac{\|Z_1(\hat{T}_1)\|_F^2}{\|Z_2(\hat{T}_1)\|_F^2}\right) \ge (1-\delta_w)\sigma\sigma_{W_2}(t-\hat{T}_1) - \int_{\hat{T}_1}^t \|D_2(s)\| ds \\ \Rightarrow &\frac{\|Z_1(t)\|_F^2}{\|Z_2(t)\|_F^2} \ge \frac{\|Z_1(\hat{T}_1)\|_F^2}{\|Z_2(\hat{T}_1)\|_F^2}\exp\big(-\int_{\hat{T}_1}^t \|D_2(s)\| ds\big) \end{aligned} \tag{66}$$

Therefore, it suffices to show that $\int_{\hat{T}_1}^{\hat{T}_1+T_1'}\|D_2(s)\| ds$ is upper bounded by a constant independent of α. To show

this, we first bound $\int_{\hat{T}_1}^{\hat{T}_1+T_1'}\|D_2(s)\|ds$ as follows

$$\begin{aligned}
\|D_2\|_F^2 =&\|EA_1^\top B_1^\top - FW_1^\top\|_F^2\\
=&\|(uu^\top + u_\perp u_\perp^\top)(EA_1^\top B_1^\top - FW_1^\top)(gg^\top + g_\perp g_\perp^\top)\|_F^2\\
\le&\|uu^\top(EA_1^\top B_1^\top - FW_1^\top)gg^\top\|_F^2\\
&+\|u_\perp u_\perp^\top(EA_1^\top B_1^\top - FW_1^\top)\|_F^2\\
&+\|uu^\top(EA_1^\top B_1^\top - FW_1^\top)g_\perp g_\perp^\top\|_F^2\\
\le&\|uu^\top(Evv^\top A_1^\top B_1^\top - FW_1^\top)gg^\top\|_F^2 + \|uu^\top Ev_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F^2\\
&+\|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F^2 + \|u_\perp u_\perp^\top FW_1^\top\|_F^2 + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F^2 + \|FW_1^\top g_\perp g_\perp^\top\|_F^2\\
\le&\|uu^\top((\Delta Y - W_2B_1A_1)vv^\top A_1^\top B_1^\top - W_2B_1A_1W_1^\top)gg^\top\|_F^2 + \|uu^\top Ev_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F^2\\
&+\|B_2A_2W_1 + B_2A_2B_1A_1\|_F^2\\
&+\|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F^2 + \|u_\perp u_\perp^\top FW_1^\top\|_F^2 + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F^2 + \|FW_1^\top g_\perp g_\perp^\top\|_F^2\\
\le&|(\sigma - \sigma_{W_2}g^\top B_1A_1v)g^\top B_1A_1v - \sigma_{W_2}\sigma_{W_1}g^\top B_1A_1v|^2 + \|uu^\top Ev_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F^2\\
&+\|B_2A_2W_1 + B_2A_2B_1A_1\|_F^2\\
&+\|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F^2 + \|u_\perp u_\perp^\top FW_1^\top\|_F^2 + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F^2 + \|FW_1^\top g_\perp g_\perp^\top\|_F^2\\
\le&\left|(\sigma - \sigma_{W_2}g^\top B_1A_1v)g^\top B_1A_1v - \sigma_{W_2}\sigma_{W_1}g^\top B_1A_1v\right|^2 + \|uu^\top Fv_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F^2\\
&+\|B_2A_2\|_F^2(\|W_1\|_F^2 + \|B_1A_1\|_F^2)\\
&+\|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F^2 + \|u_\perp u_\perp^\top FW_1^\top\|_F^2 + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F^2 + \|FW_1^\top g_\perp g_\perp^\top\|_F^2\,. \qquad (67)
\end{aligned}$$

Notice since $g^\top B_1A_1v \le \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}$, we can show that

$$\begin{aligned}
\left|(\sigma - \sigma_{W_2}g^\top B_1A_1v)g^\top B_1A_1v - \sigma_{W_2}\sigma_{W_1}g^\top B_1A_1v\right| &\le (\sigma - \sigma_{W_2}g^\top B_1A_1v)g^\top B_1A_1v + \sigma_{W_2}\sigma_{W_1}g^\top B_1A_1v\\
&\le \left(\frac{(3+\delta_w)\sigma}{4} + \sigma_{W_2}\sigma_{W_1}\right)g^\top B_1A_1v\,. \qquad (68)
\end{aligned}$$

Thus, we derive the following upper bound on $\|D_1\|$

$$\begin{aligned}
\|D_1\| \le \|D_1\|_F \le&\left(\frac{(3+\delta_w)\sigma}{4} + \sigma_{W_2}\sigma_{W_1}\right)g^\top B_1A_1v\\
&+\|uu^\top Fv_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F + \|B_2A_2\|_F(\|W_1\|_F + \|B_1A_1\|_F)\\
&+\|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F + \|u_\perp u_\perp^\top FW_1^\top\|_F + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F + \|FW_1^\top g_\perp g_\perp^\top\|_F\,. \qquad (69)
\end{aligned}$$

Notice except for the first term, the rest of the term either depends on B_2A_2 or $g_\perp g_\perp^\top B_1A_1$ or $B_1A_1v_\perp v_\perp^\top$. Based on Lemma C.6 and the conditon that $\|Z_1(t)\|_F^2\alpha^d \ge \|Z_2(t)\|_F^2$, we can conclude that the rest of the term is extremely small compared with the first term when α is sufficiently small which is formally characterized by the following lemma.

Lemma C.8. *Under the following conditions*

$$\begin{aligned}
\|Z_1(t)\|_F^2\alpha^d &\ge \|Z_2(t)\|_F^2\\
\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) &\ge 1 - d_2d_3^{-\frac{3+\delta_w}{5-\delta_w}}\alpha^{\frac{3+\delta_w}{5-\delta_w}}\,,\\
g^\top B_1A_1v &\le \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}} \qquad (70)
\end{aligned}$$

we can show that

$$\begin{aligned}&\|uu^\top Fv_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F + \|B_2A_2\|_F(\|W_1\|_F + \|B_1A_1\|_F)\\&+ \|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F + \|u_\perp u_\perp^\top FW_1^\top\|_F + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F + \|FW_1^\top g_\perp g_\perp^\top\|_F\\&\le \left(\frac{(3+\delta_w)\sigma}{4} + \sigma_{W_2}\sigma_{W_1}\right) g^\top B_1A_1v\,.\end{aligned} \tag{71}$$

We refer the readers to Appendix E.8 for the proof of Lemma C.8. Thus, based on Lemma C.8 one can conclude that

$$\|D_2\|_F \le \left(\frac{(3+\delta_w)\sigma}{2} + 2\sigma_{W_2}\sigma_{W_1}\right) g^\top B_1A_1v\,. \tag{72}$$

Thus, $\int_{\hat{T}_1}^{\hat{T}_1+T_1'}\|D_2(s)\|ds \le \left(\frac{(3+\delta_w)\sigma}{2} + 2\sigma_{W_2}\sigma_{W_1}\right)\int_{\hat{T}_1}^{\hat{T}_1+T_1'} g^\top B_1(s)A_1(s)vds$.

Therefore, it suffices to show that $\int_{\hat{T}_1}^{\hat{T}_1+T_1'} g^\top B_1(s)A_1(s)vds$ is bounded.

$$\begin{aligned}\frac{d}{dt}g^\top B_1A_1v &= g^\top \dot{B}_1A_1v + g^\top B_1\dot{A}_1v\\&= g^\top\big(\sigma\sigma_{W_2}gv^\top + D_1\big)A_1^\top A_1v + g^\top B_1B_1^\top\big(\sigma\sigma_{W_2}gv^\top + D_1\big)v\\&= \sigma\sigma_{W_2}(g^\top B_1B_1^\top g + v^\top A_1^\top A_1v) + g^\top D_1A_1^\top A_1v + g^\top B_1B_1^\top D_1^\top v\\&\ge \frac{(1+\delta_w)\sigma\sigma_{W_2}}{2}(g^\top B_1B_1^\top g + v^\top A_1^\top A_1^\top v)\\&\ge (1+\delta_w)\sigma\sigma_{W_2}g^\top B_1A_1v\end{aligned} \tag{73}$$

where in the last inequality, we use Cauchy Swartz. One directly has

$$\begin{aligned}&g^\top A_1(T_1')B_1(T_1')v \ge g^\top A_1(\hat{T}_1)B_1(\hat{T}_1)v \ge (1+\delta_w)\sigma\sigma_{W_2}\int_{\hat{T}_1}^{T_1'+\hat{T}_1} g^\top B_1(s)A_1(s)vds\\&\Rightarrow \int_{\hat{T}_1}^{T_1'+\hat{T}_1} g^\top B_1(s)A_1(s)vds \le g^\top A_1(T_1')B_1(T_1')v = \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}}\,,\end{aligned} \tag{74}$$

which completes the proof. □

D Proof of Theorem 3.2

In this section, we provide proof of Theorem 3.2. In Appendix B, we have shown that at the end of *alignment* phase, there is sufficient alignment and imbalance, and $g^\top B_1A_1v$ has reached a constant order which is independent of α. In this section, we will show that how these properties lead to linear convergence of the loss until it reaches a neighourhood around the global minimum. We first present a detailed version of Theorem 3.2.

Theorem D.1. *Under the same setting as Theorem 3.1, let* $T_2 = \frac{c_2\log\left(\frac{\sqrt{2L(0)}}{\alpha}\right)}{\sigma\sigma_{W_2}}$. *Then, there exists constants* c_6, c_7 *such that for* $\forall t \in [T_1, T_1+T_2]$, *the following holds*

1. Good Alignment of $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)$ *with* γ_1*:*

$$\cos^2\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t), \gamma_1\right) \ge \cos^{2c_6}\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1), \gamma_1\right). \tag{75}$$

2. Imbalance Between $\|Z_1(t)\|$ *and* $\|Z_2(t)\|$ *Persists:*

$$\frac{\|Z_1(t)\|}{\|Z_2(t)\|} \ge \left(\frac{\|Z_1(T_1)\|}{\|Z_2(T_1)\|}\right)^{c_7}. \tag{76}$$

3. Loss Converges Linearly:

$$L(t) \leq 2\exp\left(-\frac{\left(1-\frac{\|Z_2(T_1)\|_F^2}{\|Z_1(T_2)\|_F^2}\right)(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{16}\right)L(T_1) + c_8\left(1-\cos^2\left(\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(T_1),\gamma_1\right)\right). \tag{77}$$

Moreover, by substituting the alignment and imbalance results from Claim 3.2 and assuming $\alpha \leq \alpha^*$, we can simplify convergence rate in (15) *as follows:*

$$L(t) \leq 2\exp\left(\frac{-(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right)L(T_1) + 2\alpha^{c_3}. \tag{78}$$

Proof. We start with a decomposition of the loss into *signal* and *noise* part.

$$\begin{aligned} L =& \frac{1}{2}\|Y_{\text{ft}} - (W_2+B_2A_2)(W_1+B_1A_1)\|_F^2 \\ =& \frac{1}{2}\|(uu^\top + u_\perp u_\perp^\top)\big(Y_{\text{ft}} - (W_2+B_2A_2)(W_1+B_1A_1)\big)(vv^\top + v_\perp v_\perp^\top)\|_F^2 \\ =& \underbrace{\frac{1}{2}\|\sigma uv^\top - \sigma_{W_2} g^\top B_1A_1v - \sigma_{W_2}u^\top B_2A_2g - uu^\top B_2A_2B_1A_1vv^\top\|_F^2}_{\text{Signal loss: } L_S}, + \underbrace{\frac{1}{2}\|u_\perp u_\perp^\top F + uu^\top F v_\perp v_\perp^\top\|_F^2}_{\text{Noise loss: } L_N} \end{aligned} \tag{79}$$

where $F = W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1$. Our proof strategy it to show that L_S converges linearly until it reaches a small value depending on α while L_N remains small. Moreover, we will show that the alignment and imbalance we achieve in the *alignment* phase remains good in the *local convergence* phase.

Bounds on several key quantities. We first assume that for any $T_1 \leq t \leq T_2$, there is sufficient alignment between each column of Z_1 with γ_1, and sufficient imbalance between $\|Z_1\|, \|Z_2\|$. Moreover, the norm of Z_1 is bounded.

$$\|Z_1\|_F \leq d_5, \quad \|Z_1\|_F^2 \geq \alpha^{-d_6}\|Z_2\|_F^2, \quad \min_{j\in[r]} \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) \geq 1-\alpha^{d_7}, \tag{80}$$

where d_5, d_6, d_7 are constants only depending on $W_1, W_2, \Delta Y$ and independent of α. We will derive convergence results based on these constants. In the end of this section, we will provide expressions for d_5, d_6, d_7.

Bound on the noise loss L_N.

$$\begin{aligned} L_N =& \frac{1}{2}\|u_\perp u_\perp^\top F + uu^\top F v_\perp v_\perp^\top\|_F^2 \\ \leq& \frac{1}{2}\|u_\perp u_\perp^\top F\|_F^2 + \frac{1}{2}\|uu^\top F v_\perp v_\perp^\top\|_F^2 \\ \leq& \frac{1}{2}\|u_\perp u_\perp^\top W_2B_1A_1\|_F^2 + \frac{1}{2}\|u_\perp u_\perp^\top B_2A_2W_1\|_F^2 + \frac{1}{2}\|u_\perp u_\perp^\top B_2A_2B_1A_1\|_F^2 \\ & + \frac{1}{2}\|uu^\top W_2B_1A_1v_\perp v_\perp^\top\|_F^2 + \frac{1}{2}\|uu^\top B_2A_2W_1v_\perp v_\perp^\top\|_F^2 + \frac{1}{2}\|uu^\top B_2A_2B_1A_1v_\perp v_\perp^\top\|_F^2 \\ \leq& \frac{1}{2}\|u_\perp u_\perp^\top W_2B_1A_1\|_F^2 + \frac{1}{2}\|uu^\top W_2B_1A_1v_\perp v_\perp^\top\|_F^2 + \left(\|W_1\|^2 + \|B_1A_1\|_F^2\right)\|B_2A_2\|_F^2 \\ \leq& \frac{1}{2}\|W_2\|^2\left(\|g_\perp g_\perp^\top B_1A_1\|_F^2 + \|B_1A_1v_\perp v_\perp^\top\|_F^2\right) + \frac{1}{2}\|Z_2\|_F^2\left(\|W_1\|^2 + \frac{1}{2}\|Z_1\|_F^2\right) \\ \leq& r^2\|W_2\|^2\sqrt{\alpha^{d_7}}d_5^4 + \left(\|W_2\|^2 + \frac{d_5^2}{2}\right)\times\frac{1}{2}\alpha^{d_6}d_5^2, \end{aligned} \tag{81}$$

The above upper bound on L_N demonstrates that the noise loss is small if the initialization scale α is small.

Convergence on the signal loss L_S. For the convergence of L_S, we study $\dot{L}_S$ first.

$$\begin{aligned}\dot{L}_S &= \sum_{i=1}^{2}\langle \frac{L_S}{\partial A_i}, \dot{A}_i\rangle + \langle \frac{L_S}{\partial B_i}, \dot{B}_i\rangle \\ &= -\sum_{i=1}^{2}\langle \frac{L_S}{\partial A_i}, \frac{L}{\partial A_i}\rangle + \langle \frac{L_S}{\partial B_i}, \frac{L}{\partial B_i}\rangle \\ &= -\sum_{i=1}^{2}\|\frac{L_S}{\partial A_i}\|_F^2 + \|\frac{L_S}{\partial B_i}\|_F^2 + \langle \frac{L_S}{\partial A_i}, \frac{L_N}{\partial A_i}\rangle + \langle \frac{L_S}{\partial B_i}, \frac{L_N}{\partial B_i}\rangle \\ &\leq -\|\frac{L_S}{\partial A_1}\|_F^2 - \|\frac{L_S}{\partial B_1}\|_F^2 - \sum_{i=1}^{2}\langle \frac{L_S}{\partial A_i}, \frac{L_N}{\partial A_i}\rangle + \langle \frac{L_S}{\partial B_i}, \frac{L_N}{\partial B_i}\rangle . \end{aligned} \tag{82}$$

In the following section, we provide lemmas that provide bounds for $\|\frac{L_S}{\partial A_1}\|_F^2 + \|\frac{L_S}{\partial B_1}\|_F^2$ and $\sum_{i=1}^{2}\langle \frac{L_S}{\partial A_i}, \frac{L_N}{\partial A_i}\rangle + \langle \frac{L_S}{\partial B_i}, \frac{L_N}{\partial B_i}\rangle$ separately.

Lemma D.1. *Under the condition that $\|Z_1\|_F \leq d_5$, $\|Z_1\|_F^2 \geq \alpha^{-d_6}\|Z_2\|_F^2$, one has*

$$\|\frac{L_S}{\partial A_1}\|_F^2 + \|\frac{L_S}{\partial B_1}\|_F^2 \geq \|\gamma_1^\top Z_1\|_F^2\left(\sigma_{W_2}^2 - \frac{\alpha^{d_6}d_5^2}{2}\sigma_{W_2}\right)L_S . \tag{83}$$

Lemma D.2. *At the end of alignment phase, one can show that the error satisfies*

$$\|E(T_1)\|_F \leq \left(1 - \frac{(1-\delta_w)}{8}\right)\sigma . \tag{84}$$

Then, since the loss under GF is non-increasing, we have $\|E(t)\|_F \leq \|E(T_1)\|_F$. Based on this, one can show that for all $T_1 \leq t \leq T_2$,

$$\|\gamma_1^\top Z_1\|_F^2 \geq \frac{(1-\delta_w)\sigma}{8\sigma_{W_2}} . \tag{85}$$

Remark D.1. *In Lemma D.1, we the following is implied in the intermediate step*

$$\|\frac{L_S}{\partial A_1}\|_F^2 + \|\frac{L_S}{\partial B_1}\|_F^2 \geq \sigma_{W_2}^2\|\gamma_1^\top Z_1\|_F^2\left(1 - \frac{\|Z_2\|_F^2}{\sigma_{W_2}}\right)L_S . \tag{86}$$

To bound $\|Z_2\|_F^2$, we use $\|Z_2\|_F^2 = \frac{\|Z_2\|_F^2}{\|Z_1\|_F^2}\|Z_1\|_F^2$, and we highlight the role of $\frac{\|Z_2\|_F^2}{\|Z_1\|_F^2}$ in the theorem.

We refer the readers to Appendix E.9 and Appendix E.10 for the proof of Lemma D.1 and Lemma D.2. By combining these results together, we can show that

$$\|\frac{L_S}{\partial A_1}\|_F^2 + \|\frac{L_S}{\partial B_1}\|_F^2 \geq \frac{(1-\delta_w)\sigma}{8}\left(\sigma_{W_2} - \frac{\alpha^{d_6}d_5^2}{2}\right)L_S \geq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{16}L_S . \tag{87}$$

Then, we show that $\sum_{i=1}^{2}\langle \frac{L_S}{\partial A_i}, \frac{L_N}{\partial A_i}\rangle + \langle \frac{L_S}{\partial B_i}, \frac{L_N}{\partial B_i}\rangle$ is small.

Lemma D.3. *One can derive the following upper bound*

$$\begin{aligned}&\left|\langle \frac{\partial L_S}{\partial A_1}, \frac{\partial L_N}{\partial A_1}\rangle\right| + \left|\langle \frac{\partial L_S}{\partial B_1}, \frac{\partial L_N}{\partial B_1}\rangle\right| + \left|\langle \frac{\partial L_S}{\partial A_2}, \frac{\partial L_N}{\partial A_2}\rangle\right| + \left|\langle \frac{\partial L_S}{\partial B_2}, \frac{\partial L_N}{\partial B_2}\rangle\right| \\ \leq& 2\sqrt{2L_S}\left(2\|W_2\|_F^2 + 2\alpha^{d_6}d_5^2\right)\times d_5^2 \times \left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2 + \frac{1}{2}\alpha^{d_6}d_5^2 \times (\|W_1\| + \frac{d_5^2}{2})\right) \\ &+ 4\sqrt{2L_S}(\|W_1\|_F^2 + \frac{1}{2}d_5^2)\alpha^{d_6}d_5^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2 + \frac{1}{2}\alpha^{d_6}d_5^2 \times (\|W_1\| + \frac{d_5^2}{2})\right) . \end{aligned} \tag{88}$$

We refer the readers to Appendix E.11 for the proof of Lemma D.3. For convenience, we use f_1 to denote

$$f_1 = 2\big(2\|W_2\|_F^2 + 2\alpha^{d_6}d_5^2\big)\times d_5^2 \times \left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2 + \frac{1}{2}\alpha^{d_6}d_5^2 \times (\|W_1\| + \frac{d_5^2}{2})\right) + 4(\|W_1\|_F^2 + \frac{1}{2}d_5^2)\alpha^{d_6}d_5^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2 + \frac{1}{2}\alpha^{d_6}d_5^2 \times (\|W_1\| + \frac{d_5^2}{2})\right). \tag{89}$$

Remark D.2. *We are particular interested in the dependence of f_1 on α. One can see that $f_1^2 \sim \mathcal{O}(\alpha^{\min(d_7, 2d_6)})$.*

Then, based on (87) and Lemma D.3, we can conclude

$$\dot{L}_S \leq -\frac{(1-\delta_w)\sigma\sigma_{W_2}}{16}L_S + f_1\sqrt{2L_S}\,. \tag{90}$$

For this ODE system, we can show that

$$\sqrt{L_S} \leq \exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right)\sqrt{L(T_1)} + \frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\,, \tag{91}$$

which leads to

$$L_S(t) \leq 2\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{16}\right)L(T_1) + 2\left(\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\right)^2, \tag{92}$$

We choose T_2 such that

$$\begin{aligned}
&\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{16}\right)L(T_1) = \left(\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\right)^2 \\
\iff &T_2 = \frac{16}{(1-\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)^2\sigma^2\sigma_{W_2}^2 L(T_1)}{1024 f_1^2}\right)
\end{aligned} \tag{93}$$

In the remaining of the proof, we show the existence of d_5, d_6, d_7. First, for convenience, we use f_2 to denote the upper bound on L_N.

$$\begin{aligned}
\frac{d}{dt}\|Z_1\|_F^2 &= 2\langle B_1, (W_2 + B_2A_2)^\top E A_1^\top\rangle + 2\langle A_1, B_1^\top(W_2 + B_2A_2)^\top E\rangle \\
&\leq 4\|A_1\|_F\|B_1\|_F\|W_2 + B_2A_2\|_F\|E\|_F \\
&\leq 2\|Z_1\|_F^2\|E\|_F(\|W_2\|_F + \frac{1}{2}d_5^2\alpha^{d_6})\,.
\end{aligned} \tag{94}$$

Thus, one can show that

$$\begin{aligned}
\log(\|Z_1\|_F^2) &\leq \int_{T_1}^{T_1+T_2}(2\|W_2\|_F + \alpha^{d_6}d_5^2)\|E(s)\|_F ds \\
&\leq \int_{T_1}^{T_1+T_2}(2\|W_2\|_F + \alpha^{d_6}d_5^2)\left(\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right)\sqrt{L(T_1)} + \frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\right)ds \\
&= \log(\|Z_1(T_1)\|_F^2) + (2\|W_2\|_F + \alpha^{d_6}d_5^2)\sqrt{L(T_1)}\frac{1 - \frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}}{\frac{(1-\delta_w)\sigma\sigma_{W_2}}{32}} \\
&\quad + \left(\frac{32f_1L(T_1)}{(1-\delta_w)\sigma\sigma_{W_2}} + f_2\right)\frac{16}{(1-\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)^2\sigma^2\sigma_{W_2}^2L(T_1)}{1024f_1^2}\right).
\end{aligned} \tag{95}$$

We set $\log(d_5)$ larger or equal to the RHS of the above inequality, denoted by $R_1(\alpha)$. Notice when α goes to zero, R_1 converges to $\log(\|Z_1(T_1)\|_F^2) + \frac{64\|W_2\|}{(1-\delta_w)\sigma\sigma_{W_2}}$. Moreover, one can show that when α is sufficiently small, $\|Z_1(T_1)\|_F^2 \leq 4g^\top B_1A_1v = (1-\delta_w)\sigma\sigma_{W_2}$. We set $d_5^* = \exp\left(2\log((1-\delta_w)\sigma\sigma_{W_2}) + \frac{64\|W_2\|}{(1-\delta_w)\sigma\sigma_{W_2}}\right)$ Thus, when α is small enough, $R_1(\alpha) \leq \log(d_5^*)$, which verifies $\|Z_1(t)\|_F \leq d_5^*$ for all $T_1 \leq t \leq T_1 + T_2$.

For d_6, we have

$$\begin{aligned}\frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right)\geq&2(1-\delta_w)\sigma_{W_2}-2\|D_1\|-2\|D_2\|\\ \geq&2(1-\delta_w)\sigma_{W_2}-2\|A_2^\top B_2^\top E\|-2\|W_2\|\|F\|-2\|EA_1^\top B_1^\top\|-2\|W_1\|\|F\|\\ \geq&2(1-\delta_w)\sigma_{W_2}-(\|Z_1\|_F^2+\|Z_2\|_F^2)\|E\|_F-2(\|W_1\|+\|W_2\|)(\|\Delta Y\|+\|E(T_1)\|)\\ \geq&2(1-\delta_w)\sigma_{W_2}-2(\|W_1\|+\|W_2\|)\|\Delta Y\|\\ &-2\big((d_5^*)^2(1+\alpha^{d_6})+\|W_1\|+\|W_2\|\big)\left(\exp\left(-\frac{(1-\delta_w)\sigma\sigma_{W_2}(t-T_1)}{32}\right)\sqrt{L(T_1)}+\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\right)\end{aligned}\tag{96}$$

Thus, we have

$$\begin{aligned}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right)\geq&\left(2(1-\delta_w)\sigma_{W_2}-2(\|W_1\|+\|W_2\|)\Delta Y\|\right)\frac{16}{(1-\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)^2\sigma^2\sigma_{W_2}^2L(T_1)}{1024f_1^2}\right)\\ &-2\big((d_5^*)^2(1+\alpha^{d_6})+\|W_1\|+\|W_2\|\big)\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}\frac{16}{(1-\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)^2\sigma^2\sigma_{W_2}^2L(T_1)}{1024f_1^2}\right)\\ &-2\big((d_5^*)^2(1+\alpha^{d_6})+\|W_1\|+\|W_2\|\big)\sqrt{L(T_1)}\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}+\log\left(\frac{\|Z_1(T_1)\|_F^2}{\|Z_2(T_1)\|_F^2}\right)\end{aligned}\tag{97}$$

We use $R_2(\alpha)$ to denote the RHS of the above inequality. Then, when α goes to zero, the LHS is of the following order

$$R_2(\alpha)\sim\log\left(\frac{1}{\alpha}\right)\left(\frac{2(1-\delta_w)}{5-\delta_w}+\frac{32\min(2d_6,d_7)\left(2(1-\delta_w)\sigma_{W_2}-2(\|W_1\|+\|W_2\|)\Delta Y\|\right)}{(1-\delta_w)\sigma\sigma_{W_2}}\right)\tag{98}$$

For d_7, for all $j\in[r]$, for any $\hat{\gamma}_1\perp\gamma_1$we have

$$\begin{aligned}\frac{d}{dt}\log\left(\frac{\cos(\gamma_1,\frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1,\frac{w_{1j}}{\|w_{1j}\|})}\right)\geq&\sigma\sigma_{W_2}-2\|D_1\|\\ \geq&\sigma\sigma_{W_2}-2\|A_2^\top B_2^\top E\|-2\|W_2\|\|F\|\\ \geq&\sigma\sigma_{W_2}-2\alpha^{d_6}(d_5^*)^2\|E\|_F-2\|W_2\|(\|\Delta Y\|+\|E\|)\,.\end{aligned}\tag{99}$$

Thus, one can show that

$$\begin{aligned}&\log\left(\frac{\cos(\gamma_1,\frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1,\frac{w_{1j}}{\|w_{1j}\|})}\right)\\ \geq&\log\left(\frac{\cos(\gamma_1,\frac{w_{1j}(T_1)}{\|w_{1j}(T_1)\|})}{\cos(\hat{\gamma}_1,\frac{w_{1j}(T_1)}{\|w_{1j}(T_1)\|})}\right)+(\sigma\sigma_{W_2}-2\|W_2\|\|\Delta Y\|)T_2\\ &-(2\alpha^{d_6}(d_5^*)^2+2\|W_2\|)T_2\left(\frac{32f_1L(T_1)}{(1-\delta_w)\sigma\sigma_{W_2}}+f_2\right)\frac{16}{(1-\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)^2\sigma^2\sigma_{W_2}^2L(T_1)}{1024f_1^2}\right)\\ &-(2\alpha^{d_6}(d_5^*)^2+2\|W_2\|)E(T_1)\frac{1-\frac{32f_1}{(1-\delta_w)\sigma\sigma_{W_2}}}{\frac{(1-\delta_w)\sigma\sigma_{W_2}}{32}}\,.\end{aligned}\tag{100}$$

Let the RHS of the above inequality be $R_3(\alpha)$. We can show that when α goes to zero, R_3 is of the following order

$$R_3(\alpha)\sim\log(\frac{1}{\alpha})\left(\frac{3+\delta_w}{5-\delta_w}-\sigma(2\|W_2\|-\sigma_{W_2})\frac{32\min(2d_6,d_7)}{(1-\delta_w)\sigma\sigma_{W_2}}\right)\tag{101}$$

Following the same argument of providing lower bound on d_5, one only needs to ensure

$$\frac{2(1-\delta_w)}{5-\delta_w} + \frac{32\min(2d_6, d_7)\Big(2(1-\delta_w)\sigma_{W_2} - 2(\|W_1\| + \|W_2\|)\Delta Y\|\Big)}{(1-\delta_w)\sigma\sigma_{W_2}} \geq \frac{1}{2}d_6$$
$$\frac{3+\delta_w}{5-\delta_w} - \sigma(2\|W_2\| - \sigma_{W_2})\frac{32\min(2d_6, d_7)}{(1-\delta_w)\sigma\sigma_{W_2}} \geq \frac{1}{2}d_7\,. \tag{102}$$

It is obvious that there exists d_6, d_7 such that the above inequality holds. □

E Proof of Several Lemmas Used in Appendix C and Appendix D

E.1 Proof of Lemma C.1

Proof. We start with characterizing each column of Z_1 align with γ_1. Let $W_{1j} = Z_1 e_j, \forall j \in [r]$ where e_j is the standard basis. Then, we have

$$\dot{w}_{1j} = \sigma\sigma_{W_2} H_1 w_{1j} + \hat{D}_1 w_{1j}\,. \tag{103}$$

Then, based on (103), we can characterize the growth of $\|w_{1j}\|$ as follows

$$\begin{aligned}\frac{d}{dt}\|w_{1j}\|^2 &= 2\langle w_{1j}, \dot{w}_{1j}\rangle \\ &\leq 2(\sigma\sigma_{W_2} + \|\hat{D}_1\|)\|w_{1j}\|^2 \\ &= 2(\sigma\sigma_{W_2} + \|D_1\|)\|w_{1j}\|^2\,, \qquad \text{Apply Lemma A.4}\end{aligned} \tag{104}$$

Moreover, one can use a similar argument to derive the lower bound on the growth of $\|w_{1j}\|^2$

$$\begin{aligned}\frac{d}{dt}\|w_{1j}\|^2 &= 2\langle w_{1j}, \dot{w}_{1j}\rangle \\ &\geq 2(\sigma\sigma_{W_2} - \|\hat{D}_1\|)\|w_{1j}\|^2 \\ &= 2(\sigma\sigma_{W_2} - \|D_1\|)\|w_{1j}\|^2\,, \qquad \text{Apply Lemma A.4}\end{aligned} \tag{105}$$

Since $\|Z_1\|_F^2 = \sum_{j=1}^r \|w_{1j}\|^2$, therefore one has

$$\begin{aligned}\frac{d}{dt}\|Z_1\|_F^2 &= \sum_{j=1}^r \frac{d}{dt}\|w_{1j}\|^2 \leq 2(\sigma\sigma_{W_2} + \|D_1\|)\sum_{j=1}^r \|w_{1j}\|^2 = 2(\sigma\sigma_{W_2} + \|D_1\|)\|Z_1\|_F^2\,, \\ \frac{d}{dt}\|Z_1\|_F^2 &= \sum_{j=1}^r \frac{d}{dt}\|w_{1j}\|^2 \geq 2(\sigma\sigma_{W_2} - \|D_1\|)\sum_{j=1}^r \|w_{1j}\|^2 = 2(\sigma\sigma_{W_2} - \|D_1\|)\|Z_1\|_F^2\,,\end{aligned} \tag{106}$$

The above equations yield the following

$$2(\sigma\sigma_{W_2} - \|D_1\|) \leq \frac{d}{dt}\log(\|Z_1\|_F^2) \leq 2(\sigma\sigma_{W_2} + \|D_1\|)\,. \tag{107}$$

A similar results hold for $\|Z_2\|_F^2$ respectively. Thus, one can show the following characterization of the growth of $\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}$ as follows

$$\frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) \geq 2(\sigma\sigma_{W_2} - \|D_1\| - \sigma\sigma_{W_1} - \|D_2\|) = 2\big((1-\delta_w)\sigma\sigma_{W_2} - \|D_1\| - \|D_2\|\big)\,. \tag{108}$$

Now, we move on to the study of the angular dynamics of w_{1j}.

$$\begin{aligned}\frac{d}{dt}\frac{w_{1j}}{\|w_{1j}\|} &= \frac{\dot{w}_{1j}}{\|w_{1j}\|} - \frac{w_{1j}}{\|w_{1j}\|^2}\cdot\frac{\dot{w}_{1j}^\top w_{1j}}{\|w_{1j}\|} \\ &= \sigma\sigma_{W_2}\left(I - \frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{H_1 w_{1j}}{\|w_{1j}\|} + \left(I - \frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{\hat{D}_1 w_{1j}}{\|w_{1j}\|}\,.\end{aligned} \tag{109}$$

For γ_1, we have

$$\begin{aligned}\frac{d}{dt}\cos\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)=&\left\langle \gamma_1, \sigma\sigma_{W_2}\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{H_1w_{1j}}{\|w_{1j}\|}+\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{\hat{D}_1w_{1j}}{\|w_{1j}\|}\right\rangle\\ =&\sigma\sigma_{W_2}\gamma_1^\top\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\left(\frac{(\gamma_1\gamma_1^\top-\bar{\gamma}_1\bar{\gamma}_1^\top)w_{1j}}{\|w_{1j}\|}\right)+\frac{\gamma_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\\ =&\sigma\sigma_{W_2}\cos\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\left[1-\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\right]+\sigma\sigma_{W_2}\cos\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\cos^2\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\\ &+\frac{\gamma_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\end{aligned}\tag{110}$$

For $\bar{\gamma}_1$, we have

$$\begin{aligned}\frac{d}{dt}\cos\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)=&\left\langle \bar{\gamma}_1, \sigma\sigma_{W_2}\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{H_1w_{1j}}{\|w_{1j}\|}+\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{\hat{D}_1w_{1j}}{\|w_{1j}\|}\right\rangle\\ =&\sigma\sigma_{W_2}\bar{\gamma}_1^\top\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\left(\frac{(\gamma_1\gamma_1^\top-\bar{\gamma}_1\bar{\gamma}_1^\top)w_{1j}}{\|w_{1j}\|}\right)+\frac{\bar{\gamma}_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\\ =&-\sigma\sigma_{W_2}\cos\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\left[1-\cos^2\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\right]-\sigma\sigma_{W_2}\cos\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\\ &+\frac{\bar{\gamma}_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\end{aligned}\tag{111}$$

For any unit vector $\tilde{\gamma}_1 \perp \gamma_1, \bar{\gamma}_1$, we have

$$\begin{aligned}\frac{d}{dt}\cos\left(\tilde{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)=&\left\langle \tilde{\gamma}_1, \sigma\sigma_{W_2}\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{H_1w_{1j}}{\|w_{1j}\|}+\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\frac{\hat{D}_1w_{1j}}{\|w_{1j}\|}\right\rangle\\ =&\sigma\sigma_{W_2}\tilde{\gamma}_1^\top\left(I-\frac{w_{1j}w_{1j}^\top}{\|w_{1j}\|^2}\right)\left(\frac{(\gamma_1\gamma_1^\top-\bar{\gamma}_1\bar{\gamma}_1^\top)w_{1j}}{\|w_{1j}\|}\right)+\frac{\tilde{\gamma}_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\tilde{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\\ =&-\sigma\sigma_{W_2}\cos\left(\tilde{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\left[\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)-\cos^2\left(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\right]\\ &+\frac{\tilde{\gamma}_1^\top\hat{D}_1w_{1j}}{\|w_{1j}\|}-\cos\left(\tilde{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\frac{w_{1j}^\top\hat{D}_1w_{1j}}{\|w_{1j}\|^2}\end{aligned}\tag{112}$$

Based on (110), (111) and (112), we can show

$$\frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\bar{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right)\geq 2\sigma\sigma_{W_2}-2\|D_1\|\,,\quad \frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\tilde{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right)\geq \sigma\sigma_{W_2}-2\|D_1\|\,.\tag{113}$$

Thus, for any unit vector $\hat{\gamma}_1$ that is orthogonal to γ_1, we have

$$\frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right)\geq \sigma\sigma_{W_2}-2\|D_1\|\,.\tag{114}$$

□

E.2 Proof of Lemma C.2

Proof. Since for all $0 \le t \le T_1(\delta_1, \delta_2)$, we have $\|D_1(T_1(\delta_1,\delta_2))\| \le \delta_1, \|D_2(T_1(\delta_1,\delta_2))\| \le \delta_2$, then one can further derive the following based on Lemma C.1

$$\begin{aligned}
\frac{d}{dt}\log(\|w_{1j}\|^2) \le 2(\sigma\sigma_{W_2}+\delta_1)\,, \quad \frac{d}{dt}\log(\|Z_1\|_F^2) \le 2(\sigma\sigma_{W_2}+\delta_1) \\
\frac{d}{dt}\log\left(\frac{\|Z_1\|_F^2}{\|Z_2\|_F^2}\right) \ge 2((1-\delta_w)\sigma\sigma_{W_2}-\delta_1-\delta_2) > 0\,, \\
\frac{d}{dt}\log\left(\frac{\cos(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|})}{\cos(\hat{\gamma}_1, \frac{w_{1j}}{\|w_{1j}\|})}\right) \ge \sigma\sigma_{W_2} - 2\delta_1 > 0\,.
\end{aligned} \tag{115}$$

Therefore, we can show that for all $0 \le t \le T_1(\delta_1,\delta_2)$

$$\begin{aligned}
\|w_{1j}(t)\|^2 \le \exp\Big(2(\sigma\sigma_{W_2}+\delta_1)t\Big)\|w_{1j}(0)\|^2 \\
\|Z_1(t)\|_F^2 \le \exp\Big(2(\sigma\sigma_{W_2}+\delta_1)t\Big)\|Z_1(0)\|_F^2 \\
\frac{\|Z_1(t)\|_F^2}{\|Z_2(t)\|_F^2} \ge \exp\Big(2\big((1-\delta_w)\sigma\sigma_{W_2}-\delta_1-\delta_2\big)t\Big)\frac{\|Z_1(0)\|_F^2}{\|Z_2(0)\|_F^2} \\
\frac{\cos\left(\gamma_1, \frac{w_{1j}(t)}{\|w_{1j}(t)\|}\right)}{\cos\left(\hat{\gamma}_1, \frac{w_{1j}(t)}{\|w_{1j}(t)\|}\right)} \ge \exp\big((\sigma\sigma_{W_2}-2\delta_1)t\big)\frac{\cos\left(\gamma_1, \frac{w_{1j}(0)}{\|w_{1j}(0)\|}\right)}{\cos\left(\hat{\gamma}_1, \frac{w_{1j}(0)}{\|w_{1j}(0)\|}\right)}
\end{aligned} \tag{116}$$

We select $\hat{\gamma}_{1,2}, \hat{\gamma}_{1,3}, \cdots, \hat{\gamma}_{1,n+h}$ that forms an orthogonal basis for $\mathbb{R}^{n+h}$, since the following holds

$$\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) + \sum_{i=2}^{m+h}\cos^2\left(\gamma_{1,i}, \frac{w_{1j}}{\|w_{1j}\|}\right) = 1 \tag{117}$$

One can lower bound $\cos\left(\gamma_1, \frac{w_{1j}(t)}{\|w_{1j}(t)\|}\right)$ as follows

$$\begin{aligned}
1 &= \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) + \sum_{i=2}^{m+h}\cos^2\left(\gamma_{1,i}, \frac{w_{1j}}{\|w_{1j}\|}\right) \\
&\le \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) + (m+h-1)\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\exp\big(-2(\sigma\sigma_{W_2}-2\delta_1)t\big)c_{\max} \\
&= \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right)\times\left(1 + c_{\max}(m+h-1)\exp\big(-2(\sigma\sigma_{W_2}-2\delta_1)t\big)\right) \\
&\iff \cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) \ge \frac{1}{1+d_2\exp\big(-2(\sigma\sigma_{W_2}-2\delta_1)t\big)}
\end{aligned} \tag{118}$$

where in the above equation, $\frac{w_{1j}}{\|w_{1j}\|}$ is its value evaluated at t. We omit the time dependency for simplicity. □

E.3 Proof of Lemma C.3

Proof. Under the assumption that $\delta_1 = \delta_2 = \delta := \alpha$, we first derive an upper bound on $\|D_1\|$.

$$\begin{aligned}
\|D_1\| =& \|A_2^\top B_2^\top E - W_2^\top (W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1)\| \\
\leq& \|A_2^\top B_2^\top\| \cdot \|E\| + \|W_2\| \cdot \big(\|W_2\| \cdot \|B_2 A_2\| + \|W_1\| \cdot \|B_1 A_1\| + \|B_1 A_1\| \cdot \|B_2 A_2\|\big) \\
\leq& (\|E\| + \|W_2\|^2)\|B_2 A_2\| + \|W_1\| \cdot \|W_2\| \cdot \|B_1 A_1\| + \|W_2\| \cdot \|B_1 A_1\| \cdot \|B_2 A_2\| \\
\leq& (\|E\| + \|W_2\|^2)\|Z_2\|_F^2 + \|W_1\| \cdot \|W_2\| \cdot \|Z_1\|_F^2 + \frac{\|W_2\|}{4} \cdot \|Z_1\|_F^2 \cdot \|Z_2\|_F^2 && \text{Lemma A.2} \\
\leq& \alpha^2 z_{\max}^2 (\|E\| + \|W_2\|^2) \exp(2(\sigma\sigma_{W_1} + \delta_2)t) + \alpha^2 z_{\max}^2 \|W_1\| \cdot \|W_2\| \exp(2(\sigma\sigma_{W_2} + \delta_1)t) \\
& + \alpha^4 z_{\max}^4 \|W_2\| \exp(2((1+\delta_w)\sigma\sigma_{W_2} + \delta_1 + \delta_2)t) && \text{Lemma C.2} \\
\leq& \alpha^2 z_{\max}^2 (\|E\| + \|W_2\|^2 + \|W_1\| \cdot \|W_2\|) \exp(2(\sigma\sigma_{W_2} + \delta)t) + \alpha^4 z_{\max}^4 \|W_2\| \exp(4(\sigma\sigma_{W_2} + \delta)t)\,. && (119)
\end{aligned}$$

where in the last inequality, we use the property that $\sigma_{W_2} > \sigma_{W_1}$. Similarly, we can show that

$$\|D_2\| \leq \alpha^2 z_{\max}^2 (\|E\| + \|W_1\|^2 + \|W_1\| \cdot \|W_2\|) \exp(2(\sigma\sigma_{W_2} + \delta)t) + \alpha^4 z_{\max}^4 \|W_1\| \exp(4(\sigma\sigma_{W_2} + \delta)t)\,. \tag{120}$$

Furthermore, one can derive the following union upper bound on $\|D_1\|, \|D_2\|$

$$\begin{aligned}
\max(\|D_1\|, \|D_2\|) \leq& \alpha^2 z_{\max}^2 (\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2) \exp(2(\sigma\sigma_{W_2} + \delta)t) \\
& + \alpha^4 z_{\max}^4 (\|W_1\| + \|W_2\|) \exp(4(\sigma\sigma_{W_2} + \delta)t)\,. && (121)
\end{aligned}$$

Then, under the assumption that $\alpha \leq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$, we will derive a lower bound on $T_1(\delta, \delta)$ by studying the following inequality

$$\begin{aligned}
\delta \geq& \alpha^2 z_{\max}^2 (\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2) \exp(2(\sigma\sigma_{W_2} + \delta)t) \\
& + \alpha^4 z_{\max}^4 (\|W_1\| + \|W_2\|) \exp(4(\sigma\sigma_{W_2} + \delta)t) && (122)
\end{aligned}$$

Since we assume $\delta = \alpha \leq 1$, then both terms in the RHS are smaller than 1, therefore, we can further lower bound $T(\delta, \delta)$ as

$$\begin{aligned}
\alpha \geq& \alpha^2 z_{\max}^2 (\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2) \exp\left(\frac{t(5-\delta_w)\sigma\sigma_{W_2}}{2}\right) \\
& + \alpha^2 z_{\max}^2 \sqrt{\|W_1\| + \|W_2\|} \exp\left(\frac{t(5-\delta_w)\sigma\sigma_{W_2}}{2}\right) \\
=& \alpha^2 z_{\max}^2 \big(\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2 + \sqrt{\|W_1\| + \|W_2\|}\big) \exp\left(\frac{t(5-\delta_w)\sigma\sigma_{W_2}}{2}\right) \\
\iff \exp\left(\frac{t(5-\delta_w)\sigma\sigma_{W_2}}{2}\right) =& \frac{1}{\alpha z_{\max}^2 \big(\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2 + \sqrt{\|W_1\| + \|W_2\|}\big)} \\
\iff t =& \frac{2}{(5-\delta_w)\sigma\sigma_{W_2}} \log\left(\frac{1}{\alpha z_{\max}^2 \big(\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2 + \sqrt{\|W_1\| + \|W_2\|}\big)}\right). && (123)
\end{aligned}$$

For convenience, we define $d_3 = \frac{1}{z_{\max}^2 \left(\|E\| + \|W_1\|^2 + \|W_1\|\|W_2\| + \|W_2\|^2 + \sqrt{\|W_1\| + \|W_2\|}\right)}$. Thus, we set $T_1(\alpha, \alpha) = \frac{2}{(5-\delta_w)\sigma\sigma_{W_2}} \log\left(\frac{1}{\alpha z_{\max}^2 d_3}\right)$. Furthermore, based on Lemma C.2, we can further characterize the following properties

$$\begin{aligned}
\frac{\|Z_1(t)\|_F^2}{\|Z_2(t)\|_F^2} \geq& d_1 \exp\Big(2\big((1-\delta_w)\sigma\sigma_{W_2} - \delta_1 - \delta_2\big)t\Big)\,, \\
=& d_1 \exp\big((1-\delta_w)t\big) \\
=& d_1 d_3^{\frac{2(1-\delta_w)}{5-\delta_w}} \alpha^{-\frac{2(1-\delta_w)}{5-\delta_w}}\,, && (124)
\end{aligned}$$

Moreover, the alignment can be described as

$$\begin{aligned}\cos^2\left(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}\right) &\geq \frac{1}{1+d_2\exp\big(-2(\sigma\sigma_{W_2}-2\delta_1)t\big)}\\ &\geq \frac{1}{1+d_2\exp\big(-2(\sigma\sigma_{W_2}-2\delta_1)t\big)}\\ &\geq \frac{1}{1+d_2\exp\big(-(1-\delta_w)\sigma\sigma_{W_2}t\big)}\\ &\geq \frac{1}{1+d_2 d_3^{-\frac{2(1-\delta_w)}{5-\delta_w}}\alpha^{\frac{2(1-\delta_w)}{5-\delta_w}}}\\ &\geq 1-d_2 d_3^{-\frac{2(1-\delta_w)}{5-\delta_w}}\alpha^{\frac{2(1-\delta_w)}{5-\delta_w}}\,. \end{aligned} \tag{125}$$

Finally, we show that until $T_1(\alpha,\alpha)$, the norm of LoRA weights stay small. In Lemma C.2, we have shown that

$$\begin{aligned}\|Z_1(t)\|_F^2 &\leq \exp\bigg(2(\sigma\sigma_{W_2}+\delta_1)t\bigg)\|Z_1(0)\|_F^2\\ &\leq \alpha^2 z_{\max}^2\exp\bigg(\frac{t(5-\delta_w)\sigma\sigma_{W_2}}{2}\bigg)\\ &= \frac{\alpha}{d_3}\,. \end{aligned} \tag{126}$$

A similar argument holds for Z_2 as well. □

E.4 Proof of Lemma C.4

Proof.

$$\begin{aligned}\frac{d}{dt}g^\top B_1A_1v &= g^\top\dot{B}_1A_1v + g^\top B_1\dot{A}_1v\\ &= g^\top\big(\sigma\sigma_{W_2}gv^\top + D_1\big)A_1^\top A_1v + g^\top B_1B_1^\top\big(\sigma\sigma_{W_2}gv^\top+D_1\big)v\\ &= \sigma\sigma_{W_2}(g^\top B_1B_1^\top g + v^\top A_1^\top A_1v) + g^\top D_1A_1^\top A_1v + g^\top B_1B_1^\top D_1^\top v\\ &\geq (\sigma\sigma_{W_2}-\|D_1\|)(g^\top B_1B_1^\top g + v^\top A_1^\top A_1v)\\ &\geq (\sigma\sigma_{W_2}-\|D_1\|)\|Z_1\|_F^2\min_{j\in[r]}\cos^2\left(\gamma_1,\frac{w_{1j}}{\|w_{1j}\|}\right). \qquad \text{Apply Lemma A.6} \end{aligned} \tag{127}$$

□

E.5 Proof of Lemma C.5

Proof. We first introduce the *imbalance* quantity: $A_1^\top A_1 - B_1B_1^\top$. This quantity is constant when LoRA weights are trained via GF. To highlight its dependcy on the α, we let $A_1A_1^\top - B_1^\top B_1 := \alpha^2\Lambda_1$ where Λ_1 is independent of α and purely determined at initialization. Then, one can first show that

$$\|B_1A_1\| \leq \|B_1A_1\|_F \leq \|A_1\|_F\|B_1\|_F \leq \frac{1}{2}\|Z_1\|_F^2\,. \tag{128}$$

On the other hand,

$$\begin{aligned} &A_1A_1^\top - B_1^\top B_1 = \alpha^2\Lambda_1\\ \Rightarrow &B_1A_1A_1^\top B_1^\top = B_1B_1^\top B_1B_1^\top + \alpha^2 B_1\Lambda_1B_1^\top\\ \Rightarrow &\|B_1A_1A_1^\top B_1^\top\| \geq \|B_1B_1^\top B_1B_1^\top\| - \alpha^2\|B_1\Lambda_1B_1^\top\|\\ \iff &\|B_1A_1\|^2 \geq \|B_1\|^4 - \alpha^2\|\Lambda_1\|\|B_1\|^2\\ \Rightarrow &\|B_1A_1\|^2 \geq \frac{1}{r^2}\|B_1\|_F^4 - \alpha^2\|\Lambda_1\|\|B_1\|_F^2 \end{aligned} \tag{129}$$

Similarly, one can show $\|B_1A_1\|^2 \geq \frac{1}{r^2}\|A_1\|_F^4 - \alpha^2\|\Lambda_1\|\|A_1\|_F^2$. Combine these results together, one has

$$\begin{aligned} 2\|B_1A_1\|^2 &\geq \frac{1}{r^2}(\|A_1\|_F^4 + \|B_1\|_F^4) - \alpha^2\|\Lambda_1\|\|Z_1\|_F^2 \\ &\geq \frac{1}{2r^2}\|Z_1\|_F^4 - \alpha^2\|\Lambda_1\|\|Z_1\|_F^2 \,. \end{aligned} \tag{130}$$

By solving the above inequality, one has

$$\begin{aligned} \|Z_1\|_F^2 &\leq \alpha^2 r^2\|\Lambda_1\| + r^2\sqrt{\alpha^4\|\Lambda_1\|^2 + \frac{4}{r^2}\|B_1A_1\|^2} \\ &\leq \alpha^2 r^2\|\Lambda_1\| + \alpha^2 r^2\|\Lambda_1\| + 2r\|B_1A_1\| \qquad \text{we use } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}. \end{aligned} \tag{131}$$

Thus, we have $\|B_1A_1\| \geq \frac{\|Z_1\|_F^2 - 2\alpha^2 r^2\|\Lambda_1\|}{2r}$. □

E.6 Proof of Lemma C.6

Proof. We prove $\|g_\perp g_\perp^\top B_1A_1\|^2$ here. The same analysis can be applied to derive upper bound on $\|B_1A_1v_\perp v_\perp^\top\|^2$. Notice $\tilde{\gamma}_1 = \begin{pmatrix} g_\perp \\ 0_{n\times(h-1)} \end{pmatrix}$ is orthogonal to γ_1. Therefore, one can show that for each index j, we have

$$\cos^2(\gamma_1, \frac{w_{1j}}{\|w_{1j}\|}) + \sum_{i=1}^{h-1}\cos^2(\tilde{\gamma}_{1i}, \frac{w_{1j}}{\|w_{1j}\|}) \leq 1\,, \tag{132}$$

where we use $\tilde{\gamma}_{1i}$ to denote the i-th column of $\tilde{\gamma}_1$. This implies $\sum_{i=1}^{h-1}\cos^2(\tilde{\gamma}_{1i}, \frac{w_{1j}}{\|w_{1j}\|}) \leq \beta_2\alpha^d$. Now, we consider $\tilde{\gamma}_1\tilde{\gamma}_1^\top Z_1$

$$\begin{aligned} \|\tilde{\gamma}_1\tilde{\gamma}_1^\top Z_1\|_F^2 &= \sum_{j=1}^{r}\|\tilde{\gamma}_1\tilde{\gamma}_1^\top w_{1j}\|^2 \\ &\leq \sum_{j=1}^{r}\sum_{i=1}^{h-1}\cos^2\left(\tilde{\gamma}_{1i}, \frac{w_{1j}}{\|w_{1j}\|}\right)\|w_{1j}\|^2 \\ &\leq \beta_2\alpha^d\|Z_1\|_F^2 \,. \end{aligned} \tag{133}$$

Moreover, we can show that

$$\|\tilde{\gamma}_1\tilde{\gamma}_1^\top Z_1\|_F = \|\tilde{\gamma}_1^\top Z_1\|_F = \|g_\perp B_1\|_F \,. \tag{134}$$

Thus, we can show that

$$\|g_\perp g_\perp^\top B_1A_1\| \leq \|g_\perp B_1\|\|A_1\| \leq \sqrt{\beta_2\alpha^d}\|Z_1\|_F^2 \,. \tag{135}$$

□

E.7 Proof of Lemma C.7

Proof. We start with

$$\begin{aligned} & A_1A_1^\top - B_1^\top B_1 := \alpha^2\Lambda_1 \\ \Rightarrow & B_1A_1A_1^\top B_1^\top = B_1B_1^\top B_1B_1^\top + \alpha^2 B_1\Lambda_1B_1^\top \\ \Longleftrightarrow & g^\top B_1A_1(vv^\top + v_\perp v_\perp^\top)A_1^\top B_1^\top g = g^\top B_1B_1^\top(gg^\top + g_\perp g_\perp^\top)B_1B_1^\top g + \alpha^2 g^\top B_1\Lambda_1B_1^\top g \\ \Rightarrow & (g^\top B_1B_1^\top g)^2 \leq (v^\top A_1^\top B_1^\top g)^2 + (v^\top A_1^\top B_1^\top g_\perp)^2 + (g^\top B_1B_1^\top g_\perp)^2 + \alpha^2\|\Lambda_1\|g^\top B_1B_1^\top g \end{aligned} \tag{136}$$

Based on Lemma C.6, we have shown that

$$\|g_\perp g_\perp^\top B_1A_1\|^2, \|B_1A_1v_\perp v_\perp^\top\|^2 \leq \beta_2\alpha^{\frac{3+\delta_w}{5-\delta_w}}\|Z_1\|_F^4 \,. \tag{137}$$

Therefore, one has

$$(v^\top A_1^\top B_1^\top g_\perp)^2 \le \|g_\perp g_\perp^\top B_1 A_1\|^2 \le \beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|Z_1\|_F^4 . \tag{138}$$

Moreover, using the same technique, one can show

$$\begin{aligned}(g^\top B_1 B_1^\top g_\perp)^2 &\le \|g_\perp^\top B_1\|_F \|B_1\| \\ &\le \beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|B_1\| \|B_1\| \\ &\le \beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|Z_1\|_F^2 .\end{aligned} \tag{139}$$

Combine all the equations together, we have

$$\begin{aligned}(g^\top B_1 B_1^\top g)^2 &\le (v^\top A_1^\top B_1^\top g)^2 + (v^\top A_1^\top B_1^\top g_\perp)^2 + (g^\top B_1 B_1^\top g_\perp)^2 + \alpha^2 \|\Lambda_1\| g^\top B_1 B_1^\top g \\ &\le (v^\top A_1^\top B_1^\top g)^2 + \alpha^2 \|\Lambda_1\| g^\top B_1 B_1^\top g + 2\beta_2 \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|Z_1\|_F^4 .\end{aligned} \tag{140}$$

□

E.8 Proof of Lemma C.8

Proof. We first show that based on these assumptions, one can use $g^\top B_1 A_1 v$ to upper bound $\|Z_1\|_F^2$. Based on Lemma C.7, (59) and (63), we can show that

$$\begin{aligned}\left(1 - d_2 d_3^{-\frac{3+\delta_w}{5-\delta_w}} \alpha^{\frac{3+\delta_w}{5-\delta_w}}\right) \|Z_1\|_F^2 &\le \|\gamma_1 \gamma_1^\top Z_1\|_F^2 \\ &\le g^\top B_1 B_1^\top g + v^\top A_1^\top A_1 v + 2 g^\top B_1 A_1 v \\ &\le (2+\sqrt{2}) g^\top B_1 A_1 v + \alpha^2 \|D_1\| + 2\sqrt{\beta} \|Z_1\|_F^2 \alpha^{\frac{3+\delta_w}{10-2\delta_w}} \\ &\le (2+\sqrt{2}) g^\top B_1 A_1 v + \alpha^2 \frac{(1-\delta_w)\sigma\sigma_{W_2}}{2} + 2\sqrt{\beta} \|Z_1\|_F^2 \alpha^{\frac{3+\delta_w}{10-2\delta_w}} .\end{aligned} \tag{141}$$

Thus, when α is sufficently small, one has

$$\|Z_1\|_F^2 \le 2(2+\sqrt{2}) g^\top B_1 A_1 v \le 2(2+\sqrt{2}) \frac{(1-\delta_w)\sigma}{4\sigma_{W_2}} . \tag{142}$$

Then, we move the objective of interest.

$$\begin{aligned}&\|u u^\top F v_\perp v_\perp^\top A_1^\top B_1^\top g g^\top\|_F + \|B_2 A_2\|_F (\|W_1\|_F + \|B_1 A_1\|_F) \\ &+ \|u_\perp u_\perp^\top F A_1^\top B_1^\top\|_F + \|u_\perp u_\perp^\top F W_1^\top\|_F + \|E A_1^\top B_1^\top g_\perp g_\perp^\top\|_F + \|F W_1^\top g_\perp g_\perp^\top\|_F \\ \le &\|F v_\perp v_\perp^\top\|_F \|B_1 A_1 v_\perp v_\perp^\top\|_F + \frac{1}{2}\|Z_2\|_F^2 (\|W_1\|_F + \frac{1}{2}\|Z_1\|_F^2) \\ &+ \|u_\perp u_\perp^\top F\|_F \|B_1 A_1\|_F + \|W_1\| \|u_\perp u_\perp^\top F\|_F + \|E\|_F \|g_\perp g_\perp^\top B_1 A_1\|_F + \|W_1\| \|F v_\perp v_\perp^\top\|_F\end{aligned} \tag{143}$$

Based on Lemma C.6, we have

$$\|g_\perp g_\perp^\top B_1 A_1\|_F^2, \|B_1 A_1 v_\perp v_\perp^\top\|_F^2 \le \beta_2 r^2 \alpha^{\frac{3+\delta_w}{5-\delta_w}} \|Z_1\|_F^4 . \tag{144}$$

One can use the above bound to show that

$$\begin{aligned}\|F v_\perp v_\perp^\top\|_F &\le \|W_2 B_1 A_1 v_\perp v_\perp^\top\|_F + \|B_2 A_2 W_1 v_\perp v_\perp^\top\|_F + \|B_2 A_2 B_1 A_1 v_\perp v_\perp^\top\|_F \\ &\le \|W_2\| \|B_1 A_1 v_\perp v_\perp^\top\|_F + \|W_1\| \frac{1}{2}\|Z_2\|_F^2 + \frac{1}{4}\|Z_2\|_F^2 \|Z_1\|_F^2 \\ &\le \sqrt{\beta_2 r^2 \alpha^{\frac{3+\delta_w}{5-\delta_w}}} \|Z_1\|_F^2 \|W_2\| + \frac{\|W_1\|}{2}\alpha^d \|Z_1\|_F^2 + \frac{1}{4}\alpha^d \|Z_1\|^4 ,\end{aligned} \tag{145}$$

and similarly

$$\|u_\perp u_\perp^\top F\|_F \le \sqrt{\beta_2 r^2 \alpha^{\frac{3+\delta_w}{5-\delta_w}}} \|Z_1\|_F^2 \|W_2\| + \frac{\|W_1\|}{2}\alpha^d \|Z_1\|_F^2 + \frac{1}{4}\alpha^d \|Z_1\|^4 . \tag{146}$$

Therefore, based on (144), (145) and (146), we can show

$$
\begin{aligned}
&\|uu^\top Fv_\perp v_\perp^\top A_1^\top B_1^\top gg^\top\|_F + \|B_2A_2\|_F(\|W_1\|_F + \|B_1A_1\|_F)\\
&+ \|u_\perp u_\perp^\top FA_1^\top B_1^\top\|_F + \|u_\perp u_\perp^\top FW_1^\top\|_F + \|EA_1^\top B_1^\top g_\perp g_\perp^\top\|_F + \|FW_1^\top g_\perp g_\perp^\top\|_F\\
\le&\|Fv_\perp v_\perp^\top\|_F\|B_1A_1v_\perp v_\perp^\top\|_F + \frac{1}{2}\alpha^d\|Z_1\|_F^2(\|W_1\|_F + \frac{1}{2}\|Z_1\|_F^2)\\
&+ \|u_\perp u_\perp^\top F\|_F\frac{1}{2}\|Z_1\|_F^2 + \|W_1\|\|u_\perp u_\perp^\top F\|_F + \|E\|_F\|g_\perp g_\perp^\top B_1A_1\|_F + \|W_1\|\|Fv_\perp v_\perp^\top\|_F\\
\le&\|Z_1\|_F^2\Big\{\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|Z_1\|_F^2\|W_2\| + \frac{\|W_1\|}{2}\alpha^d\|Z_1\|_F^2 + \frac{1}{4}\alpha^d\|Z_1\|^4\big)\\
&+ \frac{1}{2}\alpha^d(\|W_1\|_F + \frac{1}{2}\|Z_1\|_F^2)\\
&+ \frac{1}{2}\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|Z_1\|_F^2\|W_2\| + \frac{\|W_1\|}{2}\alpha^d\|Z_1\|_F^2 + \frac{1}{4}\alpha^d\|Z_1\|^4\big)\\
&+ 2\|W_1\|\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|W_2\| + \frac{\|W_1\|}{2}\alpha^d + \frac{1}{4}\alpha^d\|Z_1\|^2\big)\\
&+ \|E(0)\|_F\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\Big\}\\
\le&2(2+\sqrt{2})g^\top B_1A_1v\Big\{\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|Z_1\|_F^2\|W_2\| + \frac{\|W_1\|}{2}\alpha^d\|Z_1\|_F^2 + \frac{1}{4}\alpha^d\|Z_1\|^4\big)\\
&+ \frac{1}{2}\alpha^d(\|W_1\|_F + \frac{1}{2}\|Z_1\|_F^2)\\
&+ \frac{1}{2}\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|Z_1\|_F^2\|W_2\| + \frac{\|W_1\|}{2}\alpha^d\|Z_1\|_F^2 + \frac{1}{4}\alpha^d\|Z_1\|^4\big)\\
&+ 2\|W_1\|\big(\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\|W_2\| + \frac{\|W_1\|}{2}\alpha^d + \frac{1}{4}\alpha^d\|Z_1\|^2\big)\\
&+ \|E(0)\|_F\sqrt{\beta_2r^2\alpha^{\frac{3+\delta_w}{5-\delta_w}}}\Big\}. \qquad (147)
\end{aligned}
$$

Notice all the terms in the big bracket goes to zero as α goes to zero. Thus, when α is sufficiently small, one can show that the RHS of the above inequality is upper bounded by $\left(\frac{(3+\delta_w)\sigma}{4} + \sigma_{W_2}\sigma_{W_1}\right)g^\top B_1A_1v$, which completes the proof. □

E.9 Proof of Lemma D.1

Proof. We study $\|\frac{L_S}{\partial A_1}\|_F^2$ as an example.

$$
\begin{aligned}
\|\frac{L_S}{\partial A_1}\|_F^2 &= \|B_1^\top(W_2 + B_2A_2)^\top uu^\top Evv^\top\|_F^2\\
&= (u^\top Ev)^2\mathrm{Tr}\Big(u^\top(W_2 + B_2A_2)B_1B_1^\top(W_2 + B_2A_2)^\top u\Big)\\
&= 2u^\top(W_2 + B_2A_2)B_1B_1^\top(W_2 + B_2A_2)^\top u \times L_S\,. \qquad (148)
\end{aligned}
$$

Similarly, one can show

$$
\|\frac{L_S}{\partial B_1}\|_F^2 = 2v^\top A_1^\top A_1v \times u^\top(W_2 + B_2A_2)(W_2 + B_2A_2)^\top u \times L_S\,. \qquad (149)
$$

Notice first

$$
u^\top(W_2 + B_2A_2)(W_2 + B_2A_2)^\top u = \sigma_{W_2}^2 + u^\top B_2A_2A_2^\top B_2^\top u + \sigma_{W_2}g^\top B_2A_2u \ge \sigma_{W_2}^2 - \frac{\|Z_2\|_F^2}{2}\sigma_{W_2}\,. \qquad (150)
$$

Under the condition that $\|Z_1\|_F \le d_5, \|Z_2\|_F^2 \le \alpha^{d_6}\|Z_1\|_F^2$, we can show that

$$
u^\top(W_2 + B_2A_2)(W_2 + B_2A_2)^\top u = \sigma_{W_2}^2 - \frac{\alpha^{d_6}d_5^2}{2}\sigma_{W_2}\,. \qquad (151)
$$

On the other hand, we can show that

$$g^\top B_1 B_1^\top g + v^\top A_1^\top A_1 v = \|g^\top B_1\|^2 + \|A_1 v\|^2 \geq \frac{1}{2}\|g^\top B_1 + v^\top A_1^\top\|^2 = \frac{1}{2}\|\gamma_1^\top Z_1\|_F^2 \,. \tag{152}$$

Therefore, one can conclude that

$$\|\frac{L_S}{\partial A_1}\|_F^2 + \|\frac{L_S}{\partial B_1}\|_F^2 \geq \|\gamma_1^\top Z_1\|_F^2 \left(\sigma_{W_2}^2 - \frac{\alpha^{d_6} d_5^2}{2}\sigma_{W_2}\right) L_S \,. \tag{153}$$

□

E.10 Proof of Lemma D.2

Proof. We first show that at the end of *alignment* phase, loss has decreased by a constant order.

$$\begin{aligned}
\|E\|_F =& \|\Delta Y - W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1\|_F \\
=& \|uu^\top(\Delta Y - W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1) vv^\top\|_F \\
&+ \|u_\perp u_\perp^\top(\Delta Y - W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1) v^\top\|_F \\
&+ \|uu^\top(\Delta Y - W_2 B_1 A_1 + B_2 A_2 W_1 + B_2 A_2 B_1 A_1) v_\perp v_\perp^\top\|_F \\
\leq& \|u(\sigma - \sigma_{W_2} g^\top B_1 A_1 v) v^\top\|_F + \|W_1\|\|B_2 A_2\|_F + \|B_2 A_2\|_F \|B_1 A_1\|_F \\
&+ \|W_2\|\|g_\perp B_1 A_1\|_F + \|W_1\|\|B_2 A_2\|_F + \|B_2 A_2\|_F \|B_1 A_1\|_F \\
&+ \|W_2\|\|B_1 A_1 v_\perp\|_F + \|W_1\|\|B_2 A_2\|_F + \|B_2 A_2\|_F \|B_1 A_1\|_F \\
=& \left(1 - \frac{(1-\delta_w)}{4}\right)\sigma + 3\|W_1\|\|B_2 A_2\|_F + 3\|B_2 A_2\|_F \|B_1 A_1\|_F + \|W_2\|\|g_\perp B_1 A_1\|_F + \|W_2\|\|B_1 A_1 v_\perp\|_F \\
\leq& \left(1 - \frac{(1-\delta_w)}{4}\right)\sigma + \frac{3}{2}\|W_1\|\alpha^{d_6} d_5^2 + \frac{3}{4}\alpha^{d_6} d_5^4 + \|W_2\|\|g_\perp B_1 A_1\|_F + \|W_2\|\|B_1 A_1 v_\perp\|_F
\end{aligned} \tag{154}$$

Moreover, by Lemma C.6, we can show that

$$\begin{aligned}
\|g_\perp B_1 A_1\|_F \leq \sqrt{r}\|g_\perp B_1 A_1\| \leq \sqrt{r\alpha^{d_7}}\|Z_1\|_F^2 \,, \\
\|B_1 A_1 v_\perp\|_F \leq \sqrt{r}\|B_1 A_1 v_\perp\| \leq \sqrt{r\alpha^{d_7}}\|Z_1\|_F^2 \,.
\end{aligned} \tag{155}$$

Therefore, one can show that there exists $\alpha^*(\|W_1\|, \|W_2\|, d_5, d_6, d_7)$ such that when $0 < \alpha \leq \alpha^*(\|W_1\|, \|W_2\|, d_5, d_6, d_7)$, we have

$$\begin{aligned}
\|E\|_F &\leq \left(1 - \frac{(1-\delta_w)}{4}\right)\sigma + \frac{3}{2}\|W_1\|\alpha^{d_6} d_5^2 + \frac{3}{4}\alpha^{d_6} d_5^4 + \|W_2\|\|g_\perp B_1 A_1\|_F + \|W_2\|\|B_1 A_1 v_\perp\|_F \\
&\leq \left(1 - \frac{(1-\delta_w)}{4}\right)\sigma + \frac{3}{2}\|W_1\|\alpha^{d_6} d_5^2 + \frac{3}{4}\alpha^{d_6} d_5^4 + 2\|W_2\|\sqrt{r\alpha^{d_7}} d_5^4 \\
&\leq \left(1 - \frac{(1-\delta_w)}{8}\right)\sigma \,.
\end{aligned} \tag{156}$$

Moreover, since the loss in non-increasing when trained under GF, we can see that $\|E(t)\|_F \leq \left(1 - \frac{(1-\delta_w)}{8}\right)\sigma$ for all $t \geq T_1$. In the remaining of the proof, we show that $\|E(t)\|_F \leq \left(1 - \frac{(1-\delta_w)}{8}\right)\sigma$ induces an lower bound

on $\|B_1A_1\|_F$.

$$\begin{aligned}
\|E\|_F =&\|\Delta Y - W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1\|_F\\
=&\|uu^\top(\Delta Y - W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1)vv^\top\|_F\\
&+\|u_\perp u_\perp^\top(\Delta Y - W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1)v^\top\|_F\\
&+\|uu^\top(\Delta Y - W_2B_1A_1 + B_2A_2W_1 + B_2A_2B_1A_1)v_\perp v_\perp^\top\|_F\\
\geq&\|u(\sigma - \sigma_{W_2}g^\top B_1A_1v)v^\top\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|g_\perp B_1A_1\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|B_1A_1v_\perp\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
\geq&\sigma - \|u\sigma_{W_2}g^\top B_1A_1vv^\top\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|g_\perp B_1A_1\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|B_1A_1v_\perp\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\,.
\end{aligned}\tag{157}$$

Therefore,

$$\begin{aligned}
\sigma_{W_2}g^\top B_1A_1v \geq&\ \sigma - \|E\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|g_\perp B_1A_1\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
&-\|W_2\|\|B_1A_1v_\perp\|_F - \|W_1\|\|B_2A_2\|_F - \|B_2A_2\|_F\|B_1A_1\|_F\\
\geq&\ \frac{(1-\delta_w)\sigma}{8} - \frac{3}{2}\|W_1\|\alpha^{d_6}d_5^2 - \frac{3}{4}\alpha^{d_6}d_5^4 - 2\|W_2\|\sqrt{r\alpha^{d_7}}d_5^4\\
\geq&\ \frac{(1-\delta_w)\sigma}{16}\,,
\end{aligned}\tag{158}$$

where the last line follows when α is sufficiently small, one can have

$$\frac{3}{2}\|W_1\|\alpha^{d_6}d_5^2 + \frac{3}{4}\alpha^{d_6}d_5^4 + 2\|W_2\|\sqrt{r\alpha^{d_7}}d_5^4 \leq \frac{(1-\delta_w)\sigma}{16}\,.\tag{159}$$

Finally,

$$\|\gamma_1^\top Z_1\|_F^2 = \|g^\top B_1 + v^\top A_1^\top\|^2 \geq \|g^\top B_1\|^2 + \|g^\top A_1^\top\|^2 \geq 2\|g^\top B_1A_1v\|\,.\tag{160}$$

Thus, $\|\gamma_1^\top Z_1\|_F^2 \geq 2g^\top B_1A_1v \geq \frac{(1-\delta_w)\sigma}{8}$. □

E.11 Proof of Lemma D.3

Proof. We first study $\langle \frac{\partial L_S}{\partial B_1}, \frac{\partial L_N}{\partial B_1}\rangle$.

$$\begin{aligned}
\langle \frac{\partial L_S}{\partial B_1}, \frac{\partial L_N}{\partial B_1}\rangle =&\big\langle (W_2+B_2A_2)^\top uu^\top Evv^\top A_1^\top, (W_2+B_2A_2)^\top uu^\top Ev_\perp v_\perp^\top A_1^\top\big\rangle\\
&+\big\langle (W_2+B_2A_2)^\top uu^\top Evv^\top A_1^\top, (W_2+B_2A_2)^\top u_\perp u_\perp^\top EA_1^\top\big\rangle\\
=&\mathrm{Tr}\big((W_2+B_2A_2)^\top uu^\top Evv^\top A_1^\top A_1v_\perp v_\perp^\top F^\top uu^\top(W_2+B_2A_2)\big)\\
&+\mathrm{Tr}\big((W_2+B_2A_2)^\top uu^\top Evv^\top A_1^\top A_1Fu_\perp u_\perp^\top(W_2+B_2A_2)\big)\\
=&u^\top Ev\Big\{\mathrm{Tr}\big((W_2+B_2A_2)^\top uv^\top A_1^\top A_1v_\perp v_\perp^\top F^\top uu^\top(W_2+B_2A_2)\big)\\
&+\mathrm{Tr}\big((W_2+B_2A_2)^\top uv^\top A_1^\top A_1F^\top u_\perp u_\perp^\top(W_2+B_2A_2)\big)\Big\}
\end{aligned}\tag{161}$$

Therefore, one has

$$\begin{aligned}
&\left|\langle\frac{\partial L_S}{\partial B_1},\frac{\partial L_N}{\partial B_1}\rangle\right| \\
=&\Big|u^\top Ev\Big\{\mathrm{Tr}\big((W_2+B_2A_2)^\top uv^\top A_1^\top A_1 v_\perp v_\perp^\top F^\top uu^\top (W_2+B_2A_2)\big) \\
&+\mathrm{Tr}\big((W_2+B_2A_2)^\top uv^\top A_1^\top A_1 F^\top u_\perp u_\perp^\top (W_2+B_2A_2)\big)\Big\}\Big| \\
\le&\sqrt{2L_S}\Big\{\big\|u^\top (W_2+B_2A_2)(W_2+B_2A_2)^\top u_\perp\big\|_F\times\|u^\top Fv_\perp\|_F\times\|v^\top A_1^\top A_1 v_\perp\|_F \\
&+\big\|u^\top (W_2+B_2A_2)(W_2+B_2A_2)^\top u_\perp\big\|_F\times\|u_\perp^\top F\|_F\times\|v^\top A_1^\top A_1\|_F\Big\}
\end{aligned}\tag{162}$$

Then, we provide bounds for each term on the RHS of the above equation.

First, we can upper bound

$$\begin{aligned}
\big\|u^\top (W_2+B_2A_2)(W_2+B_2A_2)^\top u_\perp\big\|_F &\le \big\|(W_2+B_2A_2)(W_2+B_2A_2)^\top\big\|_F \\
&= \|W_2+B_2A_2\|_F^2 \\
&\le 2\big(\|W_2\|_F^2+\|B_2A_2\|_F^2\big) \\
&\le 2\|W_2\|_F^2+2\|Z_2\|_F^2 \\
&\le 2\|W_2\|_F^2+2\alpha^{d_6}d_5^2\,.
\end{aligned}\tag{163}$$

Second,

$$\|v^\top A_1^\top A_1 v_\perp\|_F, \|v^\top A_1^\top A_1\|_F \le \|A_1\|_F^2\,.\tag{164}$$

Last,

$$\begin{aligned}
\|u^\top Fv_\perp\|_F &\le \|Fv_\perp\|_F \\
&= \|(W_2B_1A_1+B_2A_2W_1+B_2A_2B_1A_1)v_\perp\|_F \\
&\le \|W_2\|\|B_1A_1v_\perp\|_F+\|B_2A_2\|_F\times\big(\|W_1\|+\|B_1A_1\|_F\big) \\
&\le \|W_2\|\sqrt{\alpha^{d_7}}\|Z_1\|_F^2+\frac{1}{2}\|Z_2\|_F^2\times\big(\|W_1\|+\frac{1}{2}\|Z_1\|_F^2\big) \\
&\le \|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\,.
\end{aligned}\tag{165}$$

Similarly, one can show $\|u_\perp^\top F\|_F \le \|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})$. Combine these results together, one can show

$$\begin{aligned}
&\left|\langle\frac{\partial L_S}{\partial B_1},\frac{\partial L_N}{\partial B_1}\rangle\right| \\
\le&\sqrt{2L_S}\Big\{\big\|u^\top (W_2+B_2A_2)(W_2+B_2A_2)^\top u_\perp\big\|_F\times\|u^\top Fv_\perp\|_F\times\|v^\top A_1^\top A_1 v_\perp\|_F \\
&+\big\|u^\top (W_2+B_2A_2)(W_2+B_2A_2)^\top u_\perp\big\|_F\times\|u_\perp^\top F\|_F\times\|v^\top A_1^\top A_1\|_F\Big\} \\
\le&2\sqrt{2L_S}\big(2\|W_2\|_F^2+2\alpha^{d_6}d_5^2\big)\times\|A_1\|_F^2\times\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right).
\end{aligned}\tag{166}$$

Similarly, we can also show

$$\begin{aligned}
&\left|\langle\frac{\partial L_S}{\partial A_1},\frac{\partial L_N}{\partial B_1}\rangle\right| \\
\le&2\sqrt{2L_S}\big(2\|W_2\|_F^2+2\alpha^{d_6}d_5^2\big)\times\|B_1\|_F^2\times\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right).
\end{aligned}\tag{167}$$

Thus, we conclude

$$\left|\langle\frac{\partial L_S}{\partial A_1},\frac{\partial L_N}{\partial A_1}\rangle\right|+\left|\langle\frac{\partial L_S}{\partial B_1},\frac{\partial L_N}{\partial B_1}\rangle\right|\leq 2\sqrt{2L_S}\big(2\|W_2\|_F^2+2\alpha^{d_6}d_5^2\big)\times d_5^2\times\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right) \tag{168}$$

Next, we study $\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle$.

$$\begin{aligned}\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle=&\langle uu^\top Evv^\top(W_1+B_1A_1)^\top A_2^\top, uu^\top Ev_\perp v_\perp^\top(W_1+B_1A_1)^\top A_2^\top\rangle\\&+\langle uu^\top Evv^\top(W_1+B_1A_1)^\top A_2^\top, u_\perp u_\perp^\top E(W_1+B_1A_1)^\top A_2^\top\rangle\\=&u^\top Ev\bigg\{\mathrm{Tr}\bigg(v^\top(W_1+B_1A_1)^\top A_2^\top A_2(W_1+B_1A_1)v_\perp v_\perp^\top F^\top uu^\top\bigg)\\&+\mathrm{Tr}\bigg(v^\top(W_1+B_1A_1)^\top A_2^\top A_2(W_1+B_1A_1)F^\top u_\perp u_\perp^\top\bigg)\bigg\}.\end{aligned} \tag{169}$$

Thus, one can show that

$$\left|\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle\right|\leq\sqrt{2L_S}\bigg\{\big\|W_1+B_1A_1\big\|_F^2\times\|u^\top Fv_\perp\|_F\times\|A_2\|_F^2+\big\|W_1+B_1A_1\big\|_F^2\times\|u_\perp^\top F\|_F\times\|A_2\|_F^2\bigg\}. \tag{170}$$

We apply (165) to the above inequality

$$\begin{aligned}\left|\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle\right|&\leq\sqrt{2L_S}\bigg\{\big\|W_1+B_1A_1\big\|_F^2\times\|u^\top Fv_\perp\|_F\times\|A_2\|_F^2+\big\|W_1+B_1A_1\big\|_F^2\times\|u_\perp^\top F\|_F\times\|A_2\|_F^2\bigg\}\\&\leq 4\sqrt{2L_S}(\|W_1\|_F^2+\frac{1}{2}\|Z_1\|_F^2)\|A_2\|_F^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right).\end{aligned} \tag{171}$$

Similarly, we can also show that

$$\left|\langle\frac{\partial L_S}{\partial A_2},\frac{\partial L_N}{\partial A_2}\rangle\right|\leq 4\sqrt{2L_S}(\|W_1\|_F^2+\frac{1}{2}\|Z_1\|_F^2)\|B_2\|_F^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right). \tag{172}$$

Add them together, and one has

$$\left|\langle\frac{\partial L_S}{\partial A_2},\frac{\partial L_N}{\partial A_2}\rangle\right|+\left|\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle\right|\leq 4\sqrt{2L_S}(\|W_1\|_F^2+\frac{1}{2}\|Z_1\|_F^2)\alpha^{d_6}d_5^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right). \tag{173}$$

Based on (168) and (173), we conclude

$$\begin{aligned}&\left|\langle\frac{\partial L_S}{\partial A_1},\frac{\partial L_N}{\partial A_1}\rangle\right|+\left|\langle\frac{\partial L_S}{\partial B_1},\frac{\partial L_N}{\partial B_1}\rangle\right|+\left|\langle\frac{\partial L_S}{\partial A_2},\frac{\partial L_N}{\partial A_2}\rangle\right|+\left|\langle\frac{\partial L_S}{\partial B_2},\frac{\partial L_N}{\partial B_2}\rangle\right|\\\leq&2\sqrt{2L_S}\big(2\|W_2\|_F^2+2\alpha^{d_6}d_5^2\big)\times d_5^2\times\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right)\\&+4\sqrt{2L_S}(\|W_1\|_F^2+\frac{1}{2}\|Z_1\|_F^2)\alpha^{d_6}d_5^2\left(\|W_2\|\sqrt{\alpha^{d_7}}d_5^2+\frac{1}{2}\alpha^{d_6}d_5^2\times(\|W_1\|+\frac{d_5^2}{2})\right).\end{aligned} \tag{174}$$

□

F Proof of Theorem 4.1

Theorem F.1. *Let* $\delta_w^{(i)}=\frac{\sigma_{W_2}^{(i)}}{\sigma_{W_1}^{(i)}},\ell^{(i)}(t)=\frac{1}{2}\big(\sigma_{\Delta Y}^{(i)}-\sigma_{\bar{F}}^{(i)}\big)^2,z_1^{(i)}=(\sigma_{A_1}^{(i)})^2+(\sigma_{B_1}^{(i)})^2$ *(respectively* $z_2^{(i)}$*). In the case where* $\delta_w^{(i)}\neq 1$*, we assume* $\delta_w^{(i)}<1$ *WLOG, then the learning dynamics has two phases which can be separated by* $T_1^{(i)}=\frac{2}{(3+\delta_w^{(i)})\sigma\sigma_{W_2}^{(i)}}\log\left(\frac{(1-\delta_w^{(i)})\sigma_{\Delta Y}^{(i)}}{8\sigma_{W_2}^{(i)}z_1^{(i)}(0)}\right)$

1. Growth of Norm and Imbalance: $\forall t \le T_1^{(i)}$

$$\frac{d}{dt}\log z_1^{(i)} \ge \frac{3+\delta_w^{(i)}}{2}\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}\,,$$
$$\frac{d}{dt}\log\left(\frac{z_1^{(i)}}{z_2^{(i)}}\right) \ge \frac{3(1-\delta_w^{(i)})}{2}\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}\,. \tag{175}$$

2. Local Convergence: for $\forall t \ge T_1^{(i)}$*, the loss converges linearly*

$$\ell^{(i)}(t) \le \exp\left(-\frac{(1-\delta_w^{(i)})\sigma_{\Delta Y}^{(i)}\sigma_{W_2}^{(i)}(t-T_1)}{8}\right)\ell^{(i)}(T_1)\,.$$

Proof. Under spectral initialization, the learning dynamics of LoRA weights can be decoupled to several scalar dynamics. WLOG, we prove the learning dynamics when $i = 1$ as an example, and assume $\delta_w^{(i)} < 1$. In the rest of the proof, we will omit the superscript (i) for convenience. Throughout the paper, we will use $e = \sigma_{\Delta Y} - \sigma_f$ to denote the residual.

We first argue that $\sigma_{B_1}\sigma_{A_1}, \sigma_{B_2}\sigma_{A_2}$ will be always positive during the training. We show this is true for $\sigma_{B_1}\sigma_{A_1}$ as an example. Similar argument holds for $\sigma_{B_2}\sigma_{A_2}$.

$$\frac{d}{dt}\sigma_{B_1}\sigma_{A_1} = (\sigma_{A_1}^2 + \sigma_{B_1}^2)\sigma_{W_2}e\,. \tag{176}$$

Notice $e(0) = \sigma_{\Delta Y}$ and $\sigma_{B_1}(0)\sigma_{A_1}(0) = 0$, thus, $\frac{d}{dt}\sigma_{B_1}\sigma_{A_1} > 0$ during the training until $e = 0$ which is a stationary point.

Growth of Norm and Imbalance phase. We first can see that

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix}\sigma_{B_1}\\ \sigma_{A_1}\end{pmatrix} &= \begin{pmatrix}0 & (\sigma_{W_2}+\sigma_{B_2}\sigma_{A_2})e\\ (\sigma_{W_2}+\sigma_{B_2}\sigma_{A_2})e & 0\end{pmatrix}\begin{pmatrix}\sigma_{B_1}\\ \sigma_{A_1}\end{pmatrix}\\ &= \begin{pmatrix}0 & \sigma_{W_2}\sigma_{\Delta Y}\\ \sigma_{W_2}\sigma_{\Delta Y} & 0\end{pmatrix}\begin{pmatrix}\sigma_{A_1}\\ \sigma_{B_1}\end{pmatrix} + \underbrace{\begin{pmatrix}0 & -\sigma_{W_2}\sigma_f + e\sigma_{B_2}\sigma_{A_2}\\ -\sigma_{W_2}\sigma_f + e\sigma_{B_2}\sigma_{A_2} & 0\end{pmatrix}}_{D_1}\begin{pmatrix}\sigma_{B_1}\\ \sigma_{A_1}\end{pmatrix}.\end{aligned} \tag{177}$$

Similarly, one can show that

$$\frac{d}{dt}\begin{pmatrix}\sigma_{B_2}\\ \sigma_{A_2}\end{pmatrix} = \begin{pmatrix}0 & \sigma_{W_1}\sigma_{\Delta Y}\\ \sigma_{W_1}\sigma_{\Delta Y} & 0\end{pmatrix}\begin{pmatrix}\sigma_{B_2}\\ \sigma_{A_2}\end{pmatrix} + \underbrace{\begin{pmatrix}0 & -\sigma_{W_1}\sigma_f + e\sigma_{B_1}\sigma_{A_1}\\ -\sigma_{W_1}\sigma_f + e\sigma_{B_1}\sigma_{A_1} & 0\end{pmatrix}}_{D_2}\begin{pmatrix}\sigma_{B_2}\\ \sigma_{A_2}\end{pmatrix} \tag{178}$$

For convenience, we define the following notation $h_1 = -\sigma_{W_2}\sigma_f + e\sigma_{B_2}\sigma_{A_2}, h_2 = -\sigma_{W_1}\sigma_f + e\sigma_{B_1}\sigma_{A_1}$. It is obvious that $|h_1| = \|D_1\|, |h_2| = \|D_2\|$.

Notice at initialization, $\|D_1\| + \|D_2\| \sim \mathcal{O}(\alpha^2)$. We first cut off the time when $\|D_1\| + \|D_2\| = 2\alpha$, denoted by $\hat{T}_1$, then we can show that the imbalance between z_1 and z_2 grows monotonically for all $0 \le t \le \hat{T}_1$.

$$\frac{d}{dt}\log(\frac{z_1}{z_2}) \ge 2(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} + 2(h_1 - h_2) \ge (1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}\,, \tag{179}$$

where the last inequality holds under the assumption that $\alpha \le (1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}$. We then characterize the time it takes for $\|D_1\| + \|D_2\|$ to reach 2α.

$$\begin{aligned}\|D_1\| &= |-\sigma_{W_2}\sigma_f + e\sigma_{B_2}\sigma_{A_2}|\\ &\le \sigma_{W_2}|\sigma_f| + \frac{1}{2}|e(0)| \times z_2\\ &\le \sigma_{W_2}\left(\sigma_{W_2}|\sigma_{B_1}\sigma_{A_1}| + \sigma_{W_1}|\sigma_{B_2}\sigma_{A_2}| + |\sigma_{B_1}\sigma_{A_1}| \times |\sigma_{B_2}\sigma_{A_2}|\right) + \frac{1}{2}|e(0)| \times z_2\\ &\le \frac{1}{2}\sigma_{W_2}^2 z_1 + \sigma_{W_1}\sigma_{W_2}z_2 + \frac{1}{4}\sigma_{W_2}z_1z_2 + \frac{|e(0)|}{2}z_2\end{aligned} \tag{180}$$

Similarly, one can show that

$$\|D_2\| \leq \frac{1}{2}\sigma_{W_1}^2 z_2 + \sigma_{W_2}\sigma_{W_1} z_1 + \frac{1}{4}\sigma_{W_1} z_1 z_2 + \frac{|e(0)|}{2} z_1 \,. \tag{181}$$

For $0 \leq t \leq \hat{T}_1$, we can show that the growth of z_1, z_2 is

$$\frac{d}{dt}\log z_1 \leq (2\sigma_{W_2}\sigma_{\Delta Y} + 2\|D_1\|) \leq 2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} \,, \tag{182}$$

$$\frac{d}{dt}\log z_2 \leq (2\sigma_{W_1}\sigma_{\Delta Y} + 2\|D_2\|) \leq 2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} \,. \tag{183}$$

Let $z_{\max} = \frac{1}{\alpha^2}\max(z_1(0), z_2(0))$. Therefore, one can show that

$$\|D_1\| + \|D_2\| \leq \Big(3\sigma_{W_2}^2 + |e(0)|\Big)\alpha^2 \exp(2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) z_{\max} + \frac{\alpha^4 z_{\max}^2}{2}\exp(4(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) \,. \tag{184}$$

We let the RHS of the above equation equal to α, and derive a lower bound on $\hat{T}_1$. Notice in this case, we can further upper bound the RHS as

$$\begin{aligned}
&\Big(3\sigma_{W_2}^2 + |e(0)|\Big)\alpha^2 \exp(2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) z_{\max} + \frac{\alpha^4 z_{\max}^2}{2}\exp(4(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) \\
&\leq 2\Big(3\sigma_{W_2}^2 + |e(0)|\Big)\alpha^2 \exp(2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) z_{\max} \,.
\end{aligned} \tag{185}$$

Let $2\Big(3\sigma_{W_2}^2 + |e(0)|\Big)\alpha^2 \exp(2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} t) z_{\max} = \alpha$, we can show that

$$\exp\Big(2(2-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}\hat{T}_1\Big) = \frac{1}{2\Big(3\sigma_{W_2}^2 + |e(0)|\Big)\alpha} \,. \tag{186}$$

Based on these, one can show that there exists constants β_4 such that $z_1 \geq z_2 \beta_4 \alpha^{-\frac{1-\delta_w}{4-2\delta_w}}$.

After $\|D_1\| + \|D_2\|$ has reached 2α, we then study when $\|D_1\|$ or $\|D_2\|$reach$\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$.

If $\|D_1\|$ and $\|D_2\|$ never reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$ for all $t \leq T_1$, then one can show that z_1 grows exponentially fast

$$\frac{d}{dt} z_1 \geq (2\sigma_{W_2}\sigma_{\Delta Y} - 2\|D_1\|) z_1 \geq \frac{(3+\delta_w)\sigma\sigma_{W_2}}{2} z_1 \,, \tag{187}$$

and it leads to the following lower bound on z_1

$$\begin{aligned}
&\log z_1(t) \geq \log z_1(0) + \frac{(3+\delta_w)\sigma\sigma_{W_2}}{2} t \\
\iff &z_1(t) \geq z_1(0)\exp\Big(\frac{(3+\delta_w)\sigma\sigma_{W_2}}{2} t\Big)
\end{aligned} \tag{188}$$

Therefore, in order for $z_1(t)$ to reach $\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}$, one needs at most time

$$\begin{aligned}
&z_1(0)\exp\Big(\frac{(3+\delta_w)\sigma\sigma_{W_2}}{2} t\Big) = \frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}} \\
\iff &T_1 = \frac{2}{(3+\delta_w)\sigma\sigma_{W_2}} \log\Big(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2} z_1(0)}\Big) \,.
\end{aligned} \tag{189}$$

On the other hand, when $\|D_1\|$ or $\|D_2\|$reach$\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$ happens before z_1 reaches $\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}$. This must happen between $\hat{T}_1$ and T_1. We consider the following two cases.

First, $\|D_1\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$ before $\|D_2\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$, denoted by T_1'. Notice for all $\hat{T}_1 \le t \le T_1'$,

$$\frac{d}{dt}\log(\frac{z_1}{z_2}) \ge 2(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2} - 2(\|D_1\| + \|D_2\|) \ge 0\,, \tag{190}$$

Thus, the imbalance between z_1 and z_2 persists.

Then, one can show that

$$\begin{aligned}
&\|D_1\| > \frac{(1-\delta_w)\sigma\sigma_{W_2}}{8} \\
\iff &|-\sigma_{W_2}\sigma_f + e\sigma_{B_2}\sigma_{A_2}| > \frac{(1-\delta_w)\sigma\sigma_{W_2}}{8} \\
\Rightarrow &\sigma_{W_2}^2\sigma_{B_1}\sigma_{A_1} > \frac{(1-\delta_w)\sigma\sigma_{W_2}}{8} - \frac{1}{2}z_2|e(0)| - \sigma_{W_2}\frac{z_2}{2}\left(\sigma_{W_1} + \frac{z_1}{2}\right) \\
\Rightarrow &\sigma_{W_2}^2\sigma_{B_1}\sigma_{A_1} > \frac{(1-\delta_w)\sigma\sigma_{W_2}}{8} - \frac{1}{2\beta_4}\alpha^{\frac{1-\delta_w}{4-2\delta_w}}\frac{(1-\delta_w)\sigma_{\Delta Y}}{16\sigma_{W_2}}\left(|e(0)| + \sigma_{W_1}\sigma_{W_2} - \frac{(1-\delta_w)\sigma_{\Delta Y}}{32}\right).
\end{aligned} \tag{191}$$

Notice $z_1 = \sqrt{(\sigma_{A_1}^2 - \sigma_{B_1}^2)^2 + 4\sigma_{A_1}^2\sigma_{B_1}^2}$, therefore

$$\begin{aligned}
\sigma_{W_2}^4 z_1^2 \ge &\frac{((1-\delta_w)\sigma\sigma_{W_2})^2}{16} \\
&+ \sigma_{W_2}^4(\sigma_{A_1}^2 - \sigma_{B_1}^2)^2 \\
&+ 4\alpha^{\frac{1-\delta_w}{2-\delta_w}}\left\{\frac{1}{2\beta_4}\frac{(1-\delta_w)\sigma_{\Delta Y}}{16\sigma_{W_2}}\left(|e(0)| + \sigma_{W_1}\sigma_{W_2} - \frac{(1-\delta_w)\sigma_{\Delta Y}}{32}\right)\right\}^2 \\
&- \frac{8}{2\beta_4}\alpha^{\frac{1-\delta_w}{4-2\delta_w}}\frac{(1-\delta_w)\sigma_{\Delta Y}}{16\sigma_{W_2}}\left(|e(0)| + \sigma_{W_1}\sigma_{W_2} - \frac{(1-\delta_w)\sigma_{\Delta Y}}{32}\right)\times\frac{(1-\delta_w)\sigma\sigma_{W_2}}{8}
\end{aligned} \tag{192}$$

Notice $\sigma_{A_1}^2 - \sigma_{B_1}^2$ is preserved during GF, and its initial value is of order α^2. Thus, when α is sufficiently, one can show that

$$z_1 \ge \frac{(1-\delta_w)\sigma}{8\sigma_{W_2}}\,. \tag{193}$$

Second, $\|D_2\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$ before $\|D_1\|$, denoted by T_1''. We first assume that $z_2 \le z_1\alpha^{h_3}$ in this case before T_1 where $d_3 > 0$ is independent of α. Notice when $\|D_1\| \le \frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$, z_1 continues to grow exponentially fast $\frac{d}{dt}\log z_1 \ge \frac{(3+\delta_w)\sigma\sigma_{W_2}}{2}$. Moreover, whenever $\|D_1\|$ reaches $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{4}$ before T_1, we can both show that z_1 will reach $\frac{(1-\delta_w)\sigma}{8\sigma_{W_2}}$ before T_1. Then, the only thing that needs to show is the imbalance between z_1 and z_2 persists before z_1 reaches $\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}$. We will show that in this case, when α is sufficiently, the imbalance between z_1 and z_2 persists. For convenience, let $p_1 = \sigma_{B_1}\sigma_{A_1}, p_2 = \sigma_{B_2}\sigma_{A_2}$.

$$\begin{aligned}
\frac{d}{dt}\log(\frac{z_1}{z_2}) &= 2(1-\delta_w)\sigma\sigma_{W_2} - 2h_1 + 2h_2 \\
&\ge \frac{7(1-\delta_w)\sigma\sigma_{W_2}}{4} - 2h_2 \\
&= \frac{7(1-\delta_w)\sigma\sigma_{W_2}}{4} - 2(\sigma_{\Delta Y} + 2\sigma_{W_1}\sigma_{W_2})p_1 + 2\sigma_{W_2}p_1^2 + 2p_2\left(\sigma_{W_1}^2 + 2p_1\sigma_{W_1} + p_1^2\right) \\
&\ge \frac{7(1-\delta_w)\sigma\sigma_{W_2}}{4} - 2(\sigma_{\Delta Y} + 2\sigma_{W_1}\sigma_{W_2})p_1 \\
\Rightarrow \log\left(\frac{z_1(t)}{z_2(t)}\right) &= \log\left(\frac{z_1(\hat{T}_1)}{z_2(\hat{T}_1)}\right) + \frac{7(1-\delta_w)\sigma\sigma_{W_2}(t-\hat{T}_1)}{4} - 2(\sigma_{\Delta Y} + 2\sigma_{W_1}\sigma_{W_2})\int_{\hat{T}_1}^t p_1(s)ds
\end{aligned} \tag{194}$$

Thus, as long as one can show that $\int_{T_1'}^t p_1(s)ds$ is bounded by a constant h_4 (independent of α). One can conclude that

$$\frac{z_1(t)}{z_2(t)} \ge \exp\left(-2h_4(\sigma_{\Delta Y} + 2\sigma_{W_1}\sigma_{W_2})\right)\left(\frac{z_1(\hat{T}_1)}{z_2(\hat{T}_1)}\right)\exp\left(\frac{7(1-\delta_w)\sigma\sigma_{W_2}(t-\hat{T}_1)}{4}\right) \tag{195}$$

It suffices to show that $\int_{\hat{T}_1}^{t} p_1(s)ds$ is bounded. We first give a detailed characterization of the growth speed of z_1

$$\begin{aligned}
\frac{d}{dt}z_1 &= 2e(\sigma_{W_2}+p_2)p_1 \\
&= 2(\sigma_{\Delta Y}-\sigma_{W_2}p_1-\sigma_{W_1}p_2-p_1p_2)(\sigma_{W_2}+p_2)p_1 \\
&= 2p_1\sigma_{W_2}\sigma_{\Delta Y}-2\sigma_{W_2}^2p_1^2-2p_2\big(\sigma_{\Delta Y}p_1+\sigma_{W_2}p_1^2+p_1\sigma_{W_1}\sigma_{W_2}+p_1p_2\sigma_{W_1}+p_1^2\sigma_{W_2}+p_1^2p_2\big) \\
&\geq 2p_1\sigma_{W_2}\sigma_{\Delta Y}-z_1\sigma_{W_2}^2p_1-z_2\big(\sigma_{\Delta Y}p_1+\sigma_{W_2}\frac{z_1^2}{4}+\frac{z_1}{2}\sigma_{W_1}\sigma_{W_2}+\frac{z_1z_2}{4}\sigma_{W_1}+\frac{z_1^2}{4}\sigma_{W_2}+\frac{z_1^3}{8}\alpha^{h_3}\big) \\
&\geq 2p_1\sigma_{W_2}\sigma_{\Delta Y}-z_1\sigma_{W_2}^2p_1-\alpha^{h_3}z_1\big(\sigma_{\Delta Y}p_1+\sigma_{W_2}\frac{z_1^2}{4}+\frac{z_1}{2}\sigma_{W_1}\sigma_{W_2}+\frac{z_1z_2}{4}\sigma_{W_1}+\frac{z_1^2}{4}\sigma_{W_2}+\frac{z_1^3}{8}\alpha^{h_3}\big)\,. \qquad (196)
\end{aligned}$$

Notice we have $z_1 \leq \frac{(1-\delta_w)\sigma\sigma_{W_2}}{8}$. Thus, we can show

$$2p_1\sigma_{W_2}\sigma_{\Delta Y}-z_1\sigma_{W_2}^2p_1 \geq 2p_1\sigma_{W_2}\sigma_{\Delta Y}-\sigma_{W_2}^2p_1\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}=\frac{(15+\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}{8}\,. \qquad (197)$$

Moreover, when α is sufficiently small, one can show that there exists a constant such that

$$\alpha^{h_3}z_1\big(\sigma_{\Delta Y}p_1+\sigma_{W_2}\frac{z_1^2}{4}+\frac{z_1}{2}\sigma_{W_1}\sigma_{W_2}+\frac{z_1z_2}{4}\sigma_{W_1}+\frac{z_1^2}{4}\sigma_{W_2}+\frac{z_1^3}{8}\alpha^{h_3}\big)\leq h_5\alpha^{h_3}\,. \qquad (198)$$

Therefore, one can show that

$$\begin{aligned}
&\frac{d}{dt}z_1 \geq \frac{(15+\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}{8}p_1-h_5\alpha^{h_3} \\
\Rightarrow& z_1(t)\geq z_1(\hat{T}_1)+\int_{\hat{T}_1}^{t}\frac{(15+\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}{8}p_1(s)ds-h_5\alpha^{h_3}(t-\hat{T}_1) \\
\Rightarrow& \int_{\hat{T}_1}^{t}p_1(s)ds\leq\frac{8}{(15+\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\left(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}+h_5\alpha^{h_3}\frac{2}{(3+\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}z_1(0)}\right)\right). \qquad (199)
\end{aligned}$$

where in the last inequality we use

$$\begin{aligned}
z_1(t)&\leq\frac{2}{(3+\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}z_1(0)}\right) \\
t-\hat{T}_1&\leq\frac{2}{(3+\delta_w)\sigma\sigma_{W_2}}\log\left(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}z_1(0)}\right)\Bigg). \qquad (200)
\end{aligned}$$

Therefore, one can see that if α is sufficiently small, we have

$$\int_{\hat{T}_1}^{t}p_1(s)ds\leq\frac{16}{(15+\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\left(\frac{(1-\delta_w)\sigma_{\Delta Y}}{8\sigma_{W_2}}\right). \qquad (201)$$

***Local convergence* phase.** To follow the standard technique to prove local convergence for GF. We first assume that throughout the training, we have $z_1\alpha^{h_6}\geq z_2, z_1\leq h_7$ where h_6, h_7 are positive constants and independent of α. In the end, we will provide expression for h_6, h_7.

$$\begin{aligned}
\dot{\ell}&=-\sum_{i=1}^{2}\left(\frac{d\ell}{d\sigma_{A_i}}\right)^2+\left(\frac{d\ell}{d\sigma_{B_i}}\right)^2 \\
&\leq-\left(\frac{d\ell}{d\sigma_{A_1}}\right)^2+\left(\frac{d\ell}{d\sigma_{B_1}}\right)^2 \\
&=-e^2(\sigma_{W_2}+\sigma_{B_2}\sigma_{A_2})^2(\sigma_{A_1}^2+\sigma_{B_1}^2) \\
&\leq-2\sigma_{W_2}^2\ell(\sigma_{A_1}^2+\sigma_{B_1}^2) \\
&\leq-4\sigma_{B_1}\sigma_{A_1}\sigma_{W_2}^2\ell\,. \qquad (202)
\end{aligned}$$

Therefore, it suffices to show that $\sigma_{B_1}\sigma_{A_1}$ has a uniform lower bound for all $t \geq T_1$. Recall in the end of *Growth of Norm and Imbalance* phase, we have proved that z_1 has reached $\frac{(1-\delta_w)\sigma\sigma_{W_2}}{8}$. Thus, we can show that

$$\begin{aligned}
|e| &= |\sigma_{\Delta Y} - \sigma_{W_2}p_1 - \sigma_{W_1}p_2 - p_1p_2| \\
&\leq |\sigma_{\Delta Y} - \sigma_{W_2}p_1| + p_2(p_1 + \sigma_{W_1}) \\
&\leq \frac{(7+\delta_w)\sigma_{\Delta Y}}{8} + \frac{1}{2}z_2(\frac{z_1}{2} + \sigma_{W_1}) \\
&\leq \frac{(7+\delta_w)\sigma_{\Delta Y}}{8} + \frac{1}{2}\alpha^{h_6}h_7(\frac{h_7}{2} + \sigma_{W_1}) \\
&\leq \frac{(15+\delta_w)\sigma_{\Delta Y}}{16}\,,
\end{aligned} \tag{203}$$

where the last inequality holds when α is small enough. Since in GF, the loss is non-increasing, we have $|e(t)| \leq \frac{(15+\delta_w)\sigma_{\Delta Y}}{16}$ for all $t \geq T_1$. Then, we show that one can show that

$$\begin{aligned}
&|e| \leq \frac{(15+\delta_w)\sigma_{\Delta Y}}{16} \\
\iff &|\sigma_{\Delta Y} - \sigma_{W_2}p_1 - \sigma_{W_1}p_2 - p_1p_2| \leq \frac{(15+\delta_w)\sigma_{\Delta Y}}{16} \\
\Rightarrow &|\sigma_{W_2}p_1 + \sigma_{W_1}p_2 + p_1p_2| \geq \frac{(1-\delta_w)\sigma_{\Delta Y}}{16} \\
\Rightarrow &\sigma_{W_2}p_1 \geq \frac{(1-\delta_w)\sigma_{\Delta Y}}{16} - \frac{z_2}{2}(\sigma_{W_2} + \frac{z_1}{2}) \\
\Rightarrow &\sigma_{W_2}p_1 \geq \frac{(1-\delta_w)\sigma_{\Delta Y}}{16} - \frac{\alpha^{h_6}h_7}{2}(\sigma_{W_2} + \frac{h_7}{2}) \\
\Rightarrow &\sigma_{W_2}p_1 \geq \frac{(1-\delta_w)\sigma_{\Delta Y}}{32} \\
\iff &p_1 \geq \frac{(1-\delta_w)\sigma_{\Delta Y}}{32\sigma_{W_2}}\,,
\end{aligned} \tag{204}$$

where the last inequality holds when α is sufficiently small. We apply the above lower bound on p_1 to (202)

$$\dot{\ell} \leq -4\sigma_{B_1}\sigma_{A_1}\sigma_{W_2}^2\ell \leq -\frac{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}{8}\ell\,. \tag{205}$$

Thus, we concluded $\ell(t) \leq \exp\left(-\frac{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}(t-T_1)}{8}\right)L(T_1)$.

Finally, we provide expressions for h_6, h_7.

$$\begin{aligned}
&|e| \leq |e(0)| = \sigma_{\Delta Y} \\
\iff &\sigma_{W_2}p_1 + \sigma_{W_1}p_2 + p_1p_2 \leq 2\sigma_{\Delta Y} \\
\Rightarrow &\sigma_{W_2}p_1 \leq 2\sigma_{\Delta Y} \\
\Rightarrow &p_1 \leq \frac{2\sigma_{\Delta Y}}{\sigma_{W_2}} \\
\Rightarrow &z_1^2 = \left(\sigma_{A_1}^2 - \sigma_{B_1}^2\right)^2 + 4p_1^2 \leq \left(\sigma_{A_1}^2 - \sigma_{B_1}^2\right)^2 + \frac{16\sigma_{\Delta Y}^2}{\sigma_{W_2}^2} \\
\Rightarrow &z_1 \leq \frac{8\sigma_{\Delta Y}}{\sigma_{W_2}} := h_7\,,
\end{aligned} \tag{206}$$

where in the second to last line, we use the fact that $\left(\sigma_{A_1}^2 - \sigma_{B_1}^2\right)^2$ is of order α^4 and one can choose α sufficently small to reach the last line. The same argument can prove $p_2 \leq \frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}$.

For h_6,

$$\begin{aligned}&\frac{d}{dt}z_1 = 2e(\sigma_{W_2}+p_2)p_1 \geq 0\\ &\Rightarrow z_1(t) \geq z_1(T_1)\,.\end{aligned} \tag{207}$$

On the other hand

$$\begin{aligned}&\frac{d}{dt}\log z_2 = \frac{2e(\sigma_{W_1}+p_1)p_2}{z_2} \leq e(\sigma_{W_1}+p_1) \leq \sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\\ \Rightarrow &\log z_2(t) - \log z_2(T_1) \leq \sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\int_{T_1}^{t}\exp\left(-\frac{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}(t-T_1)}{16}\right)dt\\ \Rightarrow &\log z_2(t) - \log z_2(T_1) \leq \sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\frac{16}{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\\ \Rightarrow &z_2(t) \leq z_2(T_1)\exp\left(\sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\frac{16}{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\right).\end{aligned} \tag{208}$$

Thus,

$$\begin{aligned}\frac{z_2(t)}{z_1(t)} &\leq \frac{z_2(\hat{T}_1)}{z_1(\hat{T}_1)}\exp\left(\sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\frac{16}{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\right)\\ &\leq \frac{1}{\beta_4}\alpha^{\frac{1-\delta_w}{4-2\delta_w}}\exp\left(\sqrt{2\ell}\left(\sigma_{W_2}+\frac{2\sigma_{\Delta Y}}{\sigma_{W_2}}\right)\frac{16}{(1-\delta_w)\sigma_{\Delta Y}\sigma_{W_2}}\right)\\ &\leq \alpha^{\frac{1-\delta_w}{2-\delta_w}}\,,\end{aligned} \tag{209}$$

where the last inequality holds when α is sufficiently small, and one can set $h_6 = \frac{1-\delta_w}{2-\delta_w}$. □

G Example of spectral initialization

In this section, we provide examples where methods built purely on pre-trained weights fail to fine-tune pre-trained models for MF, highlighting the importance of incorporating the fine-tuning target matrix when designing spectral initialization for LoRA. Assume that we have found a solution $W_2 = W_1 = \begin{pmatrix}10 & 0\\ 0 & 1\end{pmatrix}$ to a pretaining task of factorizing a target matrix $Y_{\text{pre}} = \begin{pmatrix}100 & 0\\ 0 & 1\end{pmatrix}$. Then, we are interested in solving Problem 2 with LoRA rank $r = 1$. The following theorem shows that either initializing LoRA weights using the top-1 or the bottom-1 singular space of W_1, W_1, there exists a Y_{ft} such that the model cannot converge to the target matrix of the fine-tuning task through minimizing Problem 2 using GF.

Theorem G.1. *WLOG, assume one initializes LoRA weights using the top-1 singular space of W_1, W_1, then let $Y_{\text{ft}} = \begin{pmatrix}100 & 0\\ 0 & 1\end{pmatrix}$. One can show that $L(t) = 1$ for $\forall t \geq 0$.*

Proof. As one initializes LoRA weights using the top-1 singular space of W_1, W_1, therefore one can assume at initialization, we have $A_i = a_i e_1^\top, B_i = b_i e_1$ where $b_1(0) = b_2(0) = 0$. Thus, one has

$$\dot{L}(0) = \frac{d}{dt}\frac{1}{2}\left\|\begin{pmatrix}b_1(0)a_1(0) + b_2(0)a_2(0) + b_2(0)a_2(0)b_1(0)a_1(0) & 0\\ 0 & 0\end{pmatrix}\right\|_F^2 \tag{210}$$

Since $b_1(0) = b_2(0) = 0$, one can see that the initialization of LoRA weights is at an stationary point and $\dot{L}(0) = 0$. Thus, $L(t) = L(0) = 1$ for $\forall t \geq 0$. □

H Experiments on Image Classification Tasks

H.1 Experiments on Matrix Factorization

In this section, we conduct additional experiments on LoRA applied to matrix factorization (MF) with varying imbalance levels, specifically $\delta_w \in \left\{\frac{1}{1.05^2}, \frac{1}{4}, \frac{1}{100}\right\}$. We also explore scenarios where $Y_{\text{ft}} = Y_{\text{pre}} + 5uv^\top$, where

(u, v) are the top and bottom singular vectors of Y_{pre}, respectively. Figures 3 and 4 illustrate that, across different imbalance levels and varying singular components of Y_{pre} that ΔY aligns with, smaller LoRA-based initialization leads to initial alignment, followed by growth in the norm of LoRA weights. Furthermore, smaller initialization scales consistently result in lower final loss. In contrast, for spectral initialization, the loss invariably converges to machine precision. Finally, although the Frobenius norm of $\Delta Y = 5uv^\top$ is the same in both cases, we observe that GF converges overall faster when (u, v) are the top singular components of Y_{pre} compared to when (u, v) are the bottom singular components. This is because the convergence rate shown in Theorem 3.1 is inversely proportional to $\frac{1}{\sigma\sigma_{W_2}}$. When (u, v) are the top singular components of Y_{pre}, the corresponding σ_{W_2} values are larger than in the other case, resulting in faster convergence. Thus, our theory effectively captures this phenomenon.

H.2 Experiments on Image Classification

In this section, we present additional experiments on fine-tuning ResNet, VIT, and VGG models pre-trained on ImageNet for MNIST and CIFAR10. For all models, we apply LoRA to the final layer, initializing B as zero matrices and A with entries drawn from $\mathcal{N}(0, 10^{-6})$. All models are trained using SGD with a step size of 0.1. To approximate gradient GF, it is common practice to choose a very small step size for SGD, typically 10^{-4}. However, for large-scale models, such small step sizes would result in prohibitively long training times. Therefore, we choose a relatively larger step size to accelerate training. Our goal is two-folded

- The evolution of LoRA weight alignment and norm during the early stages of training.
- The impact of initialization scale on the final training/test loss and accuracy.

H.2.1 Alignment in Early Stage of Training

In this section, we focus on the evolution of the alignment and the norm of the LoRA weights during the early stages of training.

We first introduce how we measure the alignments in this case. We consider the model as $f(A, B)$ where A, B are the LoRA weights. We use W to denote the pre-trained weights of the final layer, and use $\phi = W + AB$. Let $\ell(\cdot)$ be the loss function, and we can write the optimization problem of training these models as follows

$$\min_{A,B} \ell(f(A, B)) \,. \tag{211}$$

We consider solving the above problem using GF

$$\dot{A}(t) = -\frac{\partial \ell(t)}{\partial A}, \quad \dot{B}(t) = -\frac{\partial \ell(t)}{\partial B} \,. \tag{212}$$

For simplicity, we use $A(t), B(t)$ to denote the LoRA weights at time t, and $\ell(t)$ as a shorthand for $\ell(A(t), B(t))$. Moreover, we use $\phi(t)$ as a shorthand for $W + B(t)A(t)$. Then, we approximate the gradients of LoRA weights as below

$$\frac{\partial \ell(t)}{\partial A} = B(t)^\top \frac{\partial \ell(t)}{\partial \phi} \approx B(t)^\top \frac{\partial \ell(0)}{\partial \phi}, \quad \frac{\partial \ell(t)}{\partial B} = \frac{\partial \ell(t)}{\partial \phi} A(t)^\top \tag{213}$$

Notice that at initialization, $A(0)B(0) = 0$. Therefore, $\frac{\partial \ell(0)}{\partial \phi} = \frac{\partial \ell(0)}{\partial W}$, which represents the gradient of the loss of the pre-trained model evaluated on the fine-tuning dataset with respect to the last layer of the pre-trained model. Consequently, the initial learning dynamics can be simplified as

$$\frac{d}{dt} \begin{pmatrix} B(t) \\ A(t)^\top \end{pmatrix} \approx - \begin{pmatrix} 0 & \frac{\partial \ell(0)}{\partial W} \\ \frac{\partial \ell(0)}{\partial W}^\top & 0 \end{pmatrix} \begin{pmatrix} B(t) \\ A(t)^\top \end{pmatrix} \tag{214}$$

Let SVD of $\frac{\partial \ell(0)}{\partial W}$ be $U_W \Sigma_W V_W^\top$, then we say the left singular matrices of $\begin{pmatrix} B \\ A^\top \end{pmatrix}$ align with $U_{\text{target}} = \begin{pmatrix} U_W \\ -V_W \end{pmatrix}$ To measure the alignment of the left singular spaces of $Z(t) = \begin{pmatrix} B(t) \\ A(t)^\top \end{pmatrix}$ w.r.t.U_{target}. Let SVD of $Z(t)$ be

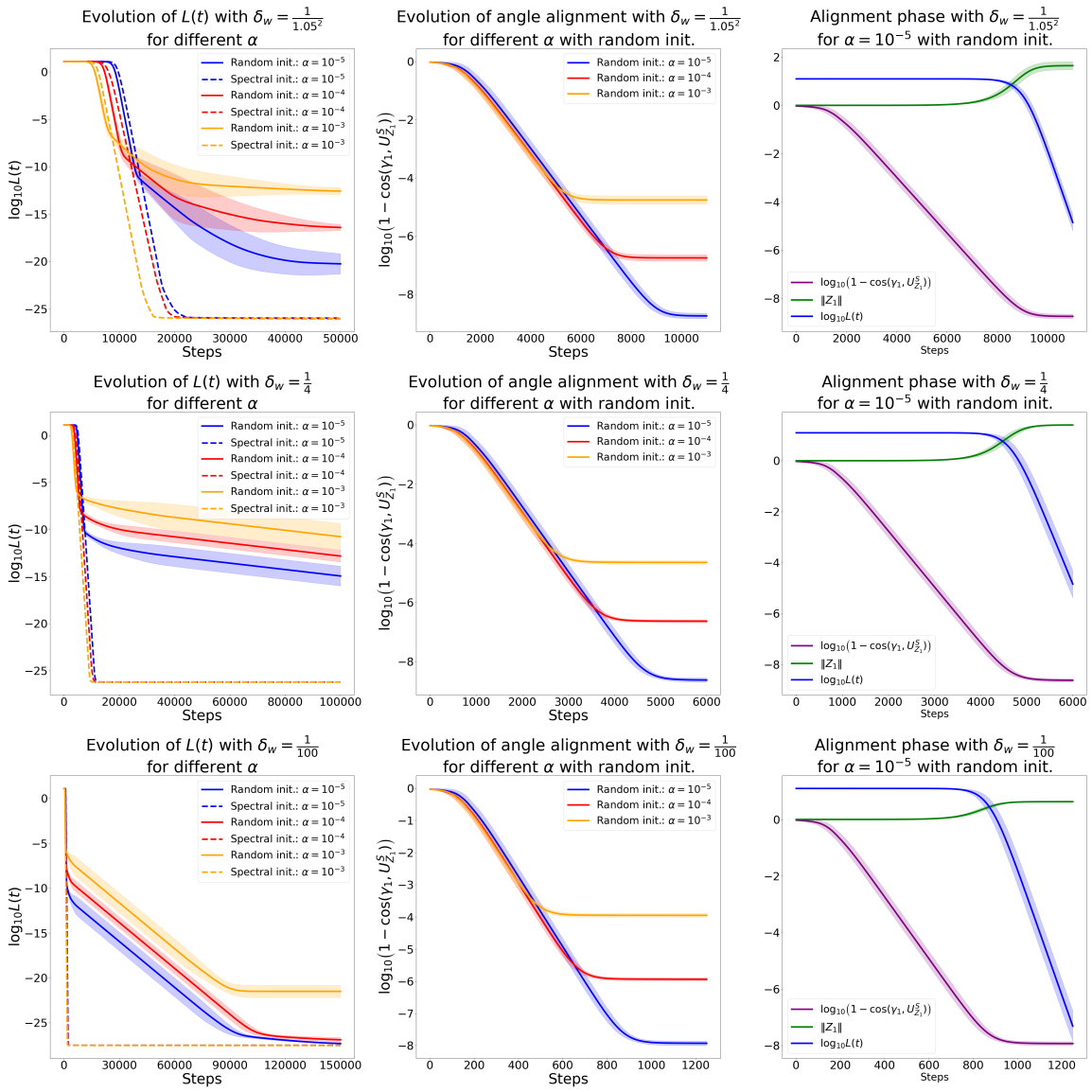

Figure 3: We simulate Problem 2 in the context of $\delta_w < 1$ using both small initialization (see §2) and small spectral initialization (see §4). We generate the data $Y_{\text{ft}} = Y_{\text{pre}} + 5uv^\top$ where u, v is the *top* principle component of Y_{pre}. Each simulation is repeated thirty times, with shaded regions representing one standard deviation above and below the mean (see §5.1 for details). The left column shows the evolution of the loss for different initialization scales α with small and spectral initialization. The middle column tracks the alignment quality between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and γ_1, measured by $\log_{10}(1 - \cos(\gamma_1, \boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)))$, where smaller values indicate better alignment. The right column focuses on small initialization with $\alpha = 10^{-5}$, illustrating how the reconstruction loss, alignment between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and γ_1, and $\|Z_1\|$ evolve during the *alignment* phase.

$U(t)\Sigma(t)V(t)^\top$, then we measure the following quantities, which is a classic metric to measure the alignment of two orthogonal matrices Chen and Chi (2013)

$$\frac{1}{\sqrt{2r}} \|U(t)U(t)^\top - U_{\text{target}} U_{\text{target}}^\top\|_F \,, \tag{215}$$

where $\frac{1}{\sqrt{2r}}$ is a normalizing constant that ensuring the above metric for alignment lies between zero and one. The smaller this number is, the better alignment is.

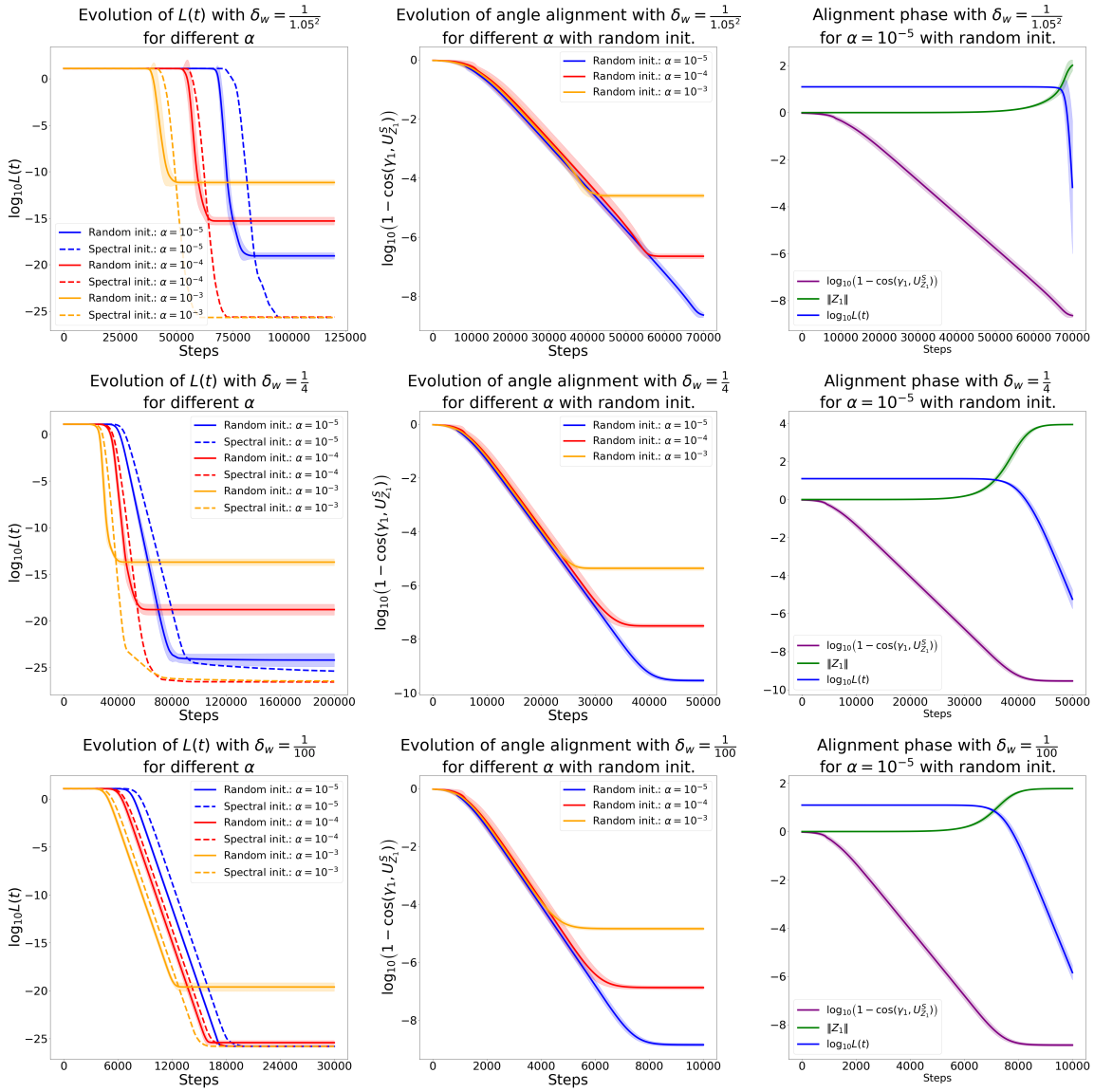

Figure 4: We simulate Problem 2 in the context of $\delta_w < 1$ using both small initialization (see §2) and small spectral initialization (see §4). We generate the data $Y_{\text{ft}} = Y_{\text{pre}} + 5uv^\top$ where u, v is the *bottom* principle component of Y_{pre}. Each simulation is repeated thirty times, with shaded regions representing one standard deviation above and below the mean (see §5.1 for details). The left column shows the evolution of the loss for different initialization scales α with small and spectral initialization. The middle column tracks the alignment quality between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and γ_1, measured by $\log_{10}(1 - \cos(\gamma_1, \boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}(t)))$, where smaller values indicate better alignment. The right column focuses on small initialization with $\alpha = 10^{-5}$, illustrating how the reconstruction loss, alignment between $\boldsymbol{U}_{\boldsymbol{Z_1}}^{\boldsymbol{S}}$ and γ_1, and $\|Z_1\|$ evolve during the *alignment* phase.

Figure 5 shows that, in the early stages of training, all models trained on both datasets exhibit strong alignment. However, the approximation in (213) is only valid when the LoRA weights remain close to zero. As training progresses, the LoRA weights deviate from zero, causing the approximation in (213) to lose accuracy. Consequently, the alignment between $U(t)$ and U_{target} ceases to improve.

H.2.2 Effect of Initial Std on Loss and Accuracy

In this section, we focus on the impact of initialization scale on the final training/test loss and accuracy.

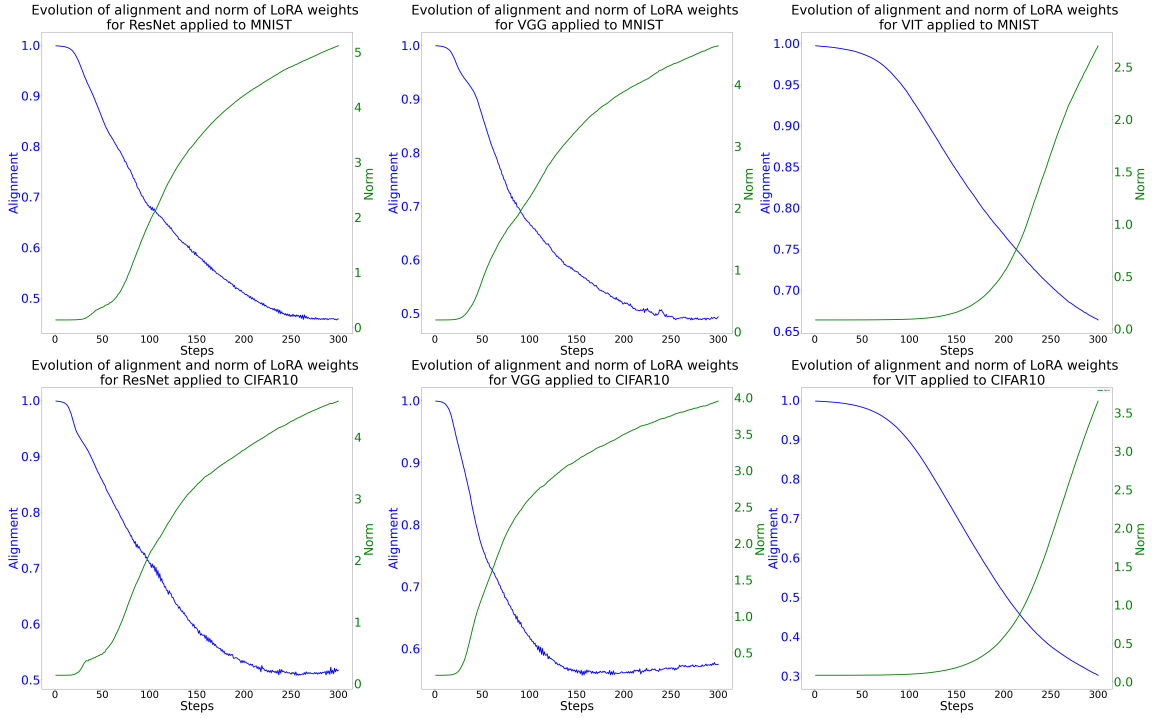

Figure 5: We run the experiments on fine-tuning ResNet, VIT and VGG pre-trained on ImageNet to MNIST and CIFAR10. We monitor the evolution of the alignment and norm of the LoRA weights in the early stage of training.

Figure 6 and Figure 7 show that for all models and dataset, smaller initialization leads to a lower final loss on both the training and test datasets. Moreover, it also results in higher accuracy on both the training and test datasets.

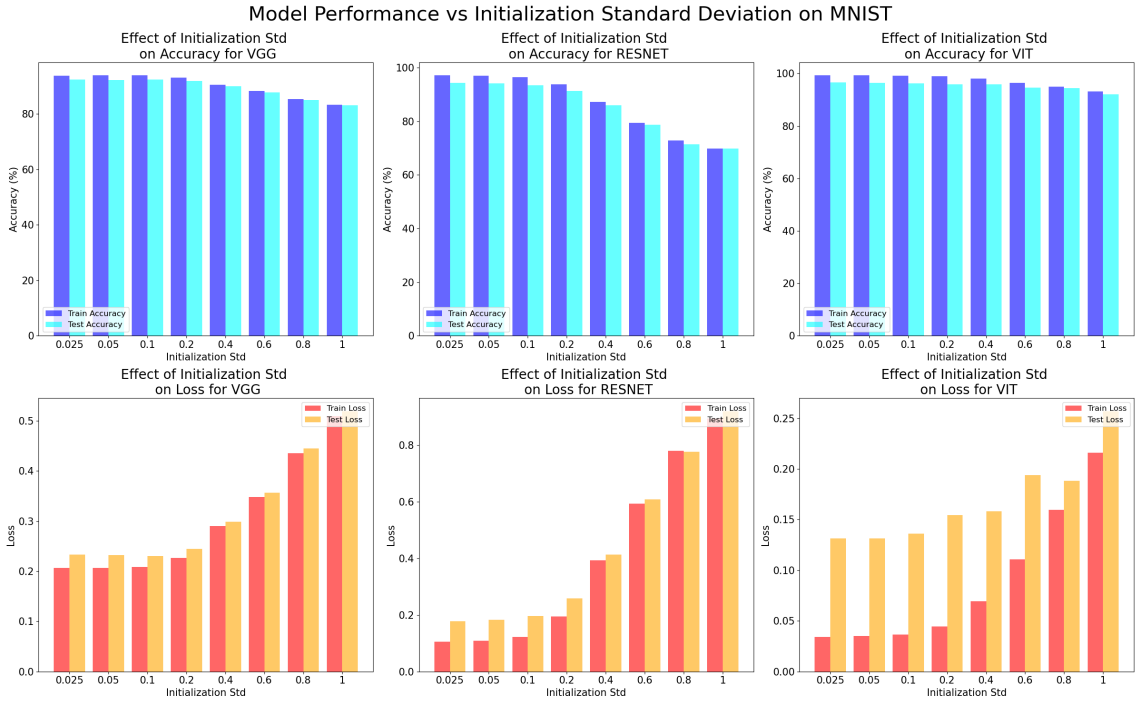

Figure 6: We run the experiments on fine-tuning ResNet, VIT and VGG pre-trained on ImageNet to MNIST. We monitor the how the scale of initialization affects the final training/test loss and accuracy.

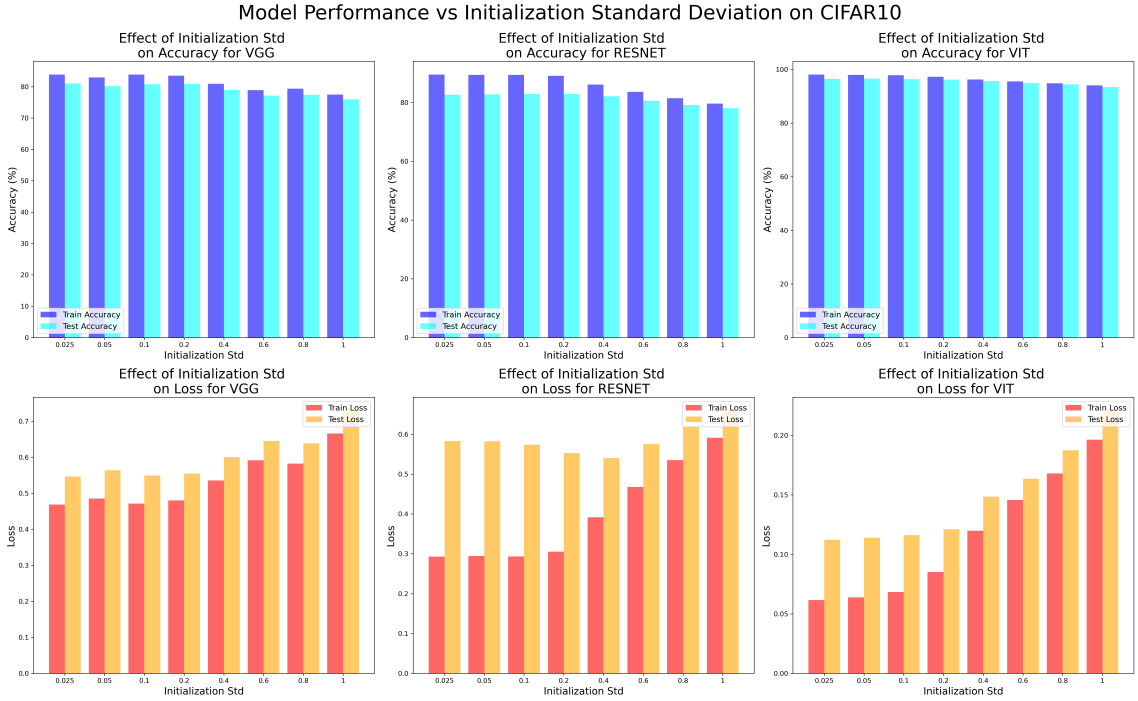

Figure 7: We run the experiments on fine-tuning ResNet, VIT and VGG pre-trained on ImageNet to CIFAR10. We monitor the how the scale of initialization affects the final training/test loss and accuracy.