# Uniform Spectral Growth under Factor-wise Gradient Orthogonalization in Low-Rank Matrix Factorization

**Changmin Kang**[*]
KAIST AI
cmkang8128@kaist.ac.kr

**Jihun Yun**[*]
KRAFTON
jihuny@krafton.com

**Baekrok Shin** **Yeseul Cho** **Chulhee Yun**
KAIST AI
{br.shin, cyseul, chulhee.yun}@kaist.ac.kr

## Abstract

Spectral gradient descent (SpecGD) orthogonalizes matrix parameter updates and has inspired practical optimizers such as Muon. However, the dynamics via such orthogonalizations remain poorly understood, especially in factorized parameterizations as orthogonalizing the updates of two matrices does not directly determine the singular value dynamics of their product. We analyze spectral gradient flow (SpecGF), a continuous-time analog of spectral gradient descent (SpecGD), for low-rank matrix factorization (MF) with orthogonalization applied to each factor gradient. Under small LoRA initialization and suitable smoothing, all active singular values asymptotically follow equal quadratic growth. Therefore, smaller target values are learned earlier, contrasting with the largest-first stepwise learning observed in standard gradient flow. We also prove that SpecGF converges to a global minimum from almost every initialization for which the factor norms remain bounded; with $\ell_2$ regularization, we obtain almost-sure global convergence for the regularized objective. Empirically, we observe that LoRA fine-tuning with orthogonalization-based optimizers including Muon exhibits qualitatively similar near-uniform growth in the product spectra of LoRA adapters.

## 1 Introduction

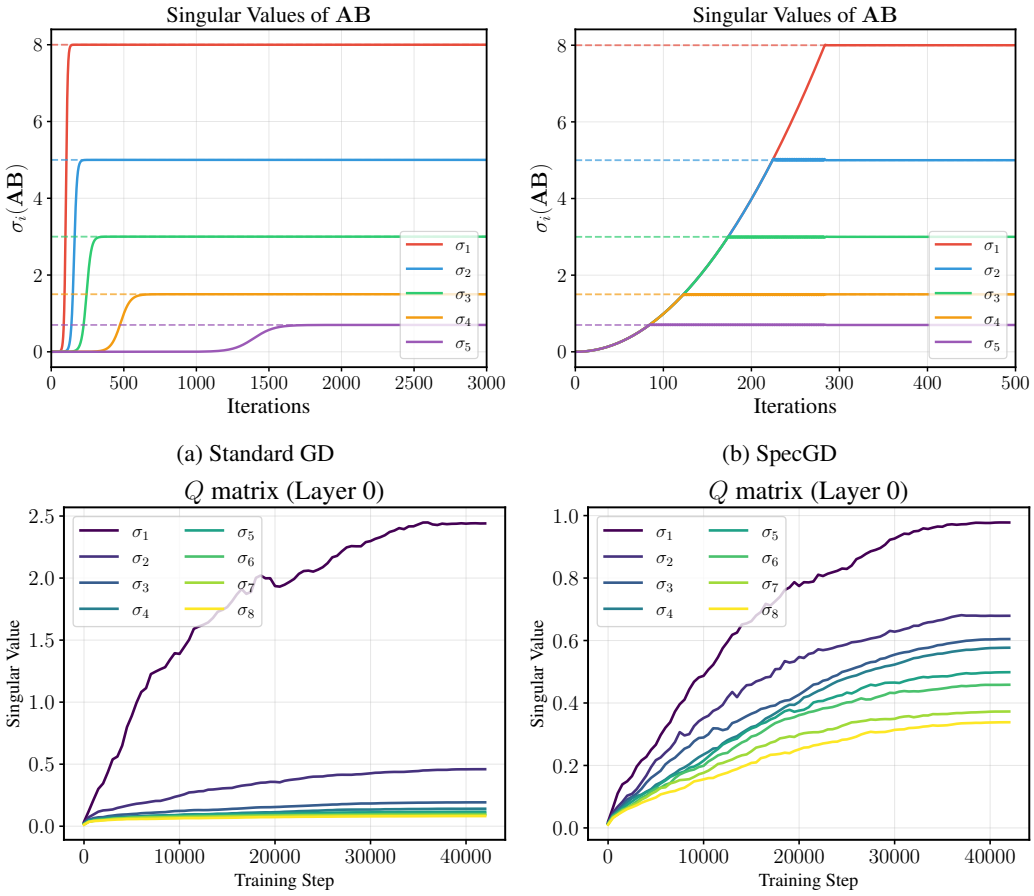
Classical optimization algorithms for training neural networks, such as stochastic gradient descent [48] and Adam(W) [36, 43], update network parameters in a coordinate-wise manner, ignoring the structure of parameters. In contrast, several modern approaches in deep learning exploit the structure of matrix parameters, focusing on *how* to update during training. The Muon optimizer [32] is a representative example. At each iteration $t$, Muon updates a weight matrix $\mathbf{W}$ for a loss function $f(\mathbf{W})$ as

$$\begin{aligned} \mathbf{M}_t &= \nabla f(\mathbf{W}_t) + \mu \mathbf{M}_{t-1}, \\ \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_t \mathcal{T}(\mathbf{M}_t), \end{aligned} \tag{1}$$

where $\mu \in [0, 1]$ tunes the momentum, $\eta_t$ is the step size, and $\mathcal{T}$ orthogonalizes the momentum (see Equation (4)). In practice, Newton-Schulz (NS) iteration approximates $\mathcal{T}$ by mapping most singular values of normalized $\mathbf{M}_t$ to an interval $[0.7, 1.3]$. If $\mu = 0$, the update follows spectral gradient descent (SpecGD) [6, 47]. Orthogonalized updates induce the spectrum of matrices to be more isotropic [60, 57]. By leveraging geometric structure, Muon has gained significant attention for training large language models, often outperforming Adam(W) [40, 60].

[*]Equal Contribution.

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




(a) Standard GD

(b) SpecGD

(c) RoBERTa, AdamW

(d) RoBERTa, Muon

Figure 1: Evolutions of the singular values of the product of factor matrices in low-rank matrix factorization (Figures 1a and 1b) and those of the LoRA $\mathbf{AB}$ adapter applied to the query projection matrix in the first self-attention layer (Figures 1c and 1d).

In single-matrix settings, exact orthogonalization can make the singular values evolve at equal speed by removing dependence on current singular-value scale. However, it is not evident whether the singular values would still evolve at the same rate if the parameter were a *product* of matrices. For a matrix product $\mathbf{W} = \mathbf{AB}$, the factor gradients are orthogonalized separately; thus, $\dot{\mathbf{W}} = \dot{\mathbf{A}}\mathbf{B} + \mathbf{A}\dot{\mathbf{B}}$ is not an orthogonalized update in the $\mathbf{W}$-space. This raises the following question:

*When orthogonalized updates are applied separately to low-rank matrix factors, what happens to the spectrum of the product?*

We provide rigorous analyses of the product dynamics induced by factor-wise orthogonalized updates in matrix factorization (MF). We model a low-rank MF objective through spectral gradient flow (SpecGF), a continuous-time model of SpecGD. Our analyses show that even though the factor gradients for $\mathbf{A}$ and $\mathbf{B}$ are orthogonalized separately, the singular values of the product $\mathbf{AB}$ can exhibit uniform growth, meaning that all the singular values grow at similar rates. We note that Vasudeva et al. [57] study a related equal-rate evolution of singular values for bilinear models, but their analysis relies on spectral initialization [64] that admits diagonalization, and thus the product dynamics decouple into independent scalar evolutions; see Section 2 for details. In contrast, we analyze low-rank MF where the product dynamics do not decouple into independent scalar evolutions.

We then use Low-Rank Adaptation (LoRA) [29] fine-tuning as a practical testbed for the same factor-wise orthogonalization mechanism. In LoRA, the update is parameterized as a low-rank product while a pretrained weight is fixed; if the factors are trained with Muon, orthogonalization is

applied separately to the factor updates. Thus, LoRA shares with our MF model the key structure: the spectrum of a learned product matrix is shaped by factor-wise orthogonalized updates.

We summarize our contributions as follows:

**Uniform Spectral Growth.** Even though SpecGF orthogonalizes the gradients of the factors $\mathbf{A}$ and $\mathbf{B}$ rather than their product, we show that the singular values of $\mathbf{AB}$ exhibit uniform quadratic growth under SpecGF and thus converge to the singular values of the target (Theorem 4.1). This holds for almost every initialization where $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{B}(0)$ is a small random matrix. The dynamics make smaller singular values converge before larger ones (Corollary 4.3), sharply contrasting with largest-first stepwise learning dynamics from standard gradient flow (GF) [3, 22]. This difference is visualized in Figures 1a and 1b.

**Convergence Guarantee.** We prove that SpecGF almost surely converges to global minima provided that the factor norms remain bounded (Theorem 5.4); with $\ell_2$ regularization, boundedness follows and SpecGF globally converges to a global minimum of the regularized objective for almost every initialization (Theorem 5.7). In addition, if SpecGF converges to a global minimum, the convergence could be exponential (Proposition 5.5), and every global minimum is Lyapunov stable.

**Empirical Evidence from LoRA Fine-Tuning.** We examine whether similar product-level spectral behavior appears in practical LoRA training. Figures 1c and 1d show that LoRA adapters trained under Muon exhibit near-uniform spectral growth, whereas AdamW shows less uniform growth, with the leading singular values growing faster. We further observe that the uniform growth phenomenon persists across other matrix optimizers and initialization choices (Section 6.2).

## 2 Related Work

**Orthogonalization-based Optimizers.** A line of research on the optimizers for matrix-valued parameters has gained attention for its empirical success. This includes Muon [32] and its variants, such as Dion [2] and MuonBP [34] that enhanced the efficiency, as well as preconditioned optimizers like Shampoo [25] and SOAP [58]. Muon normalizes the singular values of the matrix-valued updates, leading to faster convergence and higher performance [40, 60] compared to Adam [36].

Recent studies have begun to analyze orthogonalized gradients through the lens of spectral gradient descent (SpecGD), which can be viewed as Muon without momentum. Specifically, Fan et al. [17] study implicit bias and max-margin behavior under spectral descent, while Vasudeva et al. [57] analyze generalization benefits in imbalanced classification.

**Training Dynamics.** Matrix factorization (MF) is a fundamental and widely studied problem in optimization [24, 21, 3, 12, 52, 31, 35, 49] and machine learning [44, 56, 15, 51]. For linear models of depth $L$, or models with $L$ matrix factors, large singular values are typically learned earlier than smaller ones under standard GF [3, 22], aligning with broader notions of spectral bias [9, 62]. Gradient-based methods can also exhibit saddle-to-saddle trajectories in MF problems [30].

Orthogonalization-based optimizers such as Muon or SpecGD induce different dynamics. Vasudeva et al. [57] prove that, for linear and bilinear models under classification tasks with MSE loss, every singular value of the parameters increases at the same rate until saturation. Such uniform growth across all spectral components leads to better generalization by SpecGD than by gradient descent under data imbalance. However, their analysis relies on spectral initialization [64] constructed using population moment matrices; hence, the product dynamics reduce to decoupled scalar evolutions. Their result also assumes overparameterization: an intermediate width is larger than the number of spectral components, where each component corresponds to a class. Qiu et al. [46] empirically observe that Muon encourages a more uniform spectrum than AdamW in trainable matrices.

**Low-Rank Adaptation (LoRA).** Low-rank adaptation (LoRA) is an efficient approach for fine-tuning large language models (LLMs). This approach injects trainable low-rank factors into weight updates while keeping the base model fixed [29], motivated by the observation that fine-tuning updates often lie in a subspace of low intrinsic dimension [1]. LoRA reduces the number of trainable parameters from this low-rank structure without sacrificing much downstream performance.

Several recent studies report improved performance when applying Muon-style optimizers in LoRA fine-tuning [2, 61]. However, theoretical guarantees for convergence and training dynamics in this combined setting remain largely unexplored.

## 3 Theoretical Setup

**Notation.** We use bold uppercase letters (e.g., $\mathbf{X}$) for real matrices, bold lowercase letters (e.g., $\mathbf{x}$) for real vectors, and lowercase letters (e.g., $x$) for real scalars. Let $\mathbf{1} \in \mathbb{R}^d$ denote the all-ones vector and $\mathbf{I}_d$ the $d \times d$ identity matrix. We write $\mathsf{GL}(r)$ for the set of invertible $r \times r$ real matrices, and $\mathrm{Diag} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ for the diagonal operator. We denote the Frobenius, spectral, and nuclear norms by $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_2$, and $\|\cdot\|_*$, respectively. For matrices $\mathbf{M}$ and $\mathbf{N}$, we write $\langle \mathbf{M}, \mathbf{N}\rangle \coloneqq \mathrm{Tr}(\mathbf{M}\mathbf{N}^\top)$, $\mathbf{M}^\dagger$ for the pseudoinverse of $\mathbf{M}$, $\sigma_k(\mathbf{M})$ for the $k$-th largest singular value of $\mathbf{M}$, and $[k] \coloneqq \{1, \dots, k\}$. For a measurable function $f$ on an interval $I$, the essential supremum $\operatorname{ess\,sup}_{t\in I} f(t)$ is the infimum of all bounds $M$ satisfying $f(t) \le M$ for almost every $t \in I$ with respect to Lebesgue measure.

### 3.1 Problem Setup

We study the product-level dynamics through the MF objective

$$\min_{\mathbf{A}\in\mathbb{R}^{m\times r},\mathbf{B}\in\mathbb{R}^{r\times n}} \Big\{\mathcal{L}(\mathbf{A},\mathbf{B}) \coloneqq \tfrac{1}{2}\|\mathbf{A}\mathbf{B}-\mathbf{Y}\|_{\mathrm{F}}^2\Big\}, \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the low-rank factor matrices with $r \le \min\{m, n\}$ and $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is a target matrix. Let $r^\star = \mathrm{rank}(\mathbf{Y})$ and write the compact SVD $\mathbf{Y} = \mathbf{U}_{r^\star}\mathbf{\Sigma}\mathbf{V}_{r^\star}^\top$, where $\mathbf{\Sigma} \in \mathbb{R}^{r^\star\times r^\star}$ contains the nonzero singular values in non-increasing order and $\mathbf{U}_{r^\star}$, $\mathbf{V}_{r^\star}$ contain the singular vectors. If $\mathbf{A}$ and $\mathbf{B}$ depend on time $t$, we may write $\mathcal{L}(t) \coloneqq \mathcal{L}(\mathbf{A}(t), \mathbf{B}(t))$. We initialize $\mathbf{A}(0) = \mathbf{0}$ and sample each entry of $\mathbf{B}(0)$ i.i.d. from a Gaussian distribution, following a LoRA-style initialization [29, 26].

We analyze the continuous-time analog of SpecGD, *spectral gradient flow (SpecGF)*, defined by:

$$\dot{\mathbf{A}}(t) = -\mathcal{T}\left((\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})\mathbf{B}(t)^\top\right), \quad \dot{\mathbf{B}}(t) = -\mathcal{T}\left(\mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})\right), \tag{3}$$

where $\mathcal{T}$ is an orthogonalization operator applied to the matrix gradients. For a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ with compact SVD $\mathbf{U}_\mathbf{M}\mathbf{S}_\mathbf{M}\mathbf{V}_\mathbf{M}^\top$, the exact orthogonalization is written as

$$\mathcal{T}(\mathbf{M}) \coloneqq \left((\mathbf{M}\mathbf{M}^\top)^\dagger\right)^{1/2}\mathbf{M} = \mathbf{U}_\mathbf{M}\mathbf{V}_\mathbf{M}^\top, \quad \mathcal{T}(\mathbf{0}) \coloneqq \mathbf{0}. \tag{4}$$

Note that $\mathcal{T}$ normalizes all the nonzero singular values to 1. However, $\mathcal{T}$ is not analytic at rank-deficient points: when a singular value vanishes, arbitrarily small perturbations can introduce new singular directions, causing a discontinuous change in the output. To obtain an analytic vector field, we introduce a smoothed operator $\mathcal{T}_\beta$ with $\beta > 0$:

$$\mathcal{T}_\beta(\mathbf{M}) \coloneqq \left(\mathbf{M}\mathbf{M}^\top + \beta\mathbf{I}_m\right)^{-1/2}\mathbf{M}. \tag{5}$$

Adding $\beta\mathbf{I}$ makes the inverse square root well-defined, and makes SpecGF analytic everywhere; see Corollary D.2 for details. For a matrix $\mathbf{M}$ of rank $k$, we can write $\mathcal{T}_\beta(\mathbf{M})$ as

$$\mathcal{T}_\beta(\mathbf{M}) = \mathbf{U}_\mathbf{M}\mathbf{S}_\beta\mathbf{V}_\mathbf{M}^\top,\ \mathbf{S}_\beta \coloneqq \mathrm{Diag}\Big\{\tfrac{\sigma_1}{\sqrt{\sigma_1^2+\beta}}, \dots, \tfrac{\sigma_k}{\sqrt{\sigma_k^2+\beta}}\Big\},$$

when $\sigma_1, \dots, \sigma_k$ are the singular values of $\mathbf{M}$. Thus, $\mathcal{T}_\beta$ is a smooth relaxation of $\mathcal{T}$: each positive value is mapped to a value converging to 1 as $\beta \to 0$. Throughout the paper, "SpecGF with $\mathcal{S}$" refers to the flow where both $\dot{\mathbf{A}}$ and $\dot{\mathbf{B}}$ use the operator $\mathcal{S} \in \{\mathcal{T}, \mathcal{T}_\beta\}$ for orthogonalization.

*Remark* 3.1. The flow in Equation (3) is not equivalent to applying SpecGF directly to $\mathbf{W} = \mathbf{A}\mathbf{B}$. For $\mathcal{S} \in \{\mathcal{T}, \mathcal{T}_\beta\}$, the induced product dynamics are

$$\dot{\mathbf{W}}(t) = -\mathcal{S}((\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})\mathbf{B}(t)^\top)\mathbf{B}(t) - \mathbf{A}(t)\mathcal{S}(\mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})).$$

This flow is neither equal to $-\mathcal{S}(\mathbf{W}-\mathbf{Y})$ nor determined by $\mathbf{W}$ alone. Thus, product-level uniform growth cannot be obtained by directly applying single-matrix SpecGF analyses. We emphasize that the SVD of $\mathbf{Y}$ is introduced solely for analysis; the flow accesses $\mathbf{Y}$ only through the gradients.

## 4 Uniform Growth of Singular Values

In this section, we study the uniform spectral growth of the matrix product under SpecGF. For small LoRA initialization and sufficiently weak smoothing, we show that the singular values of

$\mathbf{W}(t) = \mathbf{A}(t)\mathbf{B}(t)$ exhibit approximately uniform quadratic growth before reaching their respective targets, so smaller singular values are learned earlier while larger ones continue to grow. When the factors are initialized along the target singular vectors, their dynamics reduce to scalar equations for each singular direction (Section A.2).

### 4.1 Uniform Growth from Small Initialization

We assume $r = r^\star$ and write $\mathbf{Y} = \mathbf{U}_r \mathbf{\Sigma} \mathbf{V}_r^\top$, where $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Here, $\sigma_i = \sigma_i(\mathbf{Y})$ are the target singular values. The target singular values need not be distinct. The factors are initialized as

$$\mathbf{A}(0) = \mathbf{0}, \qquad \mathbf{B}(0) = \gamma \mathbf{\Xi}, \qquad \mathbf{\Xi} \in \mathbb{R}^{r \times n}, \quad \gamma > 0. \tag{6}$$

Here, $\gamma > 0$ is the initialization scale. We fix $\mathbf{\Xi}$ such that $\mathbf{\Xi}\mathbf{V}_r$ is invertible. If the entries of $\mathbf{\Xi}$ are independent standard Gaussians, the rows of $\mathbf{\Xi}\mathbf{V}_r$ are independent standard Gaussian vectors in $\mathbb{R}^r$. Hence $\mathbf{\Xi}\mathbf{V}_r$ is invertible almost surely.

We write $\lambda_i(t) = \sigma_i(\mathbf{W}(t))$ for $i \in [r]$ and define

$$\tau_i = \sqrt{\sigma_i}, \qquad \tau_0 = \infty, \qquad \tau_{r+1} = 0, \qquad \mathbf{\Lambda}(t) = \mathrm{Diag}\big(\min\{t^2, \sigma_i\}\big)_{i=1}^r.$$

We call an index $i \leq k$ active on a non-empty interval $(\tau_{k+1}, \tau_k)$. All limits are taken along arbitrary sequences of positive values of the initialization scale $\gamma$ and the smoothing coefficient $\beta$ satisfying

$$\gamma \longrightarrow 0, \qquad \frac{\beta}{\gamma^2} \longrightarrow 0, \tag{7}$$

with $\mathbf{Y}$, the dimensions and $\mathbf{\Xi}$ fixed. The ratio $\beta/\gamma^2$ controls smoothing on the initial time scale $t \asymp \gamma$.

**Theorem 4.1.** *If $\mathbf{\Xi}\mathbf{V}_r$ is invertible, SpecGF with $\mathcal{T}_\beta$ satisfies the following under the initialization in Equation* (6) *and the limit in Equation* (7)*.*

$$\sup_{t \geq 0} \big\| \mathbf{W}(t) - \mathbf{U}_r \mathbf{\Lambda}(t) \mathbf{V}_r^\top \big\|_{\mathrm{F}} \longrightarrow 0. \tag{8}$$

*For every compact interval $I \subset (\tau_{k+1}, \tau_k)$ and every $i \leq k$,*

$$\operatorname*{ess\,sup}_{t \in I} |\dot{\lambda}_i(t) - 2t| \longrightarrow 0, \qquad \operatorname*{ess\,sup}_{t \in I} \left| \frac{\mathrm{d}}{\mathrm{d}t} \sqrt{\lambda_i(t)} - 1 \right| \longrightarrow 0. \tag{9}$$

*For $i \leq k$, the convergence $\lambda_i(t) \to t^2$ is uniform on $I$, so the active singular values are positive throughout $I$ for sufficiently small $\gamma$ and $\beta/\gamma^2$.*

Theorem 4.1 describes simultaneous quadratic growth until each singular value reaches its target. Smaller values are learned earlier while larger ones continue to grow. The product approximation holds uniformly for all $t \geq 0$, whereas the speed bounds apply on compact intervals away from initialization and the saturation times $\tau_i$. The use of essential supremum accounts for the almost-everywhere differentiability of ordered singular values.

**Proof outline.** We first rescale time by $\gamma$ to determine the initial direction of factor growth. After an orthogonal change of coordinates, the factors remain close to a common linear growth path before the first transition. We then use monotonicity of $\mathcal{T}_\beta$ to bound the error uniformly in $\beta$ across successive saturation times while allowing the saturated factor blocks to be unbalanced. The proof in Section B treats equal target singular values together and obtains Equation (9) by projecting the product velocity onto the active singular subspaces.

Theorem 4.1 gives a theoretical description of spectral growth, but evaluating the singular values throughout training requires an SVD at each step. We can instead track the diagonal entries in the target singular directions, which can be computed directly from the factors. For the target singular vectors $\mathbf{u}_i, \mathbf{v}_i$, we define

$$d_i(t) = \mathbf{u}_i^\top \mathbf{W}(t) \mathbf{v}_i = (\mathbf{u}_i^\top \mathbf{A}(t))(\mathbf{B}(t)\mathbf{v}_i).$$

The following lemma shows that $d_i(t)$ approximates $\lambda_i(t)$ and captures its growth under the same assumptions.

**Lemma 4.2.** *Under the assumptions of Theorem 4.1, for every $t \geq 0$,*

$$\max_{i\in[r]} |\lambda_i(t) - d_i(t)| \leq 2\big\|\mathbf{W}(t) - \mathbf{U}_r\mathbf{\Lambda}(t)\mathbf{V}_r^\top\big\|_{\mathrm{F}}.$$

*The diagonal approximation error therefore tends to zero uniformly in time. For each compact interval $I \subset (\tau_{k+1}, \tau_k)$ and $i \leq k$, $d_i > 0$ for sufficiently small $\gamma$ and $\beta/\gamma^2$ and*

$$\sup_{t\in I} \left| \frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} - 1 \right| \longrightarrow 0.$$

In our matrix factorization experiments, we show that $d_i(t)$ and $\lambda_i(t)$ exhibit similar growth dynamics (Section 6.1).

### 4.2 Smallest-first learning dynamics

The uniform growth rate in Theorem 4.1 leads naturally to smallest-first learning as smaller singular values approach their target values earlier. The following corollary formalizes the resulting order of learning.

**Corollary 4.3.** *Under the assumptions of Theorem 4.1, the crossing times for a fixed tolerance $0 < \varepsilon < \sigma_r$ are defined by*

$$T_i(\varepsilon) = \inf\{t \geq 0 : \lambda_i(t) \geq \sigma_i - \varepsilon\}.$$

*The crossing times satisfy $T_i(\varepsilon) \to \sqrt{\sigma_i - \varepsilon}$. If $\sigma_j < \sigma_i$, then $T_j(\varepsilon) < T_i(\varepsilon)$ for sufficiently small $\gamma$ and $\beta/\gamma^2$.*

SpecGF learns smaller singular values first, whereas standard gradient flow learns larger singular values first in the two-layer linear setting studied by Gidel et al. [22].

The asymptotic conclusions of Theorem 4.1 and Corollary 4.3 also yield guarantees when $\gamma$ and $\beta/\gamma^2$ are sufficiently small and positive. For Gaussian initialization, the following corollary shows that the growth behavior and learning order persist with high probability.

**Corollary 4.4.** *The entries of $\mathbf{\Xi}$ are independent standard Gaussians. For fixed tolerances $0 < \rho, \delta < 1$ and $0 < \varepsilon < \sigma_r$, we consider a non-empty compact set $I \subset (0, \tau_1) \setminus \{\tau_1, \ldots, \tau_r\}$. There exists $c_0 > 0$ such that the following statements hold jointly with probability at least $1 - \delta$ for each fixed pair $0 < \gamma \leq c_0$ and $0 < \beta \leq c_0\gamma^2$. The uniform product error in Equation (8) is at most $\rho$. For every $i$ with $I \cap (0, \tau_i)$ non-empty, the singular value $\lambda_i$ is positive on the intersection and both derivative errors in Equation (9) have essential supremum at most $\rho$ over the same set. The crossing times satisfy $T_j(\varepsilon) < T_i(\varepsilon)$ whenever $\sigma_j < \sigma_i$.*

The threshold $c_0$ may depend on the target and dimensions as well as the tolerances and $I$.

**Convergence and factor rank.** The growth analysis also yields convergence guarantees for fixed positive values of $\gamma$ and $\beta$. Under the initialization assumptions above, sufficiently small positive $\gamma$ and $\beta/\gamma^2$ ensure that the factors converge exponentially to finite limits satisfying $\mathbf{A}_\infty\mathbf{B}_\infty = \mathbf{Y}$. The precise statement and proof are given in Corollary B.4. When $r < r^\star$, the product cannot represent the full target and the best attainable approximation is $\mathbf{Y}_r$, formed from the target's $r$ largest singular values. With a spectral gap $\sigma_r > \sigma_{r+1}$, the uniform growth and smallest-first learning results extend to these $r$ components for almost every Gaussian initialization direction. For sufficiently small fixed positive values of $\gamma$ and $\beta/\gamma^2$, the factors also converge exponentially with $\mathbf{A}_\infty\mathbf{B}_\infty = \mathbf{Y}_r$, as detailed in Theorem C.1.

## 5 Convergence Analysis of SpecGF

Section 4 establishes uniform spectral growth and convergence for small LoRA initialization with sufficiently weak smoothing of $\mathcal{T}_\beta$. We now consider general initializations at fixed $\beta > 0$ and investigate the conditions where SpecGF converges to a global minimum. We begin by establishing the well-posedness and uniqueness of SpecGF solutions; see Proposition D.3 for details.

**Proposition 5.1** (Analyticity of $\mathcal{T}$ and $\mathcal{T}_\beta$)**.** *SpecGF with $\mathcal{T}_\beta$ is analytic everywhere and its solution uniquely exists. SpecGF with $\mathcal{T}$ is analytic and its solution uniquely exists when both factor gradients are full-rank.*

Since the vector field is analytic on its domain, the solution under SpecGF with $\mathcal{T}_\beta$ is unique. Moreover, the gradients with respect to $\mathbf{A}$ and $\mathbf{B}$ have rank at most $r$, and thus their orthogonalized gradients have Frobenius norm at most $\sqrt{r}$. Hence, the dynamics have bounded speed, and solutions do not blow up in finite time and globally exist. We then show the monotonic decrease of the loss:

**Lemma 5.2.** *For SpecGF with $\mathcal{T}$ and SpecGF with $\mathcal{T}_\beta$, $\mathcal{L}(t)$ is non-increasing. $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(t) = 0$ holds if and only if $\nabla_{\mathbf{A}}\mathcal{L}(t) = \nabla_{\mathbf{B}}\mathcal{L}(t) = 0$.*

See Lemmas D.4 and D.5 for the proof of Lemma 5.2. Using this monotonicity, Proposition 5.3 states that SpecGF with $\mathcal{T}_\beta$ either converges or diverges to infinity in norm; see Section D.2 for the proof.

**Proposition 5.3.** *Under SpecGF with $\mathcal{T}_\beta$, we have*

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(t) \le -\tfrac{1}{2r}\|\nabla\mathcal{L}(t)\|_{\mathrm{F}}\|(\dot{\mathbf{A}}(t), \dot{\mathbf{B}}(t))\|_{\mathrm{F}}. \tag{10}$$

*As a consequence of Equation* (10)*, $(\mathbf{A}(t), \mathbf{B}(t))$ either converges or diverges to infinity in norm.*

It remains to identify which critical points can be reached. The squared-loss factorization objective in Equation (2) has strict-saddle geometry: every critical point of $\mathcal{L}$ is either a global minimum or a strict saddle point [33, 4]. Moreover, every global minimum corresponds to the best rank-$r$ approximation of $\mathbf{Y}$. The initial conditions from which SpecGF with $\mathcal{T}_\beta$ converges to a strict saddle form a set of Lebesgue measure zero by Theorem D.13. Combining saddle avoidance with Proposition 5.3 gives the following theorem.

**Theorem 5.4.** *For almost all initial values $\mathbf{A}(0)$ and $\mathbf{B}(0)$, SpecGF with $\mathcal{T}_\beta$ either converges to a global minimum of $\mathcal{L}(\mathbf{A}, \mathbf{B})$ or $\|\mathbf{A}(t)\|_{\mathrm{F}}^2 + \|\mathbf{B}(t)\|_{\mathrm{F}}^2$ diverges to infinity.*

The remaining issue is boundedness. For vanilla GF, each factor is uniformly bounded due to a time-invariant $\mathbf{A}(t)^\top\mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top$ [3, 4]. However, such invariants are unknown for SpecGF—$\mathbf{A}(t)^\top\mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top$ fails to be invariant in general; see Proposition D.10. Although Corollary B.4 establishes boundedness and convergence in the small-initialization regime, boundedness for arbitrary initializations at fixed $\beta > 0$ remains unresolved. We conjecture that SpecGF has no nonconstant polynomial invariants in the factor entries. Propositions D.8 and D.9 establish this property in the scalar case, providing evidence for the conjecture.

**Convergence Rate.** Assume that SpecGF converges to a global minimum. We show that the loss satisfies the Morse-Bott property, where the optimal exponent in the Łojasiewicz gradient inequality (Lemma D.6) is $\frac{1}{2}$, leading to exponential convergence; see Section D.4 for the proof. The assumption of distinct singular values of a target matrix is considered in previous studies [22, 31].

**Proposition 5.5** (Informal)**.** *Assume that the singular values of $\mathbf{Y}$ are mutually distinct and $r \le r^\star$. If SpecGF with $\mathcal{T}_\beta$ converges to a global minimum of $\mathcal{L}$, then $\mathcal{L}(t)$ converges exponentially fast.*

*Remark* 5.6. If SpecGF with $\mathcal{T}$ converges, then it converges *in finite time*; see Proposition D.16 for details.

We also prove Lyapunov stability of global minima under SpecGF with $\mathcal{T}_\beta$, implying that any trajectories initialized sufficiently close to a global minimum remain nearby; see Proposition D.19.

**Global Convergence via Regularization.** Theorem 5.4 can be strengthened to state **global** convergence by adding an $\ell_2$ regularizer. This setting is motivated by the importance of norm control in practical Muon training. Recent studies report that large-scale Muon is sensitive to weight decay and that explicit norm control can improve training stability and performance [40, 45, 39, 50]. While practical Muon typically uses decoupled weight decay rather than the coupled $\ell_2$ regularization considered here, these observations motivate studying SpecGF under explicit norm control. Consider the regularized objective with $\lambda > 0$:

$$\mathcal{L}_\lambda(\mathbf{A}, \mathbf{B}) \coloneqq \tfrac{1}{2}\|\mathbf{A}\mathbf{B} - \mathbf{Y}\|_{\mathrm{F}}^2 + \tfrac{\lambda}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + \tfrac{\lambda}{2}\|\mathbf{B}\|_{\mathrm{F}}^2. \tag{11}$$

The regularizer makes the factor norms $\|\mathbf{A}(t)\|_{\mathrm{F}}$ and $\|\mathbf{B}(t)\|_{\mathrm{F}}$ uniformly bounded and, at global minima, induces soft-thresholding of the product spectrum with threshold $\lambda$ [56]. The strict-saddle

geometry and the avoidance of strict saddle points carry over assuming that all singular values of $\mathbf{Y}$ are distinct. Hence, regularized SpecGF with $\mathcal{T}_\beta$ cannot diverge to infinity and almost surely converges to global minima of the regularized objective; see Section D.6 for details.

**Theorem 5.7.** *If $r \leq r^\star$ and all singular values of $\mathbf{Y}$ are distinct, for almost all initial values, SpecGF with $\mathcal{T}_\beta$ under the regularized loss in Equation* (11) *converges to a global minimum.*

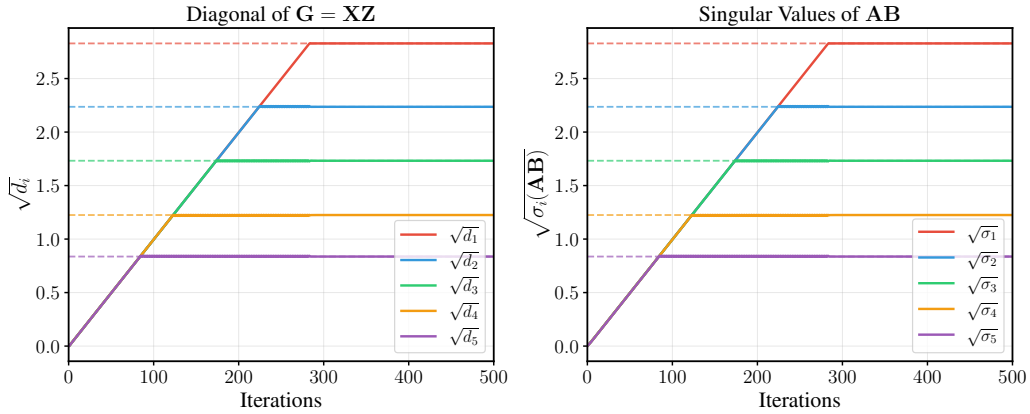
*Remark* 5.8. Theorems 5.4 and 5.7 are with respect to Lebesgue measure on the full space $\mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$. Since the LoRA initialization lies in a Lebesgue-null subset of this space as $\mathbf{A}(0) = \mathbf{0}$, these theorems do not imply an almost-sure guarantee for LoRA-style initialization. Convergence results tailored to sufficiently small LoRA initialization are provided by Corollary B.4 and Theorem C.1.

# 6 Experimental Results

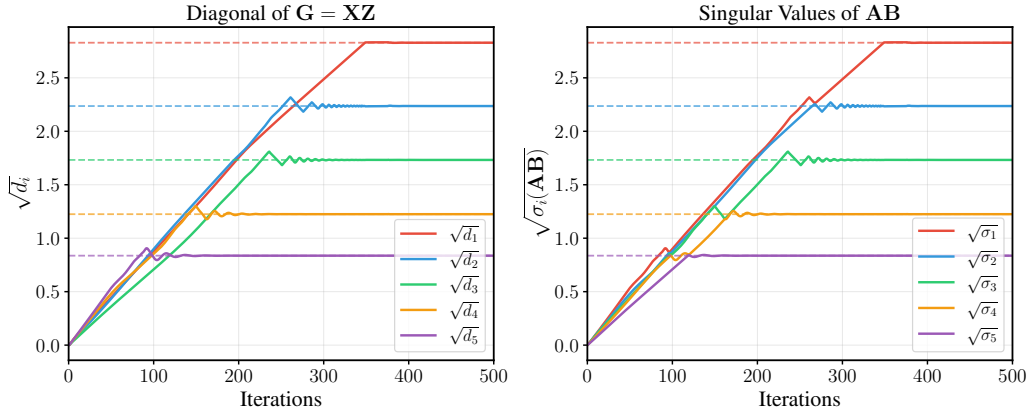
## 6.1 Validation of Our Theory

In this section, we validate our theoretical predictions and confirm the uniform spectral growth.

**Experimental Setup.** We train $(\mathbf{A}, \mathbf{B})$ with SpecGD on the objective in Equation (2). We set $(m, n, r) = (60, 70, 5)$ and then construct $\mathbf{Y} = \mathbf{U}_r\mathbf{\Sigma}\mathbf{V}_r^\top$ where $\mathbf{U}_r \in \mathbb{R}^{m\times r}$ and $\mathbf{V}_r \in \mathbb{R}^{n\times r}$ are random orthonormal matrices, and $\mathbf{\Sigma} = \mathrm{Diag}(8, 5, 3, 1.5, 0.7)$. We initialize $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{B}(0)$ with i.i.d. Gaussian entries of scale $\gamma = 10^{-3}$. We run SpecGD with a learning rate $\eta = 0.01$ and $\beta = 10^{-8}$. We track the diagonal entries $d_i(t)$ defined in Section 4 and compare them with the singular values of $\mathbf{A}(t)\mathbf{B}(t)$.




(a) SpecGD ($\beta = 10^{-8}$)




(b) Muon with NS approximation ($\mu = 0.9$)

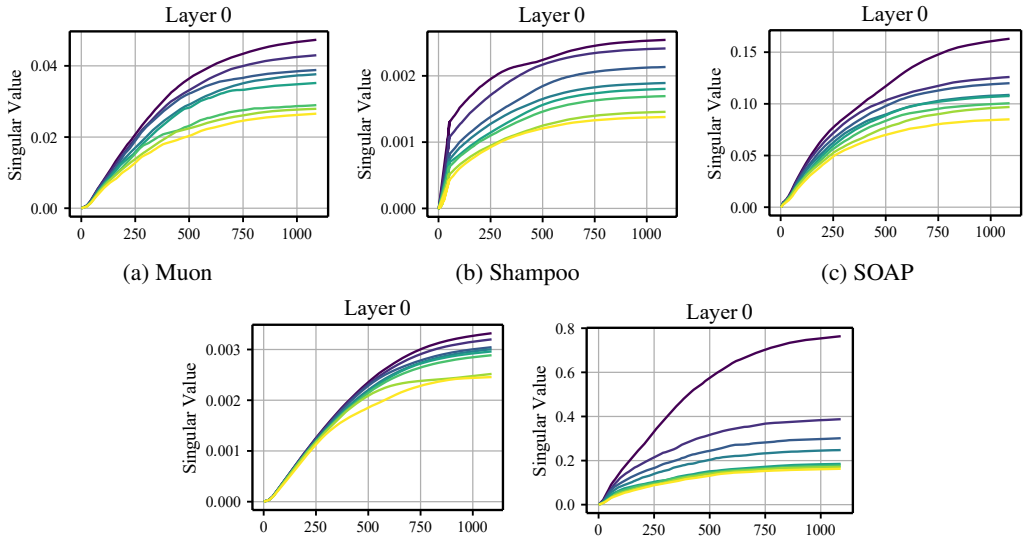
Figure 2: Comparison of SpecGD and Muon with NS approximation on matrix factorization with a rank-5 target. The dashed lines indicate the square roots of singular values of the rank-5 target.

**Results.** We highlight three observations. First, discrete-time SpecGD with $\mathcal{T}_\beta$ successfully minimizes Equation (2). This result is consistent with Corollary B.4, which establishes convergence of continuous-time SpecGF. Second, Figure 2a shows that $d_i(t) = [\mathbf{X}(t)\mathbf{Z}(t)]_{ii}$ closely tracks $\sigma_i(\mathbf{A}(t)\mathbf{B}(t))$ as predicted by Lemma 4.2. Third, the curves $\sqrt{d_i(t)}$ exhibit nearly identical slopes, consistent with our Theorem 4.1.

Since $\mathcal{T}_\beta$ is an analytic surrogate for orthogonalization, whereas Muon uses NS approximation, we also run Muon with NS approximation on the same MF objective. We observe that the diagonal approximation is still valid and the uniform growth is still present, suggesting that the mechanism is not an artifact of our smoothed operator $\mathcal{T}_\beta$. Additional ablations on orthogonalization, momentum, and training settings are provided in Section E.2. Across various settings, the uniform growth of singular values of the product $\mathbf{AB}$ persists.

### 6.2 Uniform Spectral Growth in LoRA Fine-Tuning

We present empirical evidence that near-uniform growth is consistently observed under LoRA fine-tuning with Muon, Shampoo, SOAP, and Dion. In LoRA, a frozen pretrained weight $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ is adapted by a trainable low-rank update $\Delta\mathbf{W} = \mathbf{AB}$, where $\mathbf{A} \in \mathbb{R}^{m\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n}$ with $r \ll \min\{m, n\}$. We apply these matrix optimizers separately to the LoRA factors and observe that the spectra of $\mathbf{AB}$ grow nearly uniformly. We also observe that near-uniform growth persists under non-zero LoRA initialization schemes. We consider nonzero random initialization (NRI) where $\mathbf{A}$ and $\mathbf{B}$ are both randomly sampled from the Kaiming uniform distribution, and Locally Optimal Initialization (LOI) from Bogachev et al. [8].




(a) Muon (b) Shampoo (c) SOAP

(d) Dion (e) AdamW

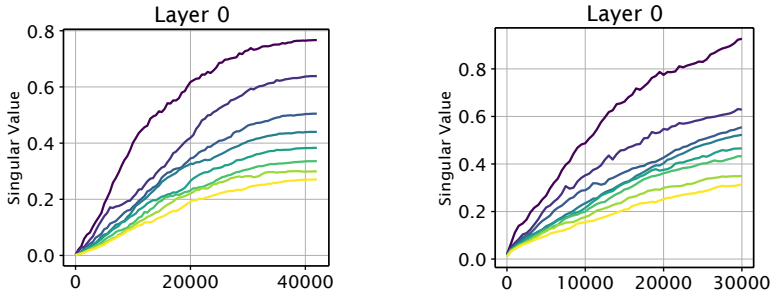
Figure 3: Persistence of uniform growth in LoRA fine-tuning under four matrix optimizers; AdamW may preferentially grow the leading singular values. The trajectories correspond to singular values of the LoRA adapter $\mathbf{AB}$ associated with the query projection matrix in the first self-attention layer of LLaMA-3.2-3B.

**Experimental Setup.** We fine-tune RoBERTa-Base [41] on the SST-2 dataset from the GLUE benchmark [59]. For a larger-scale experiment, we train LLaMA-3.2-3B [23] on the Alpaca dataset [53]. For both cases, LoRA is applied to the query and value projection matrices with rank $r = 8$. We track the singular values of the LoRA product throughout the training. We provide further training details for the experiments in Figures 1, 3 and 4, as well as the details for additional model–dataset combinations in Section E.3.1.

**Results.** We observe that near-uniform growth persists across a range of practical settings. Figure 3 illustrates that Shampoo, SOAP, and Dion also exhibit near-uniform growth of the singular values.

On the other hand, AdamW shows less uniform growth, with the leading singular values growing faster. Figure 4 shows that near-uniform growth persists under nonzero LoRA initialization schemes.

We present the complete singular value trajectories for all matrices in the settings of Figures 3 and 4, together with additional experimental settings and the corresponding singular value trajectories in Sections E.3.2, E.3.4 and E.3.5. Across all layers and settings, we consistently observe near-uniform growth of the singular values of the LoRA $\mathbf{AB}$ adapter under the four matrix optimizers.




(a) LOI, Muon (b) NRI, Muon

Figure 4: Persistence of uniform growth in LoRA fine-tuning under Muon with nonzero initializations. The trajectories correspond to singular values of the LoRA adapter $\mathbf{AB}$ associated with the query projection matrix in the first self-attention layer of RoBERTa-Base.

**Implication.** The uniform-growth mechanism under orthogonalized updates can lead to faster convergence. Standard GF in MF learns the spectral components one at a time by sequentially visiting the vicinity of saddle points of increasing rank [30, 63]. Moreover, the smaller the initialization scale is, the longer it takes to escape each saddle. In contrast, orthogonalized updates promote the evolution of every component. Hence, the time needed to learn all components is essentially that needed to learn the component with the largest singular value; see Figure 2.

On the generalization side, however, uniform growth may not always be beneficial. Its effect depends on whether the minor spectral components are task-relevant signal or irrelevant noise. If minor spectral components correspond to noise, largest-first dynamics can act as an implicit regularizer under early stopping by learning dominant signal directions before overfitting to noise. Orthogonalized updates may fit such noise earlier because the evolution is more uniform across every active component. This suggests the need for additional regularization in settings where minor components are not task-relevant. In our MF problem, $\ell_2$ regularization induces shrinkage of the product spectrum, suppressing components below the regularization scale $\lambda$ [56]. Hence, with an appropriate regularization level, noisy directions can be attenuated relative to the signal directions. Conversely, when minor components are task-relevant, such as rare classes in imbalanced data [57], uniform growth can be beneficial relative to largest-first dynamics. Although this discussion is stated for MF, the empirical spectral trends in LoRA motivate studying whether a similar optimization–generalization tradeoff arises in fine-tuning with LoRA.

## 7 Conclusion

We analyzed spectral gradient flow (SpecGF) for low-rank matrix factorization (MF) and proved that singular values of the product $\mathbf{AB}$ exhibit common quadratic growth up to their respective target values, uniformly in time in a joint small-initialization and smoothing limit. This leads to uniform spectral growth dynamics that contrast with the largest-first behavior of standard GF. We also established exponential factor convergence for sufficiently small fixed parameters in this regime. For general initializations, we proved that bounded SpecGF trajectories converge almost surely to global minima; with $\ell_2$ regularization, boundedness follows and almost-everywhere convergence holds for the regularized objective. We tested our theoretical predictions via experiments on MF and observed qualitatively near-uniform spectral growth across various matrix optimizers and initialization schemes in LoRA fine-tuning. Global convergence without regularization for general initializations remains open due to the absence of known invariants for SpecGF. Extending our analysis to discrete-time SpecGD and general network architectures is important future work.

## Acknowledgement

This work was supported by the InnoCORE program of the Ministry of Science and ICT (No. N10250156, 40%) and two National Research Foundation of Korea (NRF) grants (No. RS-2023-00211352 Development of Efficient Stochastic Optimization Algorithms via Analyzing Without-replacement Sampling Methods, 20%; No. RS-2026-25612350 Development of an End-to-End Design and Validation Framework for Mathematically Principled Next-Generation AI Architectures and Learning Algorithms, 20%) funded by the Korean government (MSIT). This work was also supported by a grant from KRAFTON AI (20%).

## References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: long papers)*, pages 7319–7328, 2021.

[2] Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

[3] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

[5] Martin Bauer, Martins Bruveris, Philipp Harms, and Peter W Michor. Smooth perturbations of the functional calculus and applications to Riemannian geometry on spaces of metrics. *Communications in Mathematical Physics*, 389(2):899–931, 2022.

[6] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[7] Sanjay P Bhat and Dennis S Bernstein. Arc-length-based Lyapunov tests for convergence and stability with applications to systems having a continuum of equilibria. *Mathematics of Control, Signals, and Systems*, 22(2):155–184, 2010.

[8] Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *The Fourteenth International Conference on Learning Representations*, 2026.

[9] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 2021.

[10] Po Chen, Rujun Jiang, and Peng Wang. A complete loss landscape analysis of regularized deep matrix factorization. *SIAM Journal on Optimization*, 36(4):2266–2295, 2026.

[11] Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Gradient descent provably escapes saddle points in the training of shallow ReLU networks. *Journal of Optimization Theory and Applications*, 203(3):2617–2648, 2024.

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

[13] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, 2023.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Ke-Lin Du, MNS Swamy, Zhang-Quan Wang, and Wai Ho Mow. Matrix factorization techniques in machine learning, signal processing, and statistics. *Mathematics*, 11(12):2674, 2023.

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

[17] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. In *High-dimensional Learning Dynamics 2025*, 2025.

[18] Paul MN Feehan. On the Morse–Bott property of analytic functions on Banach spaces with Łojasiewicz exponent one half. *Calculus of Variations and Partial Differential Equations*, 59 (2):87, 2020.

[19] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[20] Evan S Gawlik and Melvin Leok. Computing the Fréchet derivative of the polar decomposition. *arXiv preprint arXiv:1608.04491*, 2016.

[21] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[22] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[24] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

[25] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[26] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. *Advances in Neural Information Processing Systems*, 37:117015–117040, 2024.

[27] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[28] Nicholas J Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

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

[30] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

[31] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pages 15200–15238. PMLR, 2023.

[32] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

[33] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[34] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. MuonBP: Faster Muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

[35] Junsu Kim, Jaeyeon Kim, and Ernest K Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). *arXiv preprint arXiv:2502.09376*, 2025.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[37] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[38] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

[39] Kai Lion, Florian Hübler, Bingcong Li, Antonio Orvieto, and Niao He. Muown: Row-norm control for Muon optimization. *arXiv preprint arXiv:2605.10797*, 2026.

[40] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[42] Stanislaw Lojasiewicz. *Ensembles semi-analytiques*. Institut des Hautes Etudes Scientifiques, 1965.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[44] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

[45] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

[46] Zeju Qiu, Simon Buchholz, Tim Z. Xiao, Maximilian Dax, Bernhard Schölkopf, and Weiyang Liu. Reparameterized LLM training via orthogonal equivalence transformation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[48] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[49] Michael Rushka and Diego Klabjan. Rank-accuracy trade-off for LoRA: A gradient-flow analysis. *arXiv preprint arXiv:2602.10212*, 2026.

[50] Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

[51] Baekrok Shin and Chulhee Yun. Implicit bias and loss of plasticity in matrix completion: Depth promotes low-rankness. In *The Fourteenth International Conference on Learning Representations*, 2026.

[52] Dominik Stöger and Mahdi Soltanolkotabi. 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, 2021.

[53] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford Alpaca: An instruction-following LLaMA model, 2023.

[54] Teknium. OpenHermes 2.5: An open dataset of synthetic data for generalist LLM assistants, 2023. URL `https://huggingface.co/datasets/teknium/OpenHermes-2.5`.

[55] Benjamin Texier. Basic matrix perturbation theory. *L'Enseignement Mathématique*, 64(3): 249–263, Jul 2019. doi: 10.4171/lem/64-3/4-1.

[56] Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[57] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *The Fourteenth International Conference on Learning Representations*, 2026.

[58] Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *13th International Conference on Learning Representations*, 2025.

[59] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

[60] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

[61] Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Taming momentum: Rethinking optimizer states through low-rank approximation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[62] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

[63] Yedi Zhang, Andrew Saxe, and Peter E Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. *arXiv preprint arXiv:2512.20607*, 2025.

[64] Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Forty-second International Conference on Machine Learning*, 2025.

## Supplementary Materials

# A Speed of singular values of $\mathbf{A}$ and $\mathbf{B}$

In this section, we show that $\mathbf{A}$ *and* $\mathbf{B}$ increase at the same speed if the whole matrix evolution can be well-decoupled into scalar evolutions. Throughout this section, we fix $\beta$.

### A.1 Rank-1 Case

We first consider the case with random initialization but simplify the problem to the case where $\mathbf{Y}$ has rank 1 and $r = 1$:

$$\mathbf{Y} = \sigma \mathbf{u}\mathbf{v}^\top, \quad \mathbf{A} \in \mathbb{R}^{m\times 1}, \quad \mathbf{B} \in \mathbb{R}^{1\times n}.$$

There, $\mathbf{u}$ and $\mathbf{v}$ are unit vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

#### A.1.1 SpecGF with $\mathcal{T}_\beta$

We note that the dynamics exist uniquely (from analyticity of $\mathcal{T}_\beta$) and globally (from boundedness of velocity by $\mathcal{T}_\beta$). In addition, for this rank-1 case, SpecGF with $\mathcal{T}_\beta$ is:

$$\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}) = \frac{\nabla_{\mathbf{A}}\mathcal{L}}{\sqrt{\|\nabla_{\mathbf{A}}\mathcal{L}\|_{\mathrm{F}}^2 + \beta}}, \quad \mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}) = \frac{\nabla_{\mathbf{B}}\mathcal{L}}{\sqrt{\|\nabla_{\mathbf{B}}\mathcal{L}\|_{\mathrm{F}}^2 + \beta}}.$$

We sample a random unit vector $\mathbf{w}$ in $\mathbb{R}^n$; then $\alpha \triangleq \langle \mathbf{v}, \mathbf{w}\rangle \neq 0$ with probability 1.

**Assumption A.1.** $\alpha \neq 0$.

A typical LoRA training run would start at

$$\mathbf{A}(0) = \mathbf{0}, \quad \mathbf{B}(0) = \gamma \mathbf{w}^\top,$$

for sufficiently small $\gamma > 0$ with gradients

$$\nabla_{\mathbf{A}}\mathcal{L} = (\mathbf{A}\mathbf{B} - \mathbf{Y})\mathbf{B}^\top,$$
$$\nabla_{\mathbf{B}}\mathcal{L} = \mathbf{A}^\top(\mathbf{A}\mathbf{B} - \mathbf{Y}).$$

We first show that $\mathbf{A}(t) \in \mathrm{span}\{\mathbf{u}\}$ and $\mathbf{B}(t)^\top \in \mathrm{span}\{\mathbf{v}, \mathbf{w}\}$ at all $t \geq 0$ for SpecGF with $\mathcal{T}_\beta$.

**Lemma A.2.** *For all* $t \geq 0$*,* $\mathbf{A}(t) \in \mathrm{span}\{\mathbf{u}\}$ *and* $\mathbf{B}(t)^\top \in \mathrm{span}\{\mathbf{v}, \mathbf{w}\}$ *hold.*

*Proof.* Assume that $\mathbf{A}(t) = a(t)\mathbf{u}$ and $\mathbf{B}(t) = b_0(t)\mathbf{v}^\top + c_0(t)\mathbf{w}^\top$ for some $a(t), b_0(t), c_0(t) \in \mathbb{R}$. We directly compute the gradients:

$$\begin{aligned}
\nabla_{\mathbf{A}}\mathcal{L}(t) &= (\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y})\mathbf{B}(t)^\top \\
&= \mathbf{u}\left((a(t)b_0(t) - \sigma)(b_0(t) + c_0(t)\alpha) + a(t)c_0(t)(b_0(t)\alpha + c_0(t))\right), \\
\nabla_{\mathbf{B}}\mathcal{L}(t) &= \mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}) = a(t)(a(t)b_0(t) - \sigma)\mathbf{v}^\top + a(t)^2 c_0(t)\mathbf{w}^\top.
\end{aligned}$$

Normalization only modifies the coefficients, so the updates $\dot{\mathbf{A}}(t) \in \mathrm{span}\{\mathbf{u}\}$ and $\dot{\mathbf{B}}(t)^\top \in \mathrm{span}\{\mathbf{v}, \mathbf{w}\}$ hold. The proof is concluded as $\mathbf{A}(0) \in \mathrm{span}\{\mathbf{u}\}$ and $\mathbf{B}(0)^\top \in \mathrm{span}\{\mathbf{v}, \mathbf{w}\}$. □

Without loss of generality, assume $\alpha > 0$. If $\alpha < 0$, one can simply flip the sign of $\mathbf{u}$ and $\mathbf{v}$ and the argument analogously holds. If $\alpha = 1$, then $\mathbf{w} = \mathbf{v}$ and it is straightforward to see that

$$\nabla_{\mathbf{A}}\mathcal{L}(t) = (\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y})\mathbf{B}(t)^\top = \mathbf{u}(a(t)b_0(t) - \sigma)b_0(t),$$
$$\nabla_{\mathbf{B}}\mathcal{L}(t) = \mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}) = a(t)(a(t)b_0(t) - \sigma)\mathbf{v}^\top.$$

Hence, if $\gamma^2 < \sigma$, then the dynamics of $a(t)$ and $b_0(t)$ are exactly equal to those in Proposition A.13. Henceforth, consider the case $\alpha \in (0, 1)$.

For the convenience of later analysis, define

$$\mathbf{z} \triangleq \frac{1}{\sqrt{1-\alpha^2}}(\mathbf{w} - \alpha\mathbf{v}), \tag{12}$$

which gives $\mathbf{w} = \alpha\mathbf{v} + \sqrt{1-\alpha^2}\mathbf{z}$ and $\langle \mathbf{v}, \mathbf{z} \rangle = 0$. Thus, $\mathbf{B}(t) = (b_0(t) + \alpha c_0(t))\mathbf{v}^\top + \sqrt{1-\alpha^2}c_0(t)\mathbf{z}^\top$. For brevity, we write the parameters as

$$\mathbf{A}(t) = a(t)\mathbf{u}, \quad \mathbf{B}(t) \triangleq b(t)\mathbf{v}^\top + c(t)\mathbf{z}^\top. \tag{13}$$

The goal is to show that 1) $a(t)b_0(t) \to \sigma$ *with* neither $a(t) \to \infty$ nor $b_0(t) \to \infty$ and 2) $c_0(t) \to 0$; this is equivalent to showing that 1) $a(t)b(t) \to \sigma$ *with* neither $a(t) \to \infty$ nor $b(t) \to \infty$ and 2) $c(t) \to 0$. The remaining part is to show that $a(t) \approx b(t)$; due to randomness in the initialization, $\dot{a}(t)$ cannot be exactly equal to $\dot{b}(t)$. Instead, we show that $|a(t) - b(t)|$ and $|\dot{a}(t) - \dot{b}(t)|$ are both uniformly small over $t$, and the gap vanishes to 0 as $\gamma \to 0$.

We derive the time derivative of the coefficients. From

$$\begin{aligned}
\mathcal{L}(t) &= \frac{1}{2}\|\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}\|_{\mathrm{F}}^2 = \frac{1}{2}\left\|\mathbf{u}\left((a(t)b(t) - \sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\right)\right\|_{\mathrm{F}}^2 \\
&= \frac{1}{2}\|(a(t)b(t) - \sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\|_2^2 \triangleq \frac{1}{2}\|\mathbf{d}(t)\|_2^2,
\end{aligned}$$

with $\mathbf{d}(t) = a(t)\mathbf{B}(t) - \sigma\mathbf{v}^\top$. Whenever $\|\mathbf{d}(t)\|_2 = 0$, $a(t)b(t) = \sigma$ and $c(t) = 0$ and the dynamics terminate. From now on, the time interval of interest is $[0, T_{\max})$ where $T_{\max} = \inf\{t : \|\mathbf{d}(t)\|_2 = 0\}$; $T_{\max}$ could be possibly infinite.

We derive the loss gradient, with $m(t) \triangleq (a(t)b(t) - \sigma)b(t) + a(t)c(t)^2$,

$$\begin{aligned}
\nabla_{\mathbf{A}}\mathcal{L}(t) &= \mathbf{u}\,(m(t))\,, \\
\nabla_{\mathbf{B}}\mathcal{L}(t) &= a(t)\mathbf{d}(t),
\end{aligned}$$

and we obtain

$$\begin{aligned}
\dot{a}(t) &= -\frac{m(t)}{\sqrt{m(t)^2 + \beta}} \\
\dot{b}(t) &= -\frac{a(t)(a(t)b(t) - \sigma)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}} \\
\dot{c}(t) &= -\frac{a(t)^2 c(t)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}}.
\end{aligned}$$

As $c(0) = \sqrt{1-\alpha^2}\gamma > 0$, it directly follows that $c(t) \geq 0$ for all $t \geq 0$ and $c(t)$ is non-increasing. In addition, we have $a(0) = 0$ with $\dot{a}(0) > 0$ and $b(0) = \alpha\gamma > 0$. The simultaneous non-negativity of $a$ and $b$ at $t = 0$ ensures that they remain non-negative for all $t \geq 0$ under continuity. If $b$ hits 0 while $a(t) \geq 0$, it gains non-negative speed; the same holds for $a$. The only way for them to be negative is that both $a$ and $b$ go across 0 at the same time, but then the dynamics become stationary. Therefore, we can safely argue that $a(t), b(t), c(t) \geq 0$.

We return our attention to the loss function. We know that $\mathcal{L}(t)$ is non-increasing; hence, $\|\mathbf{d}(t)\|_2$ is also non-increasing. Precisely,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\mathbf{d}(t)\|_2^2 &= \left\langle \mathbf{d}(t), \dot{\mathbf{d}}(t) \right\rangle \\
&= \left\langle \mathbf{d}(t), \dot{a}(t)\mathbf{B}(t) + a(t)\dot{\mathbf{B}}(t) \right\rangle \\
&= \dot{a}(t)\left\langle \mathbf{d}(t), \mathbf{B}(t) \right\rangle + a(t)\left\langle \mathbf{d}(t), \dot{\mathbf{B}}(t) \right\rangle \\
&= -\frac{m(t)}{\sqrt{m(t)^2 + \beta}} \cdot m(t) - a(t)\frac{a(t)\|\mathbf{d}(t)\|_2^2}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}} \\
&= -\frac{m(t)^2}{\sqrt{m(t)^2 + \beta}} - \frac{a(t)^2\|\mathbf{d}(t)\|_2^2}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}} \leq 0.
\end{aligned}$$

Notice that $m(0) = -\sigma b(0) < 0$; thus, $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{d}(0)\|_2^2 < 0$. Hence, for $t > 0$, $\|\mathbf{d}(t)\|_2^2 < \|\mathbf{d}(0)\|_2^2$. Then, $\|\mathbf{d}(t)\|_2 < \sigma$ for $t > 0$ also implies that $a(t), b(t) > 0$; if either of the two is 0 at $t$, then $\|\mathbf{d}(t)\|_2 = \|(a(t)b(t) - \sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\|_2 \geq \sigma$.

**Lemma A.3.** $m(t) < 0$.

*Proof.* Recall that $m(t) = (a(t)b(t) - \sigma)b(t) + a(t)c(t)^2$ and $m(0) = -\sigma b(0) < 0$.

If there exists some time $\tau \triangleq \min\{t : m(t) \geq 0\}$, then $m(\tau) = 0$ and $\dot{a}(\tau) = 0$. In addition, from

$$\begin{aligned} m(t) &= \langle \mathbf{d}(t), \mathbf{B}(t) \rangle \\ &= a(t)\|\mathbf{B}(t)\|_{\mathrm{F}}^2 - \sigma b(t). \\ \dot{m}(t) &= \dot{a}(t)\|\mathbf{B}(t)\|_{\mathrm{F}}^2 + a(t)\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{B}(t)\|_{\mathrm{F}}^2 - \sigma\dot{b}(t) \\ &= \dot{a}(t)\|\mathbf{B}(t)\|_{\mathrm{F}}^2 - 2a(t)^2\frac{m(t)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}} - \sigma\dot{b}(t). \end{aligned}$$

Substituting $t = \tau$ makes the first two terms vanish on the right-hand side, giving $\dot{m}(\tau) = -\sigma\dot{b}(\tau) < 0$, a contradiction: If $\dot{b}(\tau) = 0$, then we have $a(\tau)b(\tau) = \sigma$ and $c(\tau) = 0$. □

Lemma A.3 directly implies that $a(t)b(t) < \sigma$; thus, $a(t)$ and $b(t)$ are strictly increasing and positive. Therefore, all the scalars $a(t), b(t)$, and $c(t)$ are bounded; thus, SpecGF with $\mathcal{T}_\beta$ converges. From Bah et al. [4], the critical point is either $(a, b, c) = (0, 0, c)$ or $\mathbf{Y}$, the point where $a(t)b(t) = \sigma$ and $c(t) = 0$. For $t > 0$, $\|\mathbf{d}(t)\|_2^2 < \|\mathbf{d}(0)\|_2^2$ and thus $(0, 0, c)$ cannot be the converging point. Therefore, SpecGF with $\mathcal{T}_\beta$ converges to the global minima, and we see that $\|\mathbf{d}(t)\|_2 \to 0$. Hence, $a(t)b(t) \to \sigma$ and $c(t) \to 0$. This recovers the limiting behavior of $b_0(t) = b(t) - \alpha c_0(t)$ and $c_0(t) = \frac{1}{\sqrt{1-\alpha^2}}c(t)$.

We now bound $\Delta(t) \triangleq a(t) - b(t)$. Recall that

$$\|\mathbf{d}(t)\|_2 = \sqrt{(a(t)b(t) - \sigma)^2 + (a(t)c(t))^2}.$$

1. $a(t) \geq b(t)$: Notice that, from
$-m(t) = -(a(t)b(t) - \sigma)b(t) - a(t)c(t)^2 \leq -(a(t)b(t) - \sigma)b(t) \leq -(a(t)b(t) - \sigma)a(t)$,

$$\dot{a}(t) = -\frac{m(t)}{\sqrt{m(t)^2 + \beta}} = \frac{-m(t)}{\sqrt{(-m(t))^2 + \beta}} \leq \frac{-(a(t)b(t) - \sigma)a(t)}{\sqrt{(a(t)b(t) - \sigma)^2 a(t)^2 + \beta}}$$

Then, letting $\tilde{a}(t) = -(a(t)b(t) - \sigma)a(t)$,

$$\begin{aligned} \dot{\Delta}(t) &= \dot{a}(t) - \dot{b}(t) \\ &\leq \frac{-(a(t)b(t) - \sigma)a(t)}{\sqrt{(a(t)b(t) - \sigma)^2 a(t)^2 + \beta}} + \frac{a(t)(a(t)b(t) - \sigma)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}} \\ &= \frac{-(a(t)b(t) - \sigma)a(t)\left(a(t)^2\|\mathbf{d}(t)\|_2^2 - (a(t)b(t) - \sigma)^2 a(t)^2\right)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}\sqrt{\tilde{a}(t)^2 + \beta}\left(\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta} + \sqrt{\tilde{a}(t)^2 + \beta}\right)} \\ &= \frac{-(a(t)b(t) - \sigma)a(t)^3\left(a(t)c(t)\right)^2}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}\sqrt{\tilde{a}(t)^2 + \beta}\left(\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta} + \sqrt{\tilde{a}(t)^2 + \beta}\right)} \\ &= -\dot{c}(t)\frac{-(a(t)b(t) - \sigma)a(t)^3 c(t)}{\sqrt{\tilde{a}(t)^2 + \beta}\left(\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta} + \sqrt{\tilde{a}(t)^2 + \beta}\right)} \\ &\leq -\dot{c}(t)\frac{-(a(t)b(t) - \sigma)a(t)^3 c(t)}{-(a(t)b(t) - \sigma)a(t)\left(a(t)\|\mathbf{d}(t)\|_2\right)} \leq -\dot{c}(t). \end{aligned}$$

2. $a(t) < b(t)$: We require additional assumptions on $\gamma$.
**Assumption A.4.** Take $T \triangleq \frac{\sqrt{1-\alpha^2}}{\alpha}\left(1 + \frac{4\sqrt{\beta}}{\sigma}\right) > 0$. $\gamma > 0$ satisfies

$$\gamma^2 < \frac{\alpha\sigma}{4(1-\alpha^2)^{3/2}} \quad \text{and} \quad \gamma < \min\left\{\frac{\sigma/2}{\sqrt{1-\alpha^2}(\alpha + T)}, 1\right\}.$$

Notice that $T$ of Assumption A.4 is independent of $\gamma$. Assumption A.4 translates to

$$c(0)^3 \leq \frac{\sigma}{4} b(0), \quad c(0)\,(b(0)+T) \leq \frac{\sigma}{2}.$$

The first one is straightforward; the second one holds as

$$c(0)\,(b(0)+T) = \gamma\sqrt{1-\alpha^2}\,(\gamma\alpha+T) \leq \gamma\sqrt{1-\alpha^2}\,(\alpha+T) < \frac{\sigma}{2}.$$

We see that

$$\begin{aligned}
-\dot{\Delta}(t) &= \dot{b}(t) - \dot{a}(t) \\
&\leq -\frac{a(t)(a(t)b(t)-\sigma)}{\sqrt{(a(t)b(t)-\sigma)^2 a(t)^2 + \beta}} + \frac{m(t)}{\sqrt{(a(t)b(t)-\sigma)^2 b(t)^2 + \beta}} \\
&= -\frac{a(t)(a(t)b(t)-\sigma)}{\sqrt{(a(t)b(t)-\sigma)^2 a(t)^2 + \beta}} \\
&\quad + \frac{b(t)(a(t)b(t)-\sigma)}{\sqrt{(a(t)b(t)-\sigma)^2 b(t)^2 + \beta}} + \frac{a(t)c(t)^2}{\sqrt{(a(t)b(t)-\sigma)^2 b(t)^2 + \beta}} \\
&\leq \frac{a(t)c(t)^2}{\sqrt{(a(t)b(t)-\sigma)^2 b(t)^2 + \beta}} \leq \frac{a(t)c(t)^2}{\sqrt{\beta}}.
\end{aligned}$$

We aim to bound the integral of $a(t)c(t)^2$. To this end, define $t_c$ when $a$ gets larger than $c$:

$$t_c \triangleq \inf\{t : a(t) \geq c(t)\} < T_{\max}.$$

This time uniquely exists as $a(t)$ is strictly increasing while $c(t)$ (strictly) decreases to 0 for $t > 0$.

- On $t \geq t_c$,

$$\begin{aligned}
a(t)c(t)^2 &= -\dot{c}(t)\frac{c(t)\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2 + \beta}}{a(t)} \\
&\leq -\dot{c}(t)\frac{c(t)}{a(t)}\left(a(t)\|\mathbf{d}(t)\|_2 + \sqrt{\beta}\right) \leq -\dot{c}(t)\left(c(t)\sigma + \sqrt{\beta}\right).
\end{aligned}$$

  This implies

$$\begin{aligned}
\int_{t_c}^{T_{\max}} a(t)c(t)^2 \mathrm{d}t &\leq \sigma \int_{t_c}^{T_{\max}} -\dot{c}(t)c(t)\mathrm{d}t + \sqrt{\beta}\int_{t_c}^{T_{\max}} -\dot{c}(t)\mathrm{d}t \\
&\leq \frac{\sigma}{2}c(0)^2 + \sqrt{\beta}c(0).
\end{aligned}$$

- For $t \leq t_c$, we show that $T \geq t_c$ from Assumption A.4. Then, the integral of $a(t)c(t)^2$ on $[0, t_c]$ is bounded by that on $[0, T]$. Suppose by contradiction that $t_c > T$. On $[0, T] \subset [0, t_c]$,

$$a(t) < c(t) \leq c(0), \quad b(t) \leq b(0) + T.$$

  The latter holds since $\dot{b}(t)$ is bounded by 1. Then, on $[0, T]$,

$$\sigma - a(t)b(t) \geq \sigma - c(0)(b(0)+T) \geq \frac{\sigma}{2},$$

  and this further implies

$$-m(t) = (\sigma - a(t)b(t))b(t) - a(t)c(t)^2 \geq b(0)\frac{\sigma}{2} - c(0)^3 \geq b(0)\frac{\sigma}{4}.$$

  Hence,

$$\dot{a}(t) = -\frac{m(t)}{\sqrt{m(t)^2+\beta}} \geq \frac{-m(t)}{-m(t)+\sqrt{\beta}} \geq \frac{b(0)\sigma}{b(0)\sigma + 4\sqrt{\beta}}.$$

  However, this implies

$$a(T) \geq T\frac{b(0)\sigma}{b(0)\sigma + 4\sqrt{\beta}} = \frac{\sqrt{1-\alpha^2}}{\alpha}\left(1 + \frac{4\sqrt{\beta}}{\sigma}\right)\frac{b(0)\sigma}{b(0)\sigma + 4\sqrt{\beta}} \geq c(0),$$

  a contradiction.

We see that

$$\int_0^{T_{\max}} a(t)c(t)^2\mathrm{d}t = \int_0^{t_c} a(t)c(t)^2\mathrm{d}t + \int_{t_c}^{T_{\max}} a(t)c(t)^2\mathrm{d}t \le Tc(0)^3 + \frac{\sigma}{2}c(0)^2 + \sqrt{\beta}c(0).$$

We summarize what we have until now.

$$\begin{cases} \text{On } I_+ \triangleq \{t : a(t) \ge b(t)\}, & \dot{\Delta}(t) \le -\dot{c}(t). \\ \text{On } I_- \triangleq \{t : a(t) < b(t)\}, & -\dot{\Delta}(t) \le \dfrac{a(t)c(t)^2}{\sqrt{\beta}}. \end{cases}$$

Notice that $\Delta(t)$ is (real-)analytic; hence, its zero set is discrete (even finite on each $[0, t]$). Let $\Delta_+(t) \triangleq \max\{\Delta(t), 0\}$. Then $\Delta_+$ is everywhere continuous and a.e. differentiable (precisely, differentiable except on a discrete set).

1. Define $p(t) \triangleq \Delta_+(t) + c(t)$; it is everywhere continuous and a.e. differentiable. On $I_+$, $\dot{p}(t) = \dot{a}(t) - \dot{b}(t) + \dot{c}(t) \le 0$. On $I_-$, $\dot{p}(t) = \dot{c}(t) \le 0$. Hence, $p(t)$ is non-increasing and
$$\Delta_+(t) + c(t) = p(t) \le p(0) = \Delta_+(0) + c(0).$$

2. Define $q(t) \triangleq (-\Delta)_+(t) - \frac{1}{\sqrt{\beta}}\int_0^t a(s)c(s)^2\mathrm{d}s$; it is everywhere continuous and a.e. differentiable. On $I_-$, $\dot{q}(t) = \dot{b}(t) - \dot{a}(t) - \frac{a(t)c(t)^2}{\sqrt{\beta}} \le 0$. On $I_+$, $\dot{q}(t) = -\frac{a(t)c(t)^2}{\sqrt{\beta}} \le 0$. Hence, $q(t)$ is non-increasing and
$$(-\Delta)_+(t) - \frac{1}{\sqrt{\beta}}\int_0^t a(s)c(s)^2\mathrm{d}s = q(t) \le q(0) = (-\Delta)_+(0).$$

Therefore,

$$\begin{aligned} |\Delta(t)| &= \Delta_+(t) + (-\Delta)_+(t) \le \Delta_+(0) + c(0) - c(t) + (-\Delta)_+(0) + \frac{1}{\sqrt{\beta}}\int_0^t a(s)c(s)^2\mathrm{d}s \\ &\le |\Delta(0)| + c(0) + \frac{1}{\sqrt{\beta}}\int_0^t a(s)c(s)^2\mathrm{d}s \\ &\le b(0) + c(0) + \frac{1}{\sqrt{\beta}}\left[Tc(0)^3 + \frac{\sigma}{2}c(0)^2 + \sqrt{\beta}c(0)\right] \\ &= \gamma\alpha + 2\gamma\sqrt{1-\alpha^2} + \frac{\sigma(1-\alpha^2)}{2\sqrt{\beta}}\gamma^2 + \frac{T}{\sqrt{\beta}}(1-\alpha^2)^{\frac{3}{2}}\gamma^3 \triangleq C(\gamma). \end{aligned}$$

Observe that the difference is uniformly small, and especially the difference vanishes to 0 as $\gamma \to 0$.

**Synchronized speed of factor coordinates.** We know that $a(t)$ is bounded; denote the upper bound of $a(t)$ as $A > 0$. $A$ is computed as follows:

$$a(t)^2 - \Delta(t)a(t) = a(t)b(t) \le \sigma,$$

and thus

$$a(t)^2 \le \sigma + |\Delta(t)|a(t) \le \sigma + C(\gamma)a(t),$$

giving

$$A = \frac{C(\gamma) + \sqrt{C(\gamma)^2 + 4\sigma}}{2}.$$

Notice that $A \to \sqrt{\sigma}$ as $\gamma \to 0$.

**Proposition A.5.**

$$\left|\dot{a}(t) - \dot{b}(t)\right| \le \frac{1}{\sqrt{\beta}}\left(C(\gamma)\sigma + A(1-\alpha^2)\gamma^2\right) + \frac{A^4(1-\alpha^2)\gamma^2}{2\beta}.$$

*The right-hand side goes to 0 as $\gamma \to 0$.*

*Proof.* Recall

$$\dot{a}(t) = -\frac{m(t)}{\sqrt{m(t)^2+\beta}}, \quad m(t) = (a(t)b(t)-\sigma)b(t) + a(t)c(t)^2,$$
$$\dot{b}(t) = -\frac{a(t)(a(t)b(t)-\sigma)}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2+\beta}}.$$

Then,

$$\begin{aligned}
\left|\dot{a}(t)-\dot{b}(t)\right| &\le \left|-\frac{m(t)}{\sqrt{m(t)^2+\beta}} - \frac{a(t)(\sigma-a(t)b(t))}{\sqrt{a(t)^2(\sigma-a(t)b(t))^2+\beta}}\right| \\
&\quad + \left|\frac{a(t)(\sigma-a(t)b(t))}{\sqrt{a(t)^2(\sigma-a(t)b(t))^2+\beta}} - \frac{a(t)(\sigma-a(t)b(t))}{\sqrt{a(t)^2\|\mathbf{d}(t)\|_2^2+\beta}}\right| \\
&\le \frac{1}{\sqrt{\beta}}|-m(t)-a(t)(\sigma-a(t)b(t))| + a(t)(\sigma-a(t)b(t))\times \\
&\quad \left(\frac{1}{\sqrt{a(t)^2(\sigma-a(t)b(t))^2+\beta}} - \frac{1}{\sqrt{a(t)^2\left((\sigma-a(t)b(t))^2+(a(t)c(t))^2\right)+\beta}}\right) \\
&\le \frac{1}{\sqrt{\beta}}\left|(b(t)-a(t))(\sigma-a(t)b(t)) - a(t)c(t)^2\right| + \frac{a(t)(\sigma-a(t)b(t))\times a(t)^4c(t)^2}{2\left(a(t)^2(\sigma-a(t)b(t))^2+\beta\right)^{\frac{3}{2}}} \\
&\le \frac{1}{\sqrt{\beta}}\left(|\Delta(t)|(\sigma-a(t)b(t)) + a(t)c(t)^2\right) + \frac{a(t)^4c(t)^2}{2\left(a(t)^2(\sigma-a(t)b(t))^2+\beta\right)} \\
&\le \frac{1}{\sqrt{\beta}}\left(C(\gamma)\sigma + Ac(0)^2\right) + \frac{A^4c(0)^2}{2\beta} \\
&= \frac{1}{\sqrt{\beta}}\left(C(\gamma)\sigma + A(1-\alpha^2)\gamma^2\right) + \frac{A^4(1-\alpha^2)\gamma^2}{2\beta}.
\end{aligned}$$

The second inequality follows from the Lipschitz continuity of $x \mapsto \frac{x}{\sqrt{x^2+\beta}}$ with constant $\frac{1}{\sqrt{\beta}}$. The third inequality follows from the mean value theorem applied to $x \mapsto \frac{1}{\sqrt{x}}$. □

### A.1.2 SpecGF with $\mathcal{T}$

In this case, SpecGF with $\mathcal{T}$ is identical to the normalized gradient flow:

$$\mathcal{T}(\nabla_{\mathbf{A}}\mathcal{L}) = \frac{\nabla_{\mathbf{A}}\mathcal{L}}{\|\nabla_{\mathbf{A}}\mathcal{L}\|_{\mathrm{F}}}, \quad \mathcal{T}(\nabla_{\mathbf{B}}\mathcal{L}) = \frac{\nabla_{\mathbf{B}}\mathcal{L}}{\|\nabla_{\mathbf{B}}\mathcal{L}\|_{\mathrm{F}}}.$$

As in the case of SpecGF with $\mathcal{T}_\beta$, we sample a random unit vector $\mathbf{w}$ in $\mathbb{R}^n$; then $\alpha \triangleq \langle \mathbf{v}, \mathbf{w}\rangle \neq 0$ with probability 1.

**Assumption A.6.** $\alpha \neq 0$.

For now, assume $\alpha > 0$; we will justify it shortly after. A typical LoRA training run would start at

$$\mathbf{A}_0 = \mathbf{0}, \quad \mathbf{B}_0 = \gamma\mathbf{w}^\top,$$

for sufficiently small $\gamma > 0$ with gradients

$$\nabla_{\mathbf{A}}\mathcal{L} = (\mathbf{AB}-\mathbf{Y})\mathbf{B}^\top,$$
$$\nabla_{\mathbf{B}}\mathcal{L} = \mathbf{A}^\top(\mathbf{AB}-\mathbf{Y}).$$

However, $\mathbf{A}$ being zero induces a zero gradient for $\mathbf{B}$, but $\mathcal{T}$ is discontinuous at zero. This motivates us to initialize the SpecGF at the next iteration of LoRA training with step size $\eta$:

$$\mathbf{A}_1 = \eta\mathbf{u}, \quad \mathbf{B}_1 = \gamma\mathbf{w}^\top.$$

For simplicity, we take $\eta = \gamma$. SpecGF with $\mathcal{T}$ starts at

$$\mathbf{A}(0) = \gamma\mathbf{u}, \quad \mathbf{B}(0) = \gamma\mathbf{w}^\top. \tag{14}$$

We first show that $\mathbf{A}_t \in \operatorname{span}\{\mathbf{u}\}$ and $\mathbf{B}_t^\top \in \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$ at all $t \ge 0$ for *SpecGD* with $\mathcal{T}$.

**Lemma A.7.** *For all $t \geq 0$, $\mathbf{A}_t \in \operatorname{span}\{\mathbf{u}\}$ and $\mathbf{B}_t^\top \in \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$ hold.*

*Proof.* Assume that $\mathbf{A}_t = a\mathbf{u}$ and $\mathbf{B}_t = b\mathbf{v}^\top + c\mathbf{w}^\top$ for some $a, b, c \in \mathbb{R}$. We directly compute the gradients:

$$\nabla_{\mathbf{A}}\mathcal{L}(t) = (\mathbf{A}_t\mathbf{B}_t - \mathbf{Y})\mathbf{B}_t^\top = \mathbf{u}\left((ab - \sigma)(b + c\alpha) + ac(b\alpha + c)\right),$$
$$\nabla_{\mathbf{B}}\mathcal{L}(t) = \mathbf{A}_t^\top(\mathbf{A}_t\mathbf{B}_t - \mathbf{Y}) = a(ab - \sigma)\mathbf{v}^\top + a^2c\mathbf{w}^\top.$$

Normalization only modifies the coefficients, so the updates $\Delta\mathbf{A}_t \in \operatorname{span}\{\mathbf{u}\}$ and $\Delta\mathbf{B}_t^\top \in \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$ hold. The proof is concluded as $\mathbf{A}_0 \in \operatorname{span}\{\mathbf{u}\}$ and $\mathbf{B}_0^\top \in \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$. □

Hence, for SpecGF with $\mathcal{T}$, we assume that there exists a solution that is the continuous version of the solution for SpecGD with $\mathcal{T}$; we study only this solution. Indeed, such $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are the solutions for the rank-1 case with suitable coefficient functions.

**Assumption A.8.** There exist absolutely continuous scalar functions $a(t), b_0(t)$, and $c_0(t)$ such that $\mathbf{A}(t) = a(t)\mathbf{u}$ and $\mathbf{B}(t) = b_0(t)\mathbf{v}^\top + c_0(t)\mathbf{w}^\top$.

We require absolute continuity to define the derivative at almost every $t$. In addition, we assume that the initialization scale $\gamma$ is sufficiently small.

**Assumption A.9.** $\gamma^2 < \alpha\sigma$.

Without loss of generality, assume $\alpha > 0$. If $\alpha < 0$, one can simply flip the sign of $\mathbf{u}$ and $\mathbf{v}$ and the argument analogously holds. If $\alpha = 1$, then $\mathbf{w} = \mathbf{v}$ and it is straightforward to see that

$$\nabla_{\mathbf{A}}\mathcal{L}(t) = (\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y})\mathbf{B}(t)^\top = \mathbf{u}(a(t)b_0(t) - \sigma)b_0(t),$$
$$\nabla_{\mathbf{B}}\mathcal{L}(t) = \mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}) = a(t)(a(t)b_0(t) - \sigma)\mathbf{v}^\top.$$

Hence, if $\gamma^2 < \sigma$, then $a(t)$ and $b_0(t)$ get a unit update, which yields

$$a(t) = \gamma + \min\{T_\gamma, t\}, \quad b_0(t) = \gamma + \min\{T_\gamma, t\} \tag{15}$$

where $T_\gamma > 0$ satisfies $(T_\gamma + \gamma)^2 = \sigma$. Moreover, in this case, $a(t) = b_0(t)$ for all $t \geq 0$. Henceforth, consider the case $\alpha \in (0, 1)$.

For the convenience of later analysis, define

$$\mathbf{z} \triangleq \frac{1}{\sqrt{1 - \alpha^2}}(\mathbf{w} - \alpha\mathbf{v}), \tag{16}$$

which gives $\mathbf{w} = \alpha\mathbf{v} + \sqrt{1 - \alpha^2}\mathbf{z}$ and $\langle \mathbf{v}, \mathbf{z} \rangle = 0$. Thus, $\mathbf{B}(t) = (b_0(t) + \alpha c_0(t))\mathbf{v}^\top + \sqrt{1 - \alpha^2}c_0(t)\mathbf{z}^\top$. For brevity, we write the parameters as

$$\mathbf{A}(t) = a(t)\mathbf{u}, \quad \mathbf{B}(t) \triangleq b(t)\mathbf{v}^\top + c(t)\mathbf{z}^\top. \tag{17}$$

For $0 < \alpha < 1$, we compare the factor coordinates $a(t)$ and $b(t)$ up to the first time either factor gradient vanishes.

We derive the time derivative of the coefficients. From

$$\mathcal{L}(t) = \frac{1}{2}\|\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}\|_{\mathrm{F}}^2 = \frac{1}{2}\left\|\mathbf{u}\left((a(t)b(t) - \sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\right)\right\|_{\mathrm{F}}^2$$
$$= \frac{1}{2}\|(a(t)b(t) - \sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\|_2^2 \triangleq \frac{1}{2}\|\mathbf{d}(t)\|_2^2,$$

with $\mathbf{d}(t) = a(t)\mathbf{B}(t) - \sigma\mathbf{v}^\top$. Whenever $\|\mathbf{d}(t)\|_2 = 0$, $a(t)b(t) = \sigma$ and $c(t) = 0$ and the dynamics terminate. For $0 < \alpha < 1$, let $\tau_\gamma$ denote the first time at which either factor gradient vanishes. We study the exact flow on $0 \leq t < \tau_\gamma$. Both factor gradients are nonzero at initialization, and the vector field is smooth on this interval.

We derive the loss gradient, with $m(t) \triangleq (a(t)b(t) - \sigma)b(t) + a(t)c(t)^2$,

$$\nabla_{\mathbf{A}}\mathcal{L}(t) = \mathbf{u}\left(m(t)\right),$$
$$\nabla_{\mathbf{B}}\mathcal{L}(t) = a(t)\mathbf{d}(t),$$

and we obtain

$$\begin{aligned}\dot{a}(t) &= -\operatorname{sgn}\left(m(t)\right)\\ \dot{b}(t) &= -\operatorname{sgn}(a(t))\frac{a(t)b(t)-\sigma}{\|\mathbf{d}(t)\|_2}\\ \dot{c}(t) &= -\operatorname{sgn}(a(t))\frac{a(t)c(t)}{\|\mathbf{d}(t)\|_2}.\end{aligned}$$

Since all the time derivatives are bounded by 1, due to $\mathcal{T}$, we see that $a, b, c$ are all (Lipschitz) continuous in $t$. As $c(0) = \sqrt{1-\alpha^2}\gamma > 0$, it directly follows that $c(t) \geq 0$ for all $t \geq 0$ and $c(t)$ is non-increasing. In addition, we have $a(0) = \gamma$ with $\dot{a}(0) = 1$ (from Assumption A.9) and $b(0) = \alpha\gamma > 0$. The simultaneous non-negativity of $a$ and $b$ at $t = 0$ ensures that they remain non-negative for all $t \geq 0$ under continuity. If $b$ hits 0 while $a(t) \geq 0$, it gains non-negative speed; the same holds for $a$. The only way for them to be negative is that both $a$ and $b$ go across 0 at the same time, but then the dynamics become stationary. Therefore, we can safely argue that $a(t), b(t), c(t) \geq 0$.

We return our attention to the loss function. We know that $\mathcal{L}(t)$ is non-increasing; hence, $\|\mathbf{d}(t)\|_2$ is also non-increasing. Precisely,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\mathbf{d}(t)\|_2^2 &= \left\langle \mathbf{d}(t), \dot{\mathbf{d}}(t)\right\rangle\\ &= \left\langle \mathbf{d}(t), \dot{a}(t)\mathbf{B}(t) + a(t)\dot{\mathbf{B}}(t)\right\rangle\\ &= \dot{a}(t)\left\langle \mathbf{d}(t), \mathbf{B}(t)\right\rangle + a(t)\left\langle \mathbf{d}(t), \dot{\mathbf{B}}(t)\right\rangle\\ &= -\operatorname{sgn}\left(m(t)\right)m(t) - a(t)\operatorname{sgn}(a(t))\frac{\|\mathbf{d}(t)\|_2^2}{\|\mathbf{d}(t)\|_2}\\ &= -|m(t)| - |a(t)|\|\mathbf{d}(t)\|_2 \leq 0.\end{aligned}$$

Notice that $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{d}(0)\|_2^2 < 0$. Hence, for $t > 0$, $\|\mathbf{d}(t)\|_2^2 < \|\mathbf{d}(0)\|_2^2$. Then, $\|\mathbf{d}(t)\|_2 < \sigma$ for $t > 0$ also implies that $a(t), b(t) > 0$; if either of the two is 0 at $t$, $\|\mathbf{d}(t)\|_2 = \|(a(t)b(t)-\sigma)\mathbf{v}^\top + a(t)c(t)\mathbf{z}^\top\|_2 \geq \sigma$. Therefore, we may discard the sign of $a(t)$ in the time derivatives for $t > 0$:

$$\begin{aligned}\dot{a}(t) &= -\operatorname{sgn}\left(m(t)\right)\\ \dot{b}(t) &= -\frac{a(t)b(t)-\sigma}{\|\mathbf{d}(t)\|_2}\\ \dot{c}(t) &= -\frac{a(t)c(t)}{\|\mathbf{d}(t)\|_2}.\end{aligned}$$

**Lemma A.10.** *Under Assumption A.9, for* $0 \leq t < \tau_\gamma$,

$$a(t) = \gamma + t, \qquad a(t)b(t) < \sigma, \qquad 0 \leq a(t) - b(t) \leq \gamma\big(1-\alpha+\sqrt{1-\alpha^2}\big).$$

*Proof.* Recall that $m(t) = (a(t)b(t)-\sigma)b(t) + a(t)c(t)^2$ and $m(0) = \gamma(\gamma^2 - \alpha\sigma) < 0$; moreover, $m(t)$ is (bounded) Lipschitz continuous. Thus $m < 0$ and $\dot{a} = 1$. This implies that $a(t)b(t) < \sigma$ and that $b(t)$ is strictly increasing and positive.

Let $\Delta(t) \triangleq a(t) - b(t)$. We will bound $|\Delta(t) - \Delta(0)|$ for $t > 0$. Recall that

$$\|\mathbf{d}(t)\|_2 = \sqrt{\left(a(t)b(t)-\sigma\right)^2 + \left(a(t)c(t)\right)^2}.$$

Then,

$$\dot{\Delta}(t) = \dot{a}(t) - \dot{b}(t) = 1 + \frac{a(t)b(t)-\sigma}{\|\mathbf{d}(t)\|_2} \geq 0.$$

Furthermore, $\|\mathbf{d}(t)\|_2 + a(t)b(t) - \sigma \leq a(t)c(t)$ yields

$$\dot{\Delta}(t) = 1 + \frac{a(t)b(t)-\sigma}{\|\mathbf{d}(t)\|_2} \leq -\dot{c}(t).$$

We see that $\left|\dot{\Delta}(t)\right| \leq -\dot{c}(t)$; hence, the total variation of $\Delta$ is at most $c(0)$, that is,

$$|\Delta(t) - \Delta(0)| \leq c(0) = \gamma\sqrt{1-\alpha^2}.$$

Plugging $\Delta(0) = \gamma(1-\alpha)$, we see that

$$|\Delta(t)| \leq \gamma(1-\alpha) + \gamma\sqrt{1-\alpha^2} \triangleq C(\gamma). \qquad \square$$

Observe that the difference is uniformly small, and especially the difference vanishes to 0 as $\gamma \to 0$.

**Growth rates of the factor coordinates.** We know that $a(t)$ is bounded; denote the upper bound of $a(t)$ as $A > 0$. $A$ is computed as follows:

$$a(t)^2 - \Delta(t)a(t) = a(t)b(t) \leq \sigma,$$

and thus

$$a(t)^2 \leq \sigma + |\Delta(t)|a(t) \leq \sigma + C(\gamma)a(t),$$

giving

$$A = \frac{C(\gamma) + \sqrt{C(\gamma)^2 + 4\sigma}}{2}.$$

Notice that $A \to \sqrt{\sigma}$ as $\gamma \to 0$.

**Proposition A.11.**

$$\left|\dot{a}(t) - \dot{b}(t)\right| \leq \frac{A\gamma\sqrt{1-\alpha^2}}{2(\sigma - a(t)b(t))}.$$

*Specifically, whenever*

$$\sigma - a(t)b(t) \geq K\sigma,$$

*for some* $0 < K < 1$*, we have*

$$\left|\dot{a}(t) - \dot{b}(t)\right| \leq \frac{A\sqrt{1-\alpha^2}}{2K\sigma}\gamma = \Theta(\gamma). \tag{18}$$

*Proof.* Recall

$$\dot{a}(t) = 1, \quad \dot{b}(t) = -\frac{a(t)b(t) - \sigma}{\|\mathbf{d}(t)\|_2}.$$

Then,

$$\begin{aligned}
&\left|\dot{a}(t) - \dot{b}(t)\right| \\
&= 1 - \frac{\sigma - a(t)b(t)}{\|\mathbf{d}(t)\|_2} \\
&= 1 - \frac{\sigma - a(t)b(t)}{\sqrt{\left(a(t)b(t) - \sigma\right)^2 + \left(a(t)c(t)\right)^2}} \\
&= \frac{\left(a(t)c(t)\right)^2}{\sqrt{\left(a(t)b(t) - \sigma\right)^2 + \left(a(t)c(t)\right)^2}\left(\sqrt{\left(a(t)b(t) - \sigma\right)^2 + \left(a(t)c(t)\right)^2} + \sigma - a(t)b(t)\right)} \\
&\leq \frac{\left(a(t)c(t)\right)^2}{a(t)c(t)2\left(\sigma - a(t)b(t)\right)} \\
&\leq \frac{A\gamma\sqrt{1-\alpha^2}}{2\left(\sigma - a(t)b(t)\right)}.
\end{aligned}$$

$\square$

**Fast rotation of $\mathbf{B}$ towards $\mathbf{v}$.** We independently state that $\mathbf{B}(t)$ quickly aligns with $\mathbf{v}^\top$ at early times. Let $\theta(t) \in (0, \frac{\pi}{2})$ be the angle between $\mathbf{B}(t)$ and $\mathbf{v}$. Then $b(t) = \|\mathbf{B}(t)\|_{\mathrm{F}} \cos\theta(t)$ and $c(t) = \|\mathbf{B}(t)\|_{\mathrm{F}} \sin\theta(t)$, hence

$$\dot{\theta}(t) = \frac{b(t)\dot{c}(t) - \dot{b}(t)c(t)}{\|\mathbf{B}(t)\|_{\mathrm{F}}^2} = \frac{-\sigma c(t)}{\|\mathbf{d}(t)\|_2 \|\mathbf{B}(t)\|_{\mathrm{F}}^2} = -\frac{\sigma \sin\theta(t)}{\|\mathbf{d}(t)\|_2 \|\mathbf{B}(t)\|_{\mathrm{F}}}.$$

From the fact that $\frac{\mathrm{d}}{\mathrm{d}\theta} \log\tan\left(\frac{\theta}{2}\right) = \csc\theta$ for $\theta \in (0, \pi)$, we see that

$$\frac{\mathrm{d}}{\mathrm{d}t} \log\tan\left(\frac{\theta(t)}{2}\right) = -\frac{\sigma}{\|\mathbf{d}(t)\|_2 \|\mathbf{B}(t)\|_{\mathrm{F}}} \leq -\frac{1}{\|\mathbf{B}(t)\|_{\mathrm{F}}}.$$

Note that, at small $t > 0$, the fraction is large in magnitude and negative, making $\theta$ quickly decrease towards 0.

### A.2 Spectral initialization Case

We first show a closed formula for singular values of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ if the singular-vector matrices of $\mathbf{Y}$ are directly applied at the initialization. Recall that $\mathbf{Y} = \mathbf{U}_{r^\star} \mathbf{\Sigma} \mathbf{V}_{r^\star}^\top$ is the compact SVD of $\mathbf{Y}$. Let $\mathbf{U}_r$ and $\mathbf{V}_r$ consist of the first $r$ columns of $\mathbf{U}_{r^\star}$ and $\mathbf{V}_{r^\star}$, respectively. The spectral initialization [64] sets

$$\mathbf{A}(0) = \mathbf{U}_r \mathbf{\Sigma}_{\mathbf{A},0} \mathbf{Q}^\top, \quad \mathbf{B}(0) = \mathbf{Q} \mathbf{\Sigma}_{\mathbf{B},0} \mathbf{V}_r^\top,$$

where $\mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_{\mathbf{B},0}$ are non-negative diagonal matrices. Zhang et al. [64] set the diagonal matrices to be the square root of the singular value matrix of the one-step gradient of full fine-tuning; any non-negative diagonal matrices suffice for our case.

**Proposition A.12.** *Let*

$$\mathbf{A}(0) = \mathbf{U}_r \mathbf{\Sigma}_{\mathbf{A},0} \mathbf{Q}^\top, \quad \mathbf{B}(0) = \mathbf{Q} \mathbf{\Sigma}_{\mathbf{B},0} \mathbf{V}_r^\top,$$

*where $\mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_{\mathbf{B},0}$ are non-negative diagonal matrices with $\|\mathbf{\Sigma}_{\mathbf{A},0}\|_2, \|\mathbf{\Sigma}_{\mathbf{B},0}\|_2 < \sqrt{\sigma_{\min}(\mathbf{Y})}$ and $\mathbf{Q} \in \mathbb{R}^{r \times r}$ is any orthogonal matrix. Then, for all $t \geq 0$, the SpecGF trajectory with $\mathcal{T}$ can be written as*

$$\mathbf{A}(t) = \mathbf{U}_r \mathbf{\Sigma}_A(t) \mathbf{Q}^\top, \quad \mathbf{B}(t) = \mathbf{Q} \mathbf{\Sigma}_B(t) \mathbf{V}_r^\top$$

*with $\mathbf{\Sigma}_A(0) = \mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_B(0) = \mathbf{\Sigma}_{\mathbf{B},0}$. Moreover, $\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) \to \mathbf{\Sigma}_r$ and $\dot{\mathbf{\Sigma}}_A(t) = \dot{\mathbf{\Sigma}}_B(t) = \mathbf{I}_r$ until saturation, where $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$ is the leading principal submatrix of $\mathbf{\Sigma}$, unless $\mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_{\mathbf{B},0}$ have zero diagonal values at the same coordinate.*

*Proof.* Let $\sigma_{\mathbf{A},i}(t)$ and $\sigma_{\mathbf{B},i}(t)$ be the $i$-th diagonal entries of $\mathbf{\Sigma}_{\mathbf{A}}(t)$ and $\mathbf{\Sigma}_{\mathbf{B}}(t)$, respectively. Similarly, denote the $i$-th singular value of $\mathbf{\Sigma}$ as $\sigma_i$. If

$$\mathbf{A}(t) = \mathbf{U}_r \mathbf{\Sigma}_A(t) \mathbf{Q}^\top, \quad \mathbf{B}(t) = \mathbf{Q} \mathbf{\Sigma}_B(t) \mathbf{V}_r^\top,$$

then

$$\nabla_{\mathbf{A}} \mathcal{L}(t) = (\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y})\mathbf{B}(t)^\top = \mathbf{U}_r(\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) - \mathbf{\Sigma}_r)\mathbf{\Sigma}_B(t)^\top \mathbf{Q}^\top,$$
$$\nabla_{\mathbf{B}} \mathcal{L}(t) = \mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}) = \mathbf{Q}\mathbf{\Sigma}_A(t)^\top(\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) - \mathbf{\Sigma}_r)\mathbf{V}_r^\top.$$

Thus,

$$\mathcal{T}(\nabla_{\mathbf{A}} \mathcal{L}(t)) = \mathbf{U}_r \mathbf{\Gamma}_A(t) \mathbf{Q}^\top = \mathbf{U}_r \mathrm{Diag}\{\mathrm{sgn}((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{B},i}(t))\}\mathbf{Q}^\top,$$
$$\mathcal{T}(\nabla_{\mathbf{B}} \mathcal{L}(t)) = \mathbf{Q} \mathbf{\Gamma}_B(t) \mathbf{V}_r^\top = \mathbf{Q}\mathrm{Diag}\{\mathrm{sgn}((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{A},i}(t))\}\mathbf{V}_r^\top,$$

for $i \in [r]$, where $\mathbf{\Gamma}_A(t)$ and $\mathbf{\Gamma}_B(t)$ are the rectangular diagonal matrices. Hence, only the singular values of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are updated, and we can write

$$\mathbf{\Sigma}_A(0) = \mathbf{\Sigma}_{\mathbf{A},0}, \quad \mathbf{\Sigma}_B(0) = \mathbf{\Sigma}_{\mathbf{B},0}.$$

Since the initialization is small, we see that all the signs in $\mathbf{\Gamma}_A(t)$ and $\mathbf{\Gamma}_B(t)$ are negative before saturation, or $\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) = \sigma_i$. Therefore, all the singular values have a speed of 1 until the saturation, and the solution is

$$\mathbf{\Sigma}_{\mathbf{A}}(t) = \mathrm{Diag}\left\{\sigma_{\mathbf{A},i}(0) + \min\{t, T_i\}\right\}_i$$
$$\mathbf{\Sigma}_{\mathbf{B}}(t) = \mathrm{Diag}\left\{\sigma_{\mathbf{B},i}(0) + \min\{t, T_i\}\right\}_i,$$

where for each $i$, $T_i > 0$ satisfies

$$(\sigma_{\mathbf{A},i}(0) + T_i)\,(\sigma_{\mathbf{B},i}(0) + T_i) = \sigma_i. \qquad \square$$

This implies that if $\mathbf{\Sigma}$ is already in non-increasing order, then SpecGF with $\mathcal{T}$ converges to the best rank-$r$ approximation of $\mathbf{Y}$. We have a similar result for SpecGF with $\mathcal{T}_\beta$.

**Proposition A.13.** *Let*

$$\mathbf{A}(0) = \mathbf{U}_r \mathbf{\Sigma}_{\mathbf{A},0} \mathbf{Q}^\top, \quad \mathbf{B}(0) = \mathbf{Q} \mathbf{\Sigma}_{\mathbf{B},0} \mathbf{V}_r^\top,$$

*where $\mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_{\mathbf{B},0}$ are nonnegative diagonal matrices with $\|\mathbf{\Sigma}_{\mathbf{A},0}\|_2, \|\mathbf{\Sigma}_{\mathbf{B},0}\|_2 < \gamma < \sqrt{\sigma_{\min}(\mathbf{Y})}$ for $\gamma > 0$ and $\mathbf{Q} \in \mathbb{R}^{r \times r}$ is any orthogonal matrix. Then, for all $t \geq 0$, the SpecGF trajectory with $\mathcal{T}_\beta$ can be written as*

$$\mathbf{A}(t) = \mathbf{U}_r \mathbf{\Sigma}_A(t) \mathbf{Q}^\top, \quad \mathbf{B}(t) = \mathbf{Q} \mathbf{\Sigma}_B(t) \mathbf{V}_r^\top$$

*with $\mathbf{\Sigma}_A(0) = \mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_B(0) = \mathbf{\Sigma}_{\mathbf{B},0}$. Moreover, $\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) \to \mathbf{\Sigma}_r$, where $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$ is the leading principal submatrix of $\mathbf{\Sigma}$, unless $\mathbf{\Sigma}_{\mathbf{A},0}$ and $\mathbf{\Sigma}_{\mathbf{B},0}$ have zero diagonal values at the same coordinate. In addition, both $|\mathbf{\Sigma}_A(t) - \mathbf{\Sigma}_B(t)|$ and $\left|\dot{\mathbf{\Sigma}}_A(t) - \dot{\mathbf{\Sigma}}_B(t)\right|$ go to $\mathbf{0}$ as $\gamma \to 0$.*

*Proof.* We note that the dynamics exist uniquely (from analyticity of $\mathcal{T}_\beta$) and globally (from boundedness of velocity by $\mathcal{T}_\beta$). Let $\sigma_{\mathbf{A},i}(t)$ and $\sigma_{\mathbf{B},i}(t)$ be the $i$-th diagonal entries of $\mathbf{\Sigma}_\mathbf{A}(t)$ and $\mathbf{\Sigma}_\mathbf{B}(t)$, respectively. Denote the $i$-th singular value of $\mathbf{\Sigma}$ as $\sigma_i$. If

$$\mathbf{A}(t) = \mathbf{U}_r \mathbf{\Sigma}_A(t) \mathbf{Q}^\top, \quad \mathbf{B}(t) = \mathbf{Q} \mathbf{\Sigma}_B(t) \mathbf{V}_r^\top,$$

then

$$\nabla_\mathbf{A} \mathcal{L}(t) = (\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y})\mathbf{B}(t)^\top = \mathbf{U}_r(\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) - \mathbf{\Sigma}_r)\mathbf{\Sigma}_B(t)^\top \mathbf{Q}^\top,$$
$$\nabla_\mathbf{B} \mathcal{L}(t) = \mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}) = \mathbf{Q}\mathbf{\Sigma}_A(t)^\top(\mathbf{\Sigma}_A(t)\mathbf{\Sigma}_B(t) - \mathbf{\Sigma}_r)\mathbf{V}_r^\top.$$

Thus,

$$\mathcal{T}_\beta(\nabla_\mathbf{A} \mathcal{L}(t)) = \mathbf{U}_r \mathbf{\Gamma}_A(t) \mathbf{Q}^\top = \mathbf{U}_r \mathrm{Diag}\{f_\beta((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{B},i}(t))\}\mathbf{Q}^\top,$$
$$\mathcal{T}_\beta(\nabla_\mathbf{B} \mathcal{L}(t)) = \mathbf{Q} \mathbf{\Gamma}_B(t) \mathbf{V}_r^\top = \mathbf{Q}\mathrm{Diag}\{f_\beta((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{A},i}(t))\}\mathbf{V}_r^\top,$$

for $i \in [r]$, where $\mathbf{\Gamma}_A(t)$ and $\mathbf{\Gamma}_B(t)$ are the rectangular diagonal matrices and

$$f_\beta(x) \triangleq \frac{x}{\sqrt{x^2 + \beta}}.$$

Hence, only the singular values of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are updated, and we can write

$$\mathbf{\Sigma}_A(0) = \mathbf{\Sigma}_{\mathbf{A},0}, \quad \mathbf{\Sigma}_B(0) = \mathbf{\Sigma}_{\mathbf{B},0}.$$

Notice that the overall dynamics are decoupled into $r$ different scalar dynamics:

$$\dot{\sigma}_{\mathbf{A},i}(t) = -f_\beta((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{B},i}(t)),$$
$$\dot{\sigma}_{\mathbf{B},i}(t) = -f_\beta((\sigma_{\mathbf{A},i}(t)\sigma_{\mathbf{B},i}(t) - \sigma_i)\sigma_{\mathbf{A},i}(t)).$$

So it suffices to prove the claim for a single $i \in [r]$; for simplicity, we fix $i$ and simplify the notation:

$$\sigma_{\mathbf{A},i}(t) \triangleq a(t), \quad \sigma_{\mathbf{B},i}(t) \triangleq b(t), \quad \sigma_i \triangleq \sigma$$

and rewrite the dynamics:

$$\dot{a}(t) = -f_\beta((a(t)b(t) - \sigma)b(t)),$$
$$\dot{b}(t) = -f_\beta((a(t)b(t) - \sigma)a(t)).$$

We first note an immediate consequence from the dynamics: $a(t) \geq 0, b(t) \geq 0$ always hold. If either one were violated, then $a$ and $b$ would both have to be zero simultaneously, which would imply that the dynamics terminate.

Another consequence is that $a(t)b(t) \leq \sigma$ and $a(t)b(t)$ is non-decreasing from the update of $a(t)b(t)$ itself:

$$\frac{\mathrm{d}}{\mathrm{d}t} a(t)b(t) = -(a(t)b(t) - \sigma)\left(\frac{a(t)^2}{\sqrt{(a(t)b(t) - \sigma)^2 a(t)^2 + \beta}} + \frac{b(t)^2}{\sqrt{(a(t)b(t) - \sigma)^2 b(t)^2 + \beta}}\right).$$

It is straightforward that $a(t)b(t)$ is non-decreasing as long as $a(t)b(t) \leq \sigma$. Thus, there exists $\sigma_\infty \in [0, \sigma]$ such that $a(t)b(t) \to \sigma_\infty$. Moreover, $a(t)$ and $b(t)$ are strictly increasing on $t > 0$ as long as $a(t)b(t) < \sigma$; the problematic case is $t = 0$, but, without loss of generality, if $a(0) = 0$ then $b(0) > 0$, and thus $\dot{a}(0) > 0$. Hence, $a(t), b(t) > 0$ for all $t > 0$, which yields strictly monotonic evolution. Therefore, both $a(t)$ and $b(t)$ are bounded and attain their respective forward limits, namely $a_\infty > 0$ and $b_\infty > 0$.

Then $a_\infty b_\infty = \sigma_\infty > 0$ with $\lim_{t\to\infty} \dot{a}(t) = -f_\beta((\sigma_\infty - \sigma)b_\infty) = 0$, which implies $\sigma_\infty = \sigma$.

Albeit the global convergence is proved, a closed form for the solution is not generally obtained. As an alternative, we show that if $a(0), b(0) \leq \gamma$ then $a(t) \approx b(t)$ and $\dot{a}(t) \approx \dot{b}(t)$ as $\gamma \to 0$.

Let $d(t) \triangleq a(t) - b(t)$. Then

$$\dot{d}(t) = f_\beta((a(t)b(t) - \sigma)a(t)) - f_\beta((a(t)b(t) - \sigma)b(t)).$$

Since $f_\beta$ is increasing and $a(t)b(t) \leq \sigma$ for all $t$, $d(t)\dot{d}(t) \leq 0$. Hence, $|d(t)| \leq |d(0)| \leq \gamma$.

We now bound $v(t) \triangleq \dot{a}(t) - \dot{b}(t)$. From above,

$$|v(t)| \leq \frac{1}{\sqrt{\beta}}|(a(t)b(t) - \sigma)d(t)| \leq \frac{\sigma}{\sqrt{\beta}}\gamma. \qquad \square$$

Note that $f_\beta$ is $\frac{1}{\sqrt{\beta}}$-Lipschitz.

# B Proofs for Uniform Growth

We first consider $r = r^\star$ and assume that $\mathbf{\Xi}\mathbf{V}_r$ is invertible. We take the limit in Equation (7). For any rectangular $\mathbf{G}$, its SVD gives

$$\mathcal{T}_\beta(\mathbf{G}) = (\mathbf{G}\mathbf{G}^\top + \beta\mathbf{I})^{-\frac{1}{2}}\mathbf{G} = \mathbf{G}(\mathbf{G}^\top\mathbf{G} + \beta\mathbf{I})^{-\frac{1}{2}} \tag{19}$$

All matrix inner products are Frobenius inner products. For a tuple of matrices, we use the norm defined by $\|(\mathbf{X}, \mathbf{Z}, \mathbf{Z}_\perp)\|^2 = \|\mathbf{X}\|_{\mathrm{F}}^2 + \|\mathbf{Z}\|_{\mathrm{F}}^2 + \|\mathbf{Z}_\perp\|_{\mathrm{F}}^2$. The constants $C, c > 0$ are independent of $\gamma$ and $\beta$. They may depend on the target and initialization direction as well as the time interval and the bounds on matrix norms.

**Dynamics in the target singular-vector basis.** Equation (19) gives

$$\|(\dot{\mathbf{A}}, \dot{\mathbf{B}})\| \leq \sqrt{2r}, \qquad \|(\mathbf{A}(t), \mathbf{B}(t))\| \leq \gamma\|\mathbf{\Xi}\|_{\mathrm{F}} + \sqrt{2r}\, t. \tag{20}$$

For $\beta > 0$, smoothness gives local existence and uniqueness. The speed bound in Equation (20) extends the solution to all times. We complete $\mathbf{V}_r$ to an orthogonal matrix $[\mathbf{V}_r \mid \mathbf{V}_\perp]$ and write

$$\mathbf{A} = \mathbf{U}_r\mathbf{X}, \qquad \mathbf{Z} = \mathbf{B}\mathbf{V}_r, \qquad \mathbf{Z}_\perp = \mathbf{B}\mathbf{V}_\perp, \qquad \mathbf{E} = \mathbf{X}\mathbf{Z} - \mathbf{\Sigma}.$$

If $\mathbf{A} = \mathbf{U}_r\mathbf{X}$, the residual and the gradient with respect to $\mathbf{A}$ satisfy

$$\mathbf{A}\mathbf{B} - \mathbf{Y} = \mathbf{U}_r(\mathbf{E}\mathbf{V}_r^\top + \mathbf{X}\mathbf{Z}_\perp\mathbf{V}_\perp^\top), \qquad (\mathbf{A}\mathbf{B} - \mathbf{Y})\mathbf{B}^\top = \mathbf{U}_r(\mathbf{E}\mathbf{Z}^\top + \mathbf{X}\mathbf{Z}_\perp\mathbf{Z}_\perp^\top),$$

The identity $\mathcal{T}_\beta(\mathbf{U}_r\mathbf{G}) = \mathbf{U}_r\mathcal{T}_\beta(\mathbf{G})$ and uniqueness preserve $\mathbf{A} = \mathbf{U}_r\mathbf{X}$ from $\mathbf{A}(0) = 0$. The reduced flow is

$$\begin{aligned} \dot{\mathbf{X}} &= -\mathcal{T}_\beta(\mathbf{E}\mathbf{Z}^\top + \mathbf{X}\mathbf{Z}_\perp\mathbf{Z}_\perp^\top), \\ [\dot{\mathbf{Z}} \mid \dot{\mathbf{Z}}_\perp] &= -\mathcal{T}_\beta([\mathbf{X}^\top\mathbf{E} \mid \mathbf{X}^\top\mathbf{X}\mathbf{Z}_\perp]). \end{aligned} \tag{21}$$

The coordinate change is orthogonal and preserves norms and speeds. Hence, the factors and $\dot{\mathbf{W}}$ are uniformly bounded on finite intervals for $\gamma \leq 1$. We write $\mathbf{G}_A = \nabla_{\mathbf{A}}\mathcal{L}$ and $\mathbf{G}_B = \nabla_{\mathbf{B}}\mathcal{L}$ for the factor gradients. The loss satisfies

$$-\dot{\mathcal{L}} = \langle \mathbf{G}_A, \mathcal{T}_\beta(\mathbf{G}_A)\rangle + \langle \mathbf{G}_B, \mathcal{T}_\beta(\mathbf{G}_B)\rangle \geq 0, \tag{22}$$

because $\langle \mathbf{G}, \mathcal{T}_\beta(\mathbf{G})\rangle = \sum_i \frac{\sigma_i(\mathbf{G})^2}{\sqrt{\sigma_i(\mathbf{G})^2 + \beta}}$. For any orthogonal $\mathbf{Q}$, the orthogonal change of basis in the rank-$r$ factor space

$$(\mathbf{X}, \mathbf{Z}, \mathbf{Z}_\perp) \mapsto (\mathbf{X}\mathbf{Q}^\top, \mathbf{Q}\mathbf{Z}, \mathbf{Q}\mathbf{Z}_\perp) \tag{23}$$

preserves Equation (21) as well as the products $\mathbf{X}\mathbf{Z}$ and $\mathbf{X}\mathbf{Z}_\perp$.

**Monotonicity and perturbation bounds for $\mathcal{T}_\beta$.** When a direction saturates, the factor gradients on the limiting path can lose rank. Monotonicity bounds the product error. To estimate the factor velocities in the active directions, we need a perturbation bound whose constant stays bounded as $\beta \downarrow 0$.

**Lemma B.1.** *For $\beta > 0$ and any two matrices $\mathbf{G}, \mathbf{H}$ of the same size,*

$$\langle \mathbf{G} - \mathbf{H}, \mathcal{T}_\beta(\mathbf{G}) - \mathcal{T}_\beta(\mathbf{H}) \rangle \geq 0. \tag{24}$$

*If the nonzero singular values of $\mathbf{G}_0$ are at least $\mu > 0$, its column and row projectors $\mathbf{P}_L, \mathbf{P}_R$ satisfy the following bounds for every $\mathbf{G}$ of the same size.*

$$\|\mathbf{P}_L(\mathcal{T}_\beta(\mathbf{G}) - \mathcal{T}_\beta(\mathbf{G}_0))\|_{\mathrm{F}} + \|(\mathcal{T}_\beta(\mathbf{G}) - \mathcal{T}_\beta(\mathbf{G}_0))\mathbf{P}_R\|_{\mathrm{F}} \leq \frac{8}{\mu}\|\mathbf{G} - \mathbf{G}_0\|_{\mathrm{F}}, \tag{25}$$

$$\|\mathcal{T}_\beta(\mathbf{G}_0) - \mathcal{T}(\mathbf{G}_0)\|_{\mathrm{F}} \leq \frac{\sqrt{\mathrm{rank}(\mathbf{G}_0)}\,\beta}{2\mu^2}. \tag{26}$$

*The projected bound also holds with $\mathcal{T}$ in place of $\mathcal{T}_\beta$. When $\mathbf{G}_0$ is square and invertible, this gives*

$$\|\mathcal{T}(\mathbf{G}) - \mathcal{T}(\mathbf{G}_0)\|_{\mathrm{F}} \leq \frac{8}{\sigma_{\min}(\mathbf{G}_0)}\|\mathbf{G} - \mathbf{G}_0\|_{\mathrm{F}}. \tag{27}$$

*Proof.* Convexity of the nuclear norm $\|\cdot\|_*$ implies convexity of $\Phi_\beta(\mathbf{G}) = \|[\mathbf{G} \mid \sqrt{\beta}\mathbf{I}]\|_*$. We set $\mathbf{S} = (\mathbf{G}\mathbf{G}^\top + \beta\mathbf{I})^{\frac{1}{2}}$ and differentiate $\mathbf{S}^2$ in the direction $\mathbf{F}$ to obtain

$$\mathbf{S}\dot{\mathbf{S}} + \dot{\mathbf{S}}\mathbf{S} = \mathbf{F}\mathbf{G}^\top + \mathbf{G}\mathbf{F}^\top,$$
$$\mathrm{D}\Phi_\beta(\mathbf{G})[\mathbf{F}] = \mathrm{tr}(\dot{\mathbf{S}}) = \frac{1}{2}\,\mathrm{tr}\big(\mathbf{S}^{-1}(\mathbf{F}\mathbf{G}^\top + \mathbf{G}\mathbf{F}^\top)\big) = \langle \mathcal{T}_\beta(\mathbf{G}), \mathbf{F}\rangle.$$

Adding the convexity inequalities at $\mathbf{G}$ and $\mathbf{H}$ proves (24).

We use the scalar function $f_\beta(s) = \frac{s}{\sqrt{s^2+\beta}}$. For $|s| \geq \mu$ and $s \neq t$,

$$\left|\frac{f_\beta(s) - f_\beta(t)}{s - t}\right| \leq \frac{4}{\mu}. \tag{28}$$

If $|s - t| \geq \frac{\mu}{2}$, the bound $|f_\beta| \leq 1$ gives

$$\left|\frac{f_\beta(s) - f_\beta(t)}{s - t}\right| \leq \frac{2}{|s - t|} \leq \frac{4}{\mu}.$$

If $|s - t| < \frac{\mu}{2}$, every $x$ between $s$ and $t$ satisfies $|x| \geq \frac{\mu}{2}$, so

$$f'_\beta(x) = \frac{\beta}{(x^2 + \beta)^{\frac{3}{2}}} \leq \frac{1}{\sqrt{x^2 + \beta}} \leq \frac{1}{|x|} \leq \frac{2}{\mu}.$$

The mean value theorem gives Equation (28), and the derivative estimate also covers $s = t$.

We apply Equation (28) to the eigenvalues of the symmetric dilation

$$\mathcal{H}(\mathbf{G}) = \begin{bmatrix} 0 & \mathbf{G} \\ \mathbf{G}^\top & 0 \end{bmatrix}, \qquad f_\beta(\mathcal{H}(\mathbf{G})) = \begin{bmatrix} 0 & \mathcal{T}_\beta(\mathbf{G}) \\ \mathcal{T}_\beta(\mathbf{G})^\top & 0 \end{bmatrix}. \tag{29}$$

The block identity follows from the SVD. We use an orthonormal eigenbasis $(\mathbf{u}_i)$ of $\mathbf{S}_0 = \mathcal{H}(\mathbf{G}_0)$ and an orthonormal eigenbasis $(\mathbf{v}_j)$ of $\mathbf{S} = \mathcal{H}(\mathbf{G})$. For the corresponding eigenvalues $s_i$ and $t_j$, we have

$$\mathbf{u}_i^\top(f_\beta(\mathbf{S}_0) - f_\beta(\mathbf{S}))\mathbf{v}_j = (f_\beta(s_i) - f_\beta(t_j))\mathbf{u}_i^\top\mathbf{v}_j,$$
$$\mathbf{u}_i^\top(\mathbf{S}_0 - \mathbf{S})\mathbf{v}_j = (s_i - t_j)\mathbf{u}_i^\top\mathbf{v}_j.$$

For $s_i \neq 0$, we apply Equation (28) entrywise. Both entries vanish if $s_i = t_j$. With $\mathbf{P} = \mathrm{Diag}(\mathbf{P}_L, \mathbf{P}_R)$, summing squares gives

$$\|\mathbf{P}(f_\beta(\mathbf{S}) - f_\beta(\mathbf{S}_0))\|_{\mathrm{F}}^2 \leq \frac{16}{\mu^2}\|\mathbf{S} - \mathbf{S}_0\|_{\mathrm{F}}^2 = \frac{32}{\mu^2}\|\mathbf{G} - \mathbf{G}_0\|_{\mathrm{F}}^2.$$

The left side is the sum of the two projected squared norms. Cauchy–Schwarz gives Equation (25). For $s \geq \mu$,

$$0 \leq 1 - \frac{s}{\sqrt{s^2+\beta}} = \frac{\beta}{\sqrt{s^2+\beta}(\sqrt{s^2+\beta}+s)} \leq \frac{\beta}{2s^2},$$

which proves Equation (26). The SVD formula allows $\beta \downarrow 0$ in Equation (25). When $\mathbf{G}_0$ is square and invertible, both projectors are the identity and the limit $\beta \downarrow 0$ gives Equation (27). If two gradients share their column or row space, the corresponding projection acts as the identity on the difference of their normalized gradients. □

The next two lemmas also cover $r < r^\star$. We assume $\sigma_r > \sigma_{r+1}$ throughout and use the convention $\sigma_{r+1} = 0$ when $r = r^\star$. We write

$$\mathbf{\Sigma}_r = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r), \qquad \mathbf{Y}_r = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top, \qquad \mathbf{Y}_\perp = \mathbf{Y} - \mathbf{Y}_r. \tag{30}$$

The relative gap at rank $r$ is $\kappa = 1 - \sigma_{r+1}/\sigma_r$. The best rank-$r$ approximation has loss $\mathcal{L}_\star = \frac{1}{2}\|\mathbf{Y}_\perp\|_{\mathrm{F}}^2$. Its factorizations form the set

$$\mathcal{M}_0 = \{(\mathbf{A}, \mathbf{B}) : \mathbf{AB} = \mathbf{Y}_r\}.$$

When $r < r^\star$, the excess loss contains the term $-\langle \mathbf{Y}_\perp, \mathbf{W} \rangle$. The gap $\sigma_r - \sigma_{r+1} > 0$ lets us bound its magnitude by a multiple of $\|\mathbf{W} - \mathbf{Y}_r\|_{\mathrm{F}}^2$. Near an optimal factorization, this also gives a lower bound on the gradient norm.

**Lemma B.2.** *We consider a matrix $\mathbf{W}_0$ with the same column and row spaces as $\mathbf{Y}_r$. If $\sigma_r(\mathbf{W}_0) \geq \sigma_r$, then every matrix $\mathbf{W}$ of rank at most $r$ satisfies*

$$|\langle \mathbf{Y}_\perp, \mathbf{W} \rangle| \leq \frac{\sigma_{r+1}}{2\sigma_r} \|\mathbf{W} - \mathbf{W}_0\|_{\mathrm{F}}^2. \tag{31}$$

*Taking $\mathbf{W}_0 = \mathbf{Y}_r$ bounds the excess loss in terms of the approximation error:*

$$\frac{\kappa}{2}\|\mathbf{W} - \mathbf{Y}_r\|_{\mathrm{F}}^2 \leq \frac{1}{2}\|\mathbf{W} - \mathbf{Y}\|_{\mathrm{F}}^2 - \mathcal{L}_\star \leq \frac{2-\kappa}{2}\|\mathbf{W} - \mathbf{Y}_r\|_{\mathrm{F}}^2. \tag{32}$$

*For each compact $K \subset \mathcal{M}_0$, there are $c, \rho > 0$ such that*

$$\|\nabla \mathcal{L}(\mathbf{A}, \mathbf{B})\| \geq c\sqrt{2(\mathcal{L}(\mathbf{A}, \mathbf{B}) - \mathcal{L}_\star)} \tag{33}$$

*whenever* $\mathrm{dist}((\mathbf{A}, \mathbf{B}), K) < \rho$.

*Proof.* The first bound holds directly when $\mathbf{Y}_\perp = 0$. Otherwise, we take $a = \sigma_r/\sigma_{r+1}$. The matrix $\mathbf{W}_0$ is a best rank-$r$ approximation of $\mathbf{W}_0 \pm a\mathbf{Y}_\perp$: the two summands have orthogonal column and row spaces, and the singular values of $a\mathbf{Y}_\perp$ are at most $\sigma_r$. The Eckart–Young inequality [16] gives

$$\|\mathbf{W} - \mathbf{W}_0 \mp a\mathbf{Y}_\perp\|_{\mathrm{F}}^2 \geq a^2 \|\mathbf{Y}_\perp\|_{\mathrm{F}}^2.$$

Expanding both signs proves Equation (31). Taking $\mathbf{W}_0 = \mathbf{Y}_r$ in $\frac{1}{2}\|\mathbf{W} - \mathbf{Y}\|_{\mathrm{F}}^2 - \mathcal{L}_\star = \frac{1}{2}\|\mathbf{W} - \mathbf{Y}_r\|_{\mathrm{F}}^2 - \langle \mathbf{Y}_\perp, \mathbf{W} \rangle$ gives Equation (32).

For $\mathbf{z} = (\mathbf{A}, \mathbf{B}) \in \mathcal{M}_0$, we define $J_{\mathbf{z}}(\mathbf{H}, \mathbf{K}) = \mathbf{HB} + \mathbf{AK}$. Both factors have full rank and $\nabla \mathcal{L}(\mathbf{z}) = 0$. If $J_{\mathbf{z}}(\mathbf{H}, \mathbf{K}) = 0$, multiplication by $\mathbf{B}^\dagger$ after projection orthogonal to $\mathrm{col}(\mathbf{A})$ gives $\mathbf{H} = \mathbf{AR}$. A left inverse of $\mathbf{A}$ then gives $\mathbf{K} = -\mathbf{RB}$. The parameterization $(\mathbf{AR}, \mathbf{R}^{-1}\mathbf{B})$ for $\mathbf{R} \in \mathsf{GL}(r)$ therefore yields

$$\ker J_{\mathbf{z}} = \{(\mathbf{AR}, -\mathbf{RB}) : \mathbf{R} \in \mathbb{R}^{r \times r}\} = T_{\mathbf{z}}\mathcal{M}_0.$$

We apply Equation (32) to $(\mathbf{A} + s\mathbf{H})(\mathbf{B} + s\mathbf{K})$ and let $s \to 0$ to obtain

$$\nabla^2 \mathcal{L}(\mathbf{z})[\boldsymbol{\xi}, \boldsymbol{\xi}] \geq \kappa \|J_{\mathbf{z}}\boldsymbol{\xi}\|_{\mathrm{F}}^2, \qquad \boldsymbol{\xi} = (\mathbf{H}, \mathbf{K}).$$

The nonzero singular values of $J_{\mathbf{z}}$ have a uniform positive lower bound on bounded closed subsets of $\mathcal{M}_0$, since $J_{\mathbf{z}}$ has constant rank. For $\mathbf{x}$ close to $K$, we choose a nearest $\mathbf{z} \in \mathcal{M}_0$ and write $\boldsymbol{\xi} = \mathbf{x} - \mathbf{z} \perp T_{\mathbf{z}}\mathcal{M}_0$. The nearest points lie in a larger compact subset of $\mathcal{M}_0$. Taylor expansion on that subset gives

$$\|\nabla \mathcal{L}(\mathbf{x})\| \geq c\|\boldsymbol{\xi}\|, \qquad \mathcal{L}(\mathbf{x}) - \mathcal{L}_\star \leq C\|\boldsymbol{\xi}\|^2$$

for sufficiently small $\|\boldsymbol{\xi}\|$, proving Equation (33). □

On an interval with $k$ active directions, we compare the product with a matrix that has singular value $t^2$ in each active direction and $\sigma_i$ in each saturated direction. We write $\mathbf{U}_k, \mathbf{V}_k$ for the first $k$ columns of $\mathbf{U}_r, \mathbf{V}_r$ and $\mathbf{P}_k = \mathrm{Diag}(\mathbf{I}_k, 0)$ for the active projector. For $0 \le k \le r$ and $t > 0$, we denote this matrix by $\mathbf{W}_k(t)$ and its diagonal in target coordinates by $\mathbf{\Lambda}_k(t)$.

$$\mathbf{\Lambda}_k(t) = t^2\mathbf{P}_k + (\mathbf{I} - \mathbf{P}_k)\mathbf{\Sigma}_r, \qquad \mathbf{W}_k(t) = \mathbf{U}_r\mathbf{\Lambda}_k(t)\mathbf{V}_r^\top.$$

We write $\mathcal{M}_k(t)$ for the factorizations $\mathbf{AB} = \mathbf{W}_k(t)$ that satisfy $\mathbf{A}\mathbf{A}^\top\mathbf{u}_i = t^2\mathbf{u}_i$ in every active direction $i \le k$.

$$\mathcal{M}_k(t) = \{(\mathbf{A}, \mathbf{B}) : \mathbf{AB} = \mathbf{W}_k(t), \quad \mathbf{U}_k^\top\mathbf{A}\mathbf{A}^\top = t^2\mathbf{U}_k^\top\}.$$

The constraint on $\mathbf{U}_k^\top\mathbf{A}\mathbf{A}^\top$ is absent for $k = 0$. Every point of $\mathcal{M}_k(t)$ has the form

$$\mathbf{A} = \mathbf{U}_r \begin{bmatrix} t\mathbf{I}_k & 0 \\ 0 & \mathbf{C} \end{bmatrix} \mathbf{Q}^\top, \qquad \mathbf{B} = \mathbf{Q} \begin{bmatrix} t\mathbf{I}_k & 0 \\ 0 & \mathbf{C}^{-1}\mathbf{\Sigma}_{>k} \end{bmatrix} \mathbf{V}_r^\top, \tag{34}$$

Here $\mathbf{Q}$ is an orthogonal change of basis in the factor space. The invertible matrix $\mathbf{C}$ allows the factors in the remaining directions to be unbalanced. We write their target singular values as $\mathbf{\Sigma}_{>k} = \mathrm{Diag}(\sigma_{k+1}, \dots, \sigma_r)$. Indeed, the product constraint fixes the column and row spaces. The first $k$ rows of $\mathbf{U}_r^\top\mathbf{A}/t$ are orthonormal and orthogonal to the remaining rows. Completing those rows to an orthonormal basis gives Equation (34). Empty blocks are omitted.

We denote the product error by $\mathbf{R}$ and the error in the active Gram constraint by $\mathbf{H}$.

$$\mathbf{R} = \mathbf{AB} - \mathbf{W}_k, \qquad \mathbf{H} = \mathbf{U}_k^\top\mathbf{A}\mathbf{A}^\top - t^2\mathbf{U}_k^\top.$$

We write $d = (\|\mathbf{R}\|_{\mathrm{F}}^2 + \|\mathbf{H}\|_{\mathrm{F}}^2)^{1/2}$ and use the energies

$$\mathcal{E}_0 = \frac{1}{2}\|\mathbf{R}\|_{\mathrm{F}}^2 - \langle \mathbf{Y}_\perp, \mathbf{AB}\rangle, \qquad \mathcal{E} = \mathcal{E}_0 + \frac{1}{2}\|\mathbf{H}\|_{\mathrm{F}}^2.$$

On the saturated subspace, the factor blocks $\mathbf{C}$ and $\mathbf{C}^{-1}\mathbf{\Sigma}_{>k}$ have product $\mathbf{\Sigma}_{>k}$ for every invertible $\mathbf{C}$. For factor pairs with norm at most $M$, the distance from $(\mathbf{A}, \mathbf{B})$ to $\mathcal{M}_k(t)$ is at most $Cd$ whenever $d$ is sufficiently small. The constant $C$ is uniform over these factorizations. When $r < r^\star$, Equation (31) controls the cross term $-\langle \mathbf{Y}_\perp, \mathbf{AB}\rangle$ in $\mathcal{E}$.

**Lemma B.3.** *For $k \in \{0, \dots, r\}$, we use $\mathcal{M}_k(t)$ on an interval $J = [t_0, T]$ satisfying $0 < t_0 \le T < \tau_k$. When $r < r^\star$, we also require $t_0 \ge \tau_r$. For each factor bound $M > 0$, there are constants $C, \rho > 0$ independent of $0 < \beta \le 1$. The following estimates hold at every $t \in J$ where $\|(\mathbf{A}, \mathbf{B})\| \le M$ and $d \le \rho$. The derivative $\dot{\mathcal{E}}$ is taken along SpecGF.*

$$C^{-1}d \le \mathrm{dist}((\mathbf{A}, \mathbf{B}), \mathcal{M}_k(t)) \le Cd,$$
$$\dot{\mathcal{E}} \le Cd^2 + C\beta d \le C\mathcal{E} + C\beta^2.$$

*If these conditions hold throughout $[s, t] \subset J$, then*

$$d(t) \le C\big(d(s) + \beta\big).$$

*Proof.* We set $\mu = \min\{t_0^2, \sigma_r\}$. For small $\rho$, $\sigma_r(\mathbf{AB}) \ge \mu/2$, so $\|\mathbf{A}^\dagger\|_2, \|\mathbf{B}^\dagger\|_2 \le 2M/\mu$. Moreover,

$$(\mathbf{I} - \mathbf{U}_r\mathbf{U}_r^\top)\mathbf{A} = (\mathbf{I} - \mathbf{U}_r\mathbf{U}_r^\top)\mathbf{R}\mathbf{B}^\dagger, \qquad \mathbf{B}(\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top) = \mathbf{A}^\dagger\mathbf{R}(\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^\top).$$

Thus the component of $\mathbf{A}$ orthogonal to $\mathrm{col}(\mathbf{U}_r)$ is $O(d)$, as is the component of $\mathbf{B}^\top$ orthogonal to $\mathrm{col}(\mathbf{V}_r)$. For $\mathbf{X} = \mathbf{U}_r^\top\mathbf{A}$ and $\mathbf{Z} = \mathbf{B}\mathbf{V}_r$, the identity $\mathbf{XZ} = \mathbf{\Lambda}_k + \mathbf{U}_r^\top\mathbf{R}\mathbf{V}_r$ bounds the inverses of $\mathbf{X}, \mathbf{Z}$ uniformly. We write $\mathbf{X}_1, \mathbf{X}_2$ for the first $k$ and remaining rows of $\mathbf{X}$. For $k > 0$, we construct a nearby factor pair in $\mathcal{M}_k(t)$. The projector onto the row space of $\mathbf{X}_1$ is

$$\mathbf{\Pi} = \mathbf{X}_1^\top(\mathbf{X}_1\mathbf{X}_1^\top)^{-1}\mathbf{X}_1.$$

We rescale the first $k$ rows so that their Gram matrix equals $t^2\mathbf{I}_k$. We then project the rows of $\mathbf{X}_2$ onto the orthogonal complement of the row space of $\mathbf{X}_1$. This gives

$$\widetilde{\mathbf{X}}_1 = t(\mathbf{X}_1\mathbf{X}_1^\top)^{-\frac{1}{2}}\mathbf{X}_1, \qquad \widetilde{\mathbf{X}}_2 = \mathbf{X}_2(\mathbf{I} - \mathbf{\Pi}).$$

The first $k$ rows of $\mathbf{X}\mathbf{X}^\top - t^2\mathbf{I}$ are $\mathbf{H}\mathbf{U}_r$. If $s_i$ are the eigenvalues of $\mathbf{X}_1\mathbf{X}_1^\top$, then

$$\|\widetilde{\mathbf{X}}_1 - \mathbf{X}_1\|_{\mathrm{F}}^2 = \sum_i (t - \sqrt{s_i})^2 \le t_0^{-2} \sum_i (t^2 - s_i)^2, \qquad \|\widetilde{\mathbf{X}}_2 - \mathbf{X}_2\|_{\mathrm{F}} \le C\|\mathbf{X}_2\mathbf{X}_1^\top\|_{\mathrm{F}}.$$

Thus $\|\widetilde{\mathbf{X}} - \mathbf{X}\|_{\mathrm{F}} \le C\|\mathbf{H}\|_{\mathrm{F}}$. For small $\rho$, $\widetilde{\mathbf{X}}$ remains invertible. The factors $\widetilde{\mathbf{A}} = \mathbf{U}_r\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}^{-1}\mathbf{\Lambda}_k\mathbf{V}_r^\top$ belong to $\mathcal{M}_k(t)$ and differ from $\mathbf{A}, \mathbf{B}$ by $O(d)$. For $k = 0$, we take $\widetilde{\mathbf{X}} = \mathbf{X}$. The map $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{R}, \mathbf{H})$ vanishes on $\mathcal{M}_k(t)$ and is Lipschitz on the bounded neighborhood, uniformly in $t \in J$. Applying this bound to a nearest point in $\mathcal{M}_k(t)$ gives $d \le C\,\mathrm{dist}((\mathbf{A}, \mathbf{B}), \mathcal{M}_k(t))$.

The energy satisfies $\frac{\kappa}{2}d^2 \le \mathcal{E} \le d^2$. For $r = r^\star$, the bound follows from the definition of $\mathcal{E}$. For $r < r^\star$, we use Equation (31) with $\sigma_r(\mathbf{W}_k(t)) \ge \sigma_r$. We write $\mathbf{D} = \mathbf{Y}_r - \mathbf{W}_k$. We split the gradients as

$$\mathbf{G}_A = \mathbf{H}_A + \mathbf{F}_A, \qquad \mathbf{G}_B = \mathbf{H}_B + \mathbf{F}_B.$$

The terms involving $\mathbf{D}$ are

$$\mathbf{H}_A = -\mathbf{D}\mathbf{B}^\top, \qquad \mathbf{H}_B = -\mathbf{A}^\top\mathbf{D}.$$

The other two terms are

$$\mathbf{F}_A = (\mathbf{R} - \mathbf{Y}_\perp)\mathbf{B}^\top, \qquad \mathbf{F}_B = \mathbf{A}^\top(\mathbf{R} - \mathbf{Y}_\perp).$$

The terms $\mathbf{F}_A, \mathbf{F}_B$ vanish on $\mathcal{M}_k(t)$. Their local Lipschitz continuity and the distance bound above imply $\|\mathbf{F}_A\|_{\mathrm{F}} + \|\mathbf{F}_B\|_{\mathrm{F}} \le Cd$. We write $\mathbf{U}_A = -\mathcal{T}_\beta(\mathbf{H}_A)$ and $\mathbf{U}_B = -\mathcal{T}_\beta(\mathbf{H}_B)$. For $k > 0$, $\mathbf{H}_A$ has column space $\mathrm{col}(\mathbf{U}_k)$ and $\mathbf{H}_B$ has row space spanned by $\mathbf{V}_k$. The nonzero singular values of $\mathbf{H}_A, \mathbf{H}_B$ are bounded below: $\sigma_i - t^2 \ge \sigma_k - T^2 > 0$, and the inverses of $\mathbf{X}, \mathbf{Z}$ are bounded. At the factors in Equation (34), normalization scales each active singular direction by $t(\sigma_i - t^2)/\sqrt{t^2(\sigma_i - t^2)^2 + \beta} = 1 + O(\beta)$. Both $\mathbf{H}_A$ and its value at $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ have column space $\mathrm{col}(\mathbf{U}_k)$. The two values of $\mathbf{H}_B$ likewise have row space $\mathrm{col}(\mathbf{V}_k)$. Thus projecting on the left leaves the difference between the two normalized $\mathbf{H}_A$ matrices unchanged. For $\mathbf{H}_B$, the same holds after projection on the right. Combining these estimates from Lemma B.1 with its smoothing bound gives

$$\begin{aligned} \|\mathbf{U}_A\mathbf{B} + \mathbf{A}\mathbf{U}_B - 2t\mathbf{U}_k\mathbf{V}_k^\top\|_{\mathrm{F}} &\le C(d + \beta), \\ \|\mathbf{U}_k^\top(\mathbf{U}_A\mathbf{A}^\top + \mathbf{A}\mathbf{U}_A^\top) - 2t\mathbf{U}_k^\top\|_{\mathrm{F}} &\le C(d + \beta). \end{aligned} \tag{35}$$

Monotonicity gives $\langle \mathbf{F}_A, \dot{\mathbf{A}} \rangle \le \langle \mathbf{F}_A, \mathbf{U}_A \rangle$ and the corresponding bound for $\mathbf{B}$. Since $\mathrm{col}(\mathbf{U}_A) \subseteq \mathrm{col}(\mathbf{U}_k)$ and the row space of $\mathbf{U}_B$ is contained in the span of $\mathbf{V}_k$, the identities $\mathbf{U}_k^\top\mathbf{Y}_\perp = 0$ and $\mathbf{Y}_\perp\mathbf{V}_k = 0$ give $\langle \mathbf{Y}_\perp, \mathbf{U}_A\mathbf{B} + \mathbf{A}\mathbf{U}_B \rangle = 0$. Therefore,

$$\begin{aligned} \dot{\mathcal{E}}_0 &\le \langle \mathbf{R} - \mathbf{Y}_\perp, \mathbf{U}_A\mathbf{B} + \mathbf{A}\mathbf{U}_B \rangle - \langle \mathbf{R}, \dot{\mathbf{W}}_k \rangle \\ &= \langle \mathbf{R}, \mathbf{U}_A\mathbf{B} + \mathbf{A}\mathbf{U}_B - \dot{\mathbf{W}}_k \rangle \le Cd^2 + C\beta d. \end{aligned}$$

For $k > 0$, we write $\dot{\mathbf{A}} = \mathbf{U}_A + \mathbf{\Delta}_A$ and let $\mathbf{P}_R$ project onto $\mathrm{col}(\mathbf{B}\mathbf{V}_k)$. Lemma B.1 yields

$$\|\mathbf{U}_k^\top\mathbf{\Delta}_A\|_{\mathrm{F}} + \|\mathbf{\Delta}_A\mathbf{P}_R\|_{\mathrm{F}} \le Cd, \qquad \|\mathbf{\Delta}_A\|_{\mathrm{F}} \le 2\sqrt{r}.$$

The projector is $(\mathbf{B}\mathbf{V}_k)((\mathbf{B}\mathbf{V}_k)^\top\mathbf{B}\mathbf{V}_k)^{-1}(\mathbf{B}\mathbf{V}_k)^\top$ and has bounded derivatives in the neighborhood. Since $\mathbf{U}_k^\top\mathbf{A}(\mathbf{I} - \mathbf{P}_R) = 0$ on $\mathcal{M}_k(t)$, the same expression is $O(d)$ nearby. Splitting

$$\mathbf{U}_k^\top\mathbf{A}\mathbf{\Delta}_A^\top = \mathbf{U}_k^\top\mathbf{A}(\mathbf{\Delta}_A\mathbf{P}_R)^\top + \mathbf{U}_k^\top\mathbf{A}(\mathbf{I} - \mathbf{P}_R)\mathbf{\Delta}_A^\top$$

gives $\|\mathbf{U}_k^\top(\mathbf{\Delta}_A\mathbf{A}^\top + \mathbf{A}\mathbf{\Delta}_A^\top)\|_{\mathrm{F}} \le Cd$. Indeed,

$$\begin{aligned} \dot{\mathbf{H}} = \mathbf{U}_k^\top(\mathbf{U}_A\mathbf{A}^\top + \mathbf{A}\mathbf{U}_A^\top) - 2t\mathbf{U}_k^\top \\ + \mathbf{U}_k^\top(\mathbf{\Delta}_A\mathbf{A}^\top + \mathbf{A}\mathbf{\Delta}_A^\top). \end{aligned}$$

The first line is bounded by $C(d + \beta)$ by Equation (35), and the second by $Cd$ by the preceding estimate. This gives $\|\dot{\mathbf{H}}\|_{\mathrm{F}} \le C(d + \beta)$. When $k = 0$, the Gram constraint is absent and $\mathbf{H}_A = \mathbf{H}_B = 0$. Monotonicity then gives $\dot{\mathcal{E}}_0 \le 0$. In either case,

$$\dot{\mathcal{E}} \le C\mathcal{E} + C\beta^2, \qquad \mathcal{E}(t) \le e^{C(t-s)}\big(\mathcal{E}(s) + C\beta^2(t - s)\big).$$

Using $\frac{\kappa}{2}d^2 \le \mathcal{E} \le d^2$ gives $d(t) \le C(d(s) + \beta)$. □

*Proof of Theorem 4.1.* At initialization, the gradient with respect to $\mathbf{B}$ is zero. The gradient with respect to $\mathbf{A}$ has size $O(\gamma)$, so we begin on the time scale $t = \gamma s$.

**Growth from initialization.** We write $\mathbf{M} = \mathbf{\Xi}\mathbf{V}_r$ and take $\mathbf{Q}$ to be the orthogonal polar factor of $\mathbf{\Sigma}\mathbf{M}^\top$. By Equation (23), we may rotate $\mathbf{\Xi}$ and $\mathbf{M}$ by left multiplication with $\mathbf{Q}$. We keep the same notation in these coordinates, where $\mathbf{\Sigma}\mathbf{M}^\top$ is positive definite. For an arbitrary parameter sequence satisfying Equation (7), we set

$$\widehat{\mathbf{X}}_j(s) = \frac{\mathbf{X}_j(\gamma_j s)}{\gamma_j}, \quad \widehat{\mathbf{Z}}_j(s) = \frac{\mathbf{Z}_j(\gamma_j s)}{\gamma_j}, \quad \widehat{\mathbf{Z}}_{\perp,j}(s) = \frac{\mathbf{Z}_{\perp,j}(\gamma_j s)}{\gamma_j}, \quad \eta_j = \frac{\beta_j}{\gamma_j^2}.$$

The scaling identity $\mathcal{T}_\beta(\gamma\mathbf{G}) = \mathcal{T}_{\frac{\beta}{\gamma^2}}(\mathbf{G})$ gives

$$\begin{aligned}\widehat{\mathbf{X}}_j' &= \mathcal{T}_{\eta_j}\big(\mathbf{\Sigma}\widehat{\mathbf{Z}}_j^\top - \gamma_j^2\widehat{\mathbf{X}}_j(\widehat{\mathbf{Z}}_j\widehat{\mathbf{Z}}_j^\top + \widehat{\mathbf{Z}}_{\perp,j}\widehat{\mathbf{Z}}_{\perp,j}^\top)\big),\\ [\widehat{\mathbf{Z}}_j' \mid \widehat{\mathbf{Z}}_{\perp,j}'] &= \mathcal{T}_{\eta_j}\big([\widehat{\mathbf{X}}_j^\top\mathbf{\Sigma} - \gamma_j^2\widehat{\mathbf{X}}_j^\top\widehat{\mathbf{X}}_j\widehat{\mathbf{Z}}_j \mid -\gamma_j^2\widehat{\mathbf{X}}_j^\top\widehat{\mathbf{X}}_j\widehat{\mathbf{Z}}_{\perp,j}]\big).\end{aligned}$$

The rescaled flows start at $(0, \mathbf{M}, \mathbf{\Xi}\mathbf{V}_\perp)$ and have uniformly bounded derivatives by Equation (20). Arzelà–Ascoli gives locally uniform subsequential limits $(\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{Z}}_\perp)$. Near zero, $\mathbf{\Sigma}\widehat{\mathbf{Z}}(s)^\top$ is invertible. Lemma B.1 gives $\widehat{\mathbf{X}}' = \mathcal{T}(\mathbf{\Sigma}\widehat{\mathbf{Z}}^\top)$, $\widehat{\mathbf{X}}'(0) = \mathbf{I}$ and $\widehat{\mathbf{X}}(s)/s \to \mathbf{I}$. Hence $\widehat{\mathbf{X}}(s)^\top\mathbf{\Sigma}$ is invertible for small $s > 0$, allowing passage to the other integral equations. We first pass to the limit in the integral equations on $[\varepsilon, s]$. The uniform velocity bound makes the contribution on $[0, \varepsilon]$ at most $C\varepsilon$, so letting $\varepsilon \downarrow 0$ gives

$$\widehat{\mathbf{X}}' = \mathcal{T}(\mathbf{\Sigma}\widehat{\mathbf{Z}}^\top), \qquad \widehat{\mathbf{Z}}' = \mathcal{T}(\widehat{\mathbf{X}}^\top\mathbf{\Sigma}), \qquad \widehat{\mathbf{Z}}_\perp' = 0. \tag{36}$$

The functions $(\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{Z}}_\perp) = (s\mathbf{I}, \mathbf{M} + s\mathbf{I}, \mathbf{\Xi}\mathbf{V}_\perp)$ solve Equation (36): the matrices inside $\mathcal{T}$ are positive definite for $s > 0$, so both polar factors equal $\mathbf{I}$.

To compare the two solutions, we write their differences as $\mathbf{E}_X = \widehat{\mathbf{X}} - s\mathbf{I}$ and $\mathbf{E}_Z = \widehat{\mathbf{Z}} - \mathbf{M} - s\mathbf{I}$. Here $\mu = \lambda_{\min}(\mathbf{\Sigma}\mathbf{M}^\top) > 0$. Along this solution, $\mathbf{\Sigma}(\mathbf{M} + s\mathbf{I})^\top$ has smallest singular value at least $\mu$. The smallest singular value of $s\mathbf{\Sigma}$ is $s\sigma_r$. Lemma B.1 gives

$$\begin{aligned}\|\mathbf{E}_X'(s)\|_{\mathrm{F}} &= \|\mathcal{T}(\mathbf{\Sigma}\widehat{\mathbf{Z}}^\top) - \mathcal{T}(\mathbf{\Sigma}(\mathbf{M} + s\mathbf{I})^\top)\|_{\mathrm{F}} \le \frac{8\|\mathbf{\Sigma}\|_2}{\mu}\|\mathbf{E}_Z(s)\|_{\mathrm{F}},\\ \|\mathbf{E}_Z'(s)\|_{\mathrm{F}} &= \|\mathcal{T}(\widehat{\mathbf{X}}^\top\mathbf{\Sigma}) - \mathcal{T}(s\mathbf{\Sigma})\|_{\mathrm{F}} \le \frac{8\|\mathbf{\Sigma}\|_2}{\sigma_r}\frac{\|\mathbf{E}_X(s)\|_{\mathrm{F}}}{s}.\end{aligned}$$

We choose $C$ to bound both coefficients and set $E(s) = \sup_{0<u\le s}\frac{\|\mathbf{E}_X(u)\|_{\mathrm{F}}}{u} < \infty$. Integration from zero gives

$$\begin{aligned}\|\mathbf{E}_Z(u)\|_{\mathrm{F}} &\le C\int_0^u \frac{\|\mathbf{E}_X(v)\|_{\mathrm{F}}}{v}\,\mathrm{d}v \le CuE(u),\\ \|\mathbf{E}_X(u)\|_{\mathrm{F}} &\le C\int_0^u \|\mathbf{E}_Z(v)\|_{\mathrm{F}}\,\mathrm{d}v \le \frac{C^2u^2}{2}E(u), \qquad E(s) \le \frac{C^2 s}{2}E(s).\end{aligned}$$

For small $s$, the inequality forces $E(s) = 0$. The matrices $\mathbf{\Sigma}(\mathbf{M} + s\mathbf{I})^\top$ and $s\mathbf{\Sigma}$ remain invertible for $s > 0$. The vector field is therefore locally Lipschitz along this solution at every positive time, which extends uniqueness to every finite interval. All subsequential limits agree. Thus for each fixed $s_0 > 0$,

$$\frac{\mathbf{X}(s_0\gamma)}{s_0\gamma} \to \mathbf{I}, \qquad \frac{\mathbf{Z}(s_0\gamma)}{s_0\gamma} \to \mathbf{I} + \frac{\mathbf{M}}{s_0}, \qquad \frac{\mathbf{Z}_\perp(s_0\gamma)}{s_0\gamma} \to \frac{\mathbf{\Xi}\mathbf{V}_\perp}{s_0}. \tag{37}$$

**Stability before the first transition.** To measure deviations from linear growth, we set

$$\widehat{\mathbf{X}} = \frac{\mathbf{X}}{t}, \qquad \widehat{\mathbf{Z}} = \frac{\mathbf{Z}}{t}, \qquad \widehat{\mathbf{Z}}_\perp = \frac{\mathbf{Z}_\perp}{t}.$$

Using logarithmic time $u = \log t$ and writing $\eta = \beta/t^2$, we obtain

$$\begin{aligned}\widehat{\mathbf{X}}_u &= -\mathcal{T}_\eta\big((t^2\widehat{\mathbf{X}}\widehat{\mathbf{Z}} - \mathbf{\Sigma})\widehat{\mathbf{Z}}^\top + t^2\widehat{\mathbf{X}}\widehat{\mathbf{Z}}_\perp\widehat{\mathbf{Z}}_\perp^\top\big) - \widehat{\mathbf{X}},\\ [\widehat{\mathbf{Z}}_u \mid \widehat{\mathbf{Z}}_{\perp,u}] &= -\mathcal{T}_\eta\big([\widehat{\mathbf{X}}^\top(t^2\widehat{\mathbf{X}}\widehat{\mathbf{Z}} - \mathbf{\Sigma}) \mid t^2\widehat{\mathbf{X}}^\top\widehat{\mathbf{X}}\widehat{\mathbf{Z}}_\perp]\big) - [\widehat{\mathbf{Z}} \mid \widehat{\mathbf{Z}}_\perp].\end{aligned} \tag{38}$$

For $t < \tau_r$ and $\eta = 0$, every point of $\mathcal{G} = \{(\mathbf{Q}^\top, \mathbf{Q}, 0) : \mathbf{Q} \in \mathrm{O}(r)\}$ is stationary under Equation (38). We fix $T < \tau_r$ and work on $0 < t \le T$. Minimizing the squared distance to $\mathcal{G}$ is equivalent to maximizing $\mathrm{tr}(\mathbf{R}^\top(\widehat{\mathbf{Z}} + \widehat{\mathbf{X}}^\top))$ over orthogonal $\mathbf{R}$. Invertibility of $\widehat{\mathbf{Z}} + \widehat{\mathbf{X}}^\top$ near $\mathcal{G}$ makes the orthogonal polar factor the unique maximizer [28, Theorem 2.2 and Corollary 2.3]. We set

$$\mathbf{Q} = \mathcal{T}(\widehat{\mathbf{Z}} + \widehat{\mathbf{X}}^\top), \qquad \mathbf{H} = \widehat{\mathbf{X}}\mathbf{Q} - \mathbf{I}, \quad \mathbf{K} = \mathbf{Q}^\top\widehat{\mathbf{Z}} - \mathbf{I}, \quad \mathbf{L} = \mathbf{Q}^\top\widehat{\mathbf{Z}}_\perp,$$
$$\mathcal{E} = \mathrm{dist}((\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{Z}}_\perp), \mathcal{G})^2 = \|\mathbf{H}\|_{\mathrm{F}}^2 + \|\mathbf{K}\|_{\mathrm{F}}^2 + \|\mathbf{L}\|_{\mathrm{F}}^2.$$

We write $\mathrm{sym}\,\mathbf{H} = \frac{\mathbf{H}+\mathbf{H}^\top}{2}$ and $\mathrm{skew}\,\mathbf{H} = \frac{\mathbf{H}-\mathbf{H}^\top}{2}$. The symmetry of $\mathbf{Q}^\top(\widehat{\mathbf{Z}}+\widehat{\mathbf{X}}^\top) = 2\mathbf{I}+\mathbf{K}+\mathbf{H}^\top$ implies $\mathrm{skew}\,\mathbf{H} = \mathrm{skew}\,\mathbf{K}$.

With $\mathbf{D} = \mathbf{\Sigma} - t^2\mathbf{I}$, the derivative of the polar map follows from the expansion

$$-\mathbf{D} + \epsilon\mathbf{F} = (-\mathbf{I} + \epsilon\mathbf{\Omega} + O(\epsilon^2))(\mathbf{D} + \epsilon\mathbf{S} + O(\epsilon^2)),$$

where $\mathbf{\Omega}^\top = -\mathbf{\Omega}$ and $\mathbf{S}^\top = \mathbf{S}$. Equating linear terms gives $\mathbf{F} = \mathbf{\Omega}\mathbf{D} - \mathbf{S}$, so

$$\mathbf{F} - \mathbf{F}^\top = \mathbf{\Omega}\mathbf{D} + \mathbf{D}\mathbf{\Omega}, \qquad \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}\big(-\mathcal{T}(-\mathbf{D} + \epsilon\mathbf{F})\big)\Big|_{\epsilon=0}\right]_{ij} = -\frac{F_{ij} - F_{ji}}{D_{ii} + D_{jj}}.$$

For a perturbation in the remaining columns,

$$-\mathcal{T}([-\mathbf{D} \mid \epsilon\mathbf{L}]) = -(\mathbf{D}^2 + \epsilon^2\mathbf{L}\mathbf{L}^\top)^{-\frac{1}{2}}[-\mathbf{D} \mid \epsilon\mathbf{L}],$$

whose last block has derivative $-\mathbf{D}^{-1}\mathbf{L}$ at zero [20, Lemmas 6–7]. At $(\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{Z}}_\perp) = (\mathbf{I} + \mathbf{H}, \mathbf{I} + \mathbf{K}, \mathbf{L})$, the linear gradient perturbations in (38) are

$$\mathbf{F}_X = t^2(\mathbf{H} + \mathbf{K}) - \mathbf{D}\mathbf{K}^\top, \qquad \mathbf{F}_Z = t^2(\mathbf{H} + \mathbf{K}) - \mathbf{H}^\top\mathbf{D}, \qquad t^2\mathbf{L}.$$

For $\mathbf{C} = \mathrm{skew}\,\mathbf{H} + \mathrm{skew}\,\mathbf{K}$, linearization in a fixed orthogonal basis gives

$$\begin{aligned}(\mathrm{sym}\,\mathbf{H})_u &= -\,\mathrm{sym}\,\mathbf{H}, \quad (\mathrm{sym}\,\mathbf{K})_u = -\,\mathrm{sym}\,\mathbf{K}, \quad \mathbf{L}_u = -(\mathbf{I} + t^2\mathbf{D}^{-1})\mathbf{L},\\ (\mathbf{C}_u)_{ij} &= -2\left(1 + \frac{2t^2}{D_{ii} + D_{jj}}\right)C_{ij} + \frac{D_{jj} - D_{ii}}{D_{ii} + D_{jj}}[\mathrm{sym}(\mathbf{H} - \mathbf{K})]_{ij}.\end{aligned} \tag{39}$$

The skew terms in the two gradients combine as

$$(\mathbf{F}_X - \mathbf{F}_X^\top) + (\mathbf{F}_Z - \mathbf{F}_Z^\top) = 4t^2\mathbf{C} + \mathbf{K}\mathbf{D} - \mathbf{D}\mathbf{K}^\top + \mathbf{D}\mathbf{H} - \mathbf{H}^\top\mathbf{D}.$$

Since $\mathrm{skew}\,\mathbf{H} = \mathrm{skew}\,\mathbf{K}$,

$$\mathcal{E} = \|\,\mathrm{sym}\,\mathbf{H}\|_{\mathrm{F}}^2 + \|\,\mathrm{sym}\,\mathbf{K}\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathbf{C}\|_{\mathrm{F}}^2 + \|\mathbf{L}\|_{\mathrm{F}}^2.$$

Using $|D_{jj} - D_{ii}| \le D_{ii} + D_{jj}$, the linearized flow satisfies

$$\begin{aligned}\mathcal{E}_u &\le -2\|\,\mathrm{sym}\,\mathbf{H}\|_{\mathrm{F}}^2 - 2\|\,\mathrm{sym}\,\mathbf{K}\|_{\mathrm{F}}^2 - 2\|\mathbf{C}\|_{\mathrm{F}}^2 - 2\|\mathbf{L}\|_{\mathrm{F}}^2 + \|\mathbf{C}\|_{\mathrm{F}}\|\,\mathrm{sym}(\mathbf{H} - \mathbf{K})\|_{\mathrm{F}}\\ &\le -\|\,\mathrm{sym}\,\mathbf{H}\|_{\mathrm{F}}^2 - \|\,\mathrm{sym}\,\mathbf{K}\|_{\mathrm{F}}^2 - \frac{3}{2}\|\mathbf{C}\|_{\mathrm{F}}^2 - 2\|\mathbf{L}\|_{\mathrm{F}}^2 \le -\mathcal{E}.\end{aligned}$$

The omitted term $\frac{1}{2}\|\,\mathrm{skew}\,\mathbf{H} - \mathrm{skew}\,\mathbf{K}\|_{\mathrm{F}}^2$ and its derivative vanish when $\mathrm{skew}\,\mathbf{H} = \mathrm{skew}\,\mathbf{K}$. To account for the time dependence of $\mathbf{Q}$, we write $\mathbf{\Omega} = \mathbf{Q}^\top\mathbf{Q}_u$. The additional contribution to $\frac{1}{2}\mathcal{E}_u$ is

$$\langle\mathbf{H}, (\mathbf{I} + \mathbf{H})\mathbf{\Omega}\rangle - \langle\mathbf{K}, \mathbf{\Omega}(\mathbf{I} + \mathbf{K})\rangle - \langle\mathbf{L}, \mathbf{\Omega}\mathbf{L}\rangle = \langle\mathbf{H} - \mathbf{K}, \mathbf{\Omega}\rangle = 0.$$

The quadratic terms vanish by orthogonality of symmetric and skew-symmetric matrices. The linear term vanishes because $\mathbf{H} - \mathbf{K}$ is symmetric. Taylor expansion of the vector field gives a remainder $O(\mathcal{E})$ and hence $\mathcal{E}_u \le -\mathcal{E} + C\mathcal{E}^{3/2}$. The lower bound $D_{ii} \ge \sigma_r - T^2 > 0$ keeps the matrices inside $\mathcal{T}$ full rank near the compact set $\mathcal{G}$. The polar maps therefore have bounded second derivatives on a fixed neighborhood, uniformly for $0 \le t \le T$. Equation (26) bounds the change in the vector field caused by smoothing by $C\eta$. Since $\|\mathrm{D}\mathcal{E}\| = 2\sqrt{\mathcal{E}}$, its contribution to $\mathcal{E}_u$ is at most $C\eta\sqrt{\mathcal{E}}$. On a sufficiently small neighborhood, the negative linear term dominates $C\mathcal{E}^{3/2}$. We obtain

$$\mathcal{E}_u \le -c\mathcal{E} + C\frac{\beta}{t^2}\sqrt{\mathcal{E}} \le -\frac{c}{2}\mathcal{E} + C\beta^2 e^{-4u}.$$

Integration from $u_0 = \log t_0$ yields

$$\begin{aligned} \mathcal{E}(u) &\le e^{-\frac{c}{2}(u-u_0)}\mathcal{E}(u_0) + C\beta^2 \int_{u_0}^{u} e^{-\frac{c}{2}(u-v)} e^{-4v}\,\mathrm{d}v \le \mathcal{E}(u_0) + C\frac{\beta^2}{t_0^4}, \\ \sup_{t_0 \le t \le T} \sqrt{\mathcal{E}(t)} &\le C\left(\sqrt{\mathcal{E}(t_0)} + \frac{\beta}{t_0^2}\right). \end{aligned} \tag{40}$$

If the right side is sufficiently small, the trajectory cannot leave the neighborhood. We take $t_0 = s_0\gamma$. Equation (37) gives $\limsup \sqrt{\mathcal{E}(t_0)} \le \|\mathbf{\Xi}\|_{\mathrm{F}}/s_0$ and $\beta/t_0^2 \to 0$. We first take the parameter limit and then let $s_0 \to \infty$. The orthogonal coordinate change gives $\operatorname{dist}((\mathbf{A},\mathbf{B}), \mathcal{M}_r(t)) \to 0$ uniformly on compact subintervals of $(0, \tau_r)$.

**Continuing through the transitions.** A phase is an open interval between consecutive distinct transition times. At a transition time $\tau$, the numbers of active directions just before and just after $\tau$ are

$$k = \#\{i : \tau_i \ge \tau\}, \qquad \ell = \#\{i : \tau_i > \tau\}.$$

The indices $\ell + 1, \dots, k$ all have target value $\tau^2$. By the induction hypothesis, the distance from $(\mathbf{A}(t), \mathbf{B}(t))$ to $\mathcal{M}_k(t)$ tends to zero uniformly on compact subintervals of $(\tau_{k+1}, \tau)$. For the next phase, we fix a compact interval $I \subset (\tau, \tau_\ell)$ with right endpoint $T = \sup I$. We take

$$0 < \varepsilon < \frac{1}{2}\min\{\tau - \tau_{k+1}, \inf I - \tau\}, \qquad t_\pm = \tau \pm \varepsilon.$$

At $t_-$ the trajectory lies within $o(1)$ of a comparison pair in $\mathcal{M}_k(t_-)$. In the block representation (34) of this pair, we keep $\mathbf{Q}$ and $\mathbf{C}$ fixed. The first $k$ entries of each diagonal $t\mathbf{I}_k$ block increase at unit speed until $\tau$. After $\tau$, entries $\ell + 1, \dots, k$ stay at $\tau$ while the first $\ell$ continue to increase at unit speed. The comparison path lies in $\mathcal{M}_\ell$ after $\tau$. Both paths have speed at most $\sqrt{2r}$, so their distance at $t_+$ is $o(1) + C\varepsilon$.

We use $M = 2 + \|\mathbf{\Xi}\|_{\mathrm{F}} + \sqrt{2r}\,T$. For fixed $\varepsilon$ and sufficiently small parameters, Equation (20) ensures that both the actual and comparison factor pairs have norm less than $M$. We apply Lemma B.3 on $[\tau, T]$ with index $\ell$. The constants in the lemma are independent of $\varepsilon, \beta$. The map $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{R}, \mathbf{H})$ is Lipschitz on the bounded set with a common constant for $t \in [\tau, T]$. It vanishes on the comparison path after $\tau$. The factor distance $o(1) + C\varepsilon$ at $t_+$ therefore gives

$$d_\ell(t_+) \le o(1) + C\varepsilon.$$

We choose $\varepsilon$ and then the parameters so that $C(d_\ell(t_+) + \beta) < \frac{\rho}{2}$. A first time with $d_\ell = \rho$ would contradict Lemma B.3, which gives $d_\ell < \frac{\rho}{2}$ up to that time. The factor bound also holds throughout the interval, so

$$\sup_{t \in I} d_\ell(t) \le C_I\big(o(1) + C\varepsilon + \beta\big). \tag{41}$$

We first take the parameter limit at fixed $\varepsilon$ and then let $\varepsilon \downarrow 0$. This proves convergence on the next phase and also covers the final phase $\ell = 0$.

Thus $\mathbf{W}(t) \to \mathbf{U}_r\mathbf{\Lambda}(t)\mathbf{V}_r^\top$ uniformly on compact intervals away from zero and the transition times. Near each transition, both products are Lipschitz with a common bound independent of $\gamma$ and $\beta$. Extending the estimate across an interval of width $\varepsilon$ therefore adds at most $C\varepsilon$ to the error. Near zero,

$$\|\mathbf{W}(t)\|_{\mathrm{F}} \le \sqrt{r}\,t\big(\gamma\|\mathbf{\Xi}\|_{\mathrm{F}} + \sqrt{r}\,t\big). \tag{42}$$

Taking the parameter limit and then shrinking these neighborhoods proves uniform convergence on every finite time interval. For fixed $T > \tau_1$, loss decrease in Equation (22) gives

$$\sup_{t \ge T} \|\mathbf{W}(t) - \mathbf{Y}\|_{\mathrm{F}} \le \|\mathbf{W}(T) - \mathbf{Y}\|_{\mathrm{F}} \to 0.$$

The limiting product equals $\mathbf{Y}$ for $t \ge T$, which proves Equation (8).

**Singular-value growth rates.** On a compact interval $I \subset (\tau_{k+1}, \tau_k)$ with $k \geq 1$, we estimate the active block $\mathbf{U}_k^\top \dot{\mathbf{W}} \mathbf{V}_k$. At each $t \in I$, we choose a pair in $\mathcal{M}_k(t)$ nearest to $(\mathbf{A}(t), \mathbf{B}(t))$ and use its orthogonal factor $\mathbf{Q}$ from Equation (34). The distance estimate gives factor differences of order $d$. Local Lipschitz continuity of the gradients gives the same bound for their differences. At this pair, the gradients in target coordinates are $-t(\mathbf{\Sigma} - \mathbf{\Lambda}_k)\mathbf{Q}^\top$ and $[-t\mathbf{Q}(\mathbf{\Sigma} - \mathbf{\Lambda}_k) \mid 0]$. Their column and row projectors are $(\mathbf{P}_k, \mathbf{Q}\mathbf{P}_k\mathbf{Q}^\top)$ and $(\mathbf{Q}\mathbf{P}_k\mathbf{Q}^\top, \mathrm{Diag}(\mathbf{P}_k, 0))$. The positive singular values $t(\sigma_i - t^2)$ for $i \leq k$ have a positive lower bound uniform on $I$. Lemma B.1 gives

$$\|\mathbf{P}_k \dot{\mathbf{X}} \mathbf{Q} - \mathbf{P}_k\|_{\mathrm{F}} + \|\mathbf{Q}^\top \dot{\mathbf{Z}} \mathbf{P}_k - \mathbf{P}_k\|_{\mathrm{F}} \leq C_I(d + \beta).$$

Since $\mathbf{Z}\mathbf{P}_k = t\mathbf{Q}\mathbf{P}_k + O(d)$ and $\mathbf{P}_k\mathbf{X} = t\mathbf{P}_k\mathbf{Q}^\top + O(d)$,

$$\mathbf{P}_k(\dot{\mathbf{X}}\mathbf{Z} + \mathbf{X}\dot{\mathbf{Z}})\mathbf{P}_k = t\mathbf{P}_k\dot{\mathbf{X}}\mathbf{Q}\mathbf{P}_k + t\mathbf{P}_k\mathbf{Q}^\top\dot{\mathbf{Z}}\mathbf{P}_k + O(d) = 2t\mathbf{P}_k + O_I(d + \beta).$$

The active block of the product velocity therefore satisfies

$$\sup_{t \in I} \|\mathbf{U}_k^\top \dot{\mathbf{W}}(t) \mathbf{V}_k - 2t\mathbf{I}_k\|_{\mathrm{F}} \longrightarrow 0. \tag{43}$$

To obtain singular-value derivatives, we express the singular values as eigenvalues of the symmetric block matrix $\mathcal{H}(\mathbf{W})$. A singular pair satisfying $\mathbf{W}\mathbf{v} = s\mathbf{u}$ and $\mathbf{W}^\top\mathbf{u} = s\mathbf{v}$ obeys

$$\mathcal{H}(\mathbf{W}) \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = s \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}.$$

The positive eigenvalues of $\mathcal{H}(\mathbf{W})$ in Equation (29) are the positive singular values of $\mathbf{W}$. The largest $k$ eigenvalues converge uniformly to $t^2$ on $I$ and remain separated from zero and the smaller limiting values $\sigma_{k+1}, \ldots, \sigma_r$ by a fixed gap. We write $\mathbf{P}(t)$ for the spectral projector onto the eigenspaces of the $k$ eigenvalues converging to $t^2$, counted with multiplicity. The limiting projector is

$$\mathbf{P}_0 = \mathbf{S}\mathbf{S}^\top, \qquad \mathbf{S} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{U}_k \\ \mathbf{V}_k \end{bmatrix}.$$

We write $\mathbf{S}_0 = \mathcal{H}(\mathbf{U}_r\mathbf{\Lambda}\mathbf{V}_r^\top)$ and $\mathbf{S}_1 = \mathcal{H}(\mathbf{W})$. We take a unit eigenvector $\mathbf{e}$ of $\mathbf{S}_0$ outside the active group and a unit eigenvector $\mathbf{v}$ of $\mathbf{S}_1$ inside the active group. For the associated eigenvalues $a, b$,

$$(a - b)\mathbf{e}^\top\mathbf{v} = \mathbf{e}^\top(\mathbf{S}_0 - \mathbf{S}_1)\mathbf{v}.$$

For sufficiently small parameters, $|a - b|$ is at least half the uniform gap. Summing the squared identities over orthonormal bases and using the equal ranks of the projectors gives

$$\|(\mathbf{I} - \mathbf{P}_0)\mathbf{P}\|_{\mathrm{F}} \leq C_I\|\mathbf{S}_0 - \mathbf{S}_1\|_{\mathrm{F}}, \qquad \|\mathbf{P} - \mathbf{P}_0\|_{\mathrm{F}}^2 = 2\|(\mathbf{I} - \mathbf{P}_0)\mathbf{P}\|_{\mathrm{F}}^2,$$
$$\sup_I \|\mathbf{P} - \mathbf{P}_0\|_{\mathrm{F}} \leq C_I \sup_I \|\mathbf{W} - \mathbf{U}_r\mathbf{\Lambda}\mathbf{V}_r^\top\|_{\mathrm{F}} \to 0.$$

Equation (43) gives $\mathbf{S}^\top\dot{\mathcal{H}}(\mathbf{W})\mathbf{S} = \mathrm{sym}(\mathbf{U}_k^\top\dot{\mathbf{W}}\mathbf{V}_k) \to 2t\mathbf{I}_k$. Uniform boundedness of $\dot{\mathbf{W}}$ then implies

$$\sup_{t \in I} \|\mathbf{P}(t)\dot{\mathcal{H}}(\mathbf{W}(t))\mathbf{P}(t) - 2t\mathbf{P}(t)\|_2 \longrightarrow 0. \tag{44}$$

Weyl's inequality makes the ordered eigenvalues of a Lipschitz symmetric matrix curve Lipschitz and hence differentiable almost everywhere. We fix a time $t$ at which all the ordered eigenvalues are differentiable. For an active eigenvalue $\mu$, $\mathbf{E}$ denotes the projector onto its entire eigenspace. We write $\mathbf{M}(h) = \mathcal{H}(\mathbf{W}(t + h))$ and let $\mathbf{v}_h$ be a unit eigenvector for the same ordered eigenvalue $\mu(h)$. The eigenvector equation gives

$$(\mathbf{M}(0) - \mu(h)\mathbf{I})(\mathbf{I} - \mathbf{E})\mathbf{v}_h = -(\mathbf{I} - \mathbf{E})(\mathbf{M}(h) - \mathbf{M}(0))\mathbf{v}_h.$$

For small $h$, the restriction of $\mathbf{M}(0) - \mu(h)\mathbf{I}$ to $\mathrm{range}(\mathbf{I} - \mathbf{E})$ has a bounded inverse. It follows that $\|(\mathbf{I} - \mathbf{E})\mathbf{v}_h\| = O(|h|)$. We set $\mathbf{u}_h = \frac{\mathbf{E}\mathbf{v}_h}{\|\mathbf{E}\mathbf{v}_h\|}$ and project onto $\mathrm{range}\,\mathbf{E}$. The expansion $\mathbf{M}(h) = \mathbf{M}(0) + h\dot{\mathcal{H}}(\mathbf{W}(t)) + o(|h|)$ yields

$$\frac{\mu(h) - \mu}{h} = \mathbf{u}_h^\top\dot{\mathcal{H}}(\mathbf{W}(t))\mathbf{u}_h + o(1).$$

Since $\mathbf{u}_h \in \operatorname{range}\mathbf{E} \subseteq \operatorname{range}\mathbf{P}(t)$,

$$|\dot{\mu}(t) - 2t| \le \|\mathbf{P}(t)\dot{\mathcal{H}}(\mathbf{W}(t))\mathbf{P}(t) - 2t\mathbf{P}(t)\|_2.$$

The bound holds for every ordered eigenvalue in a repeated eigenspace [55]. Equation (44) gives the limit for $\dot{\lambda}_i$ without requiring separation within the active group. Since $\lambda_i \to t^2 > 0$ uniformly on $I$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\lambda_i(t)} = \frac{\dot{\lambda}_i(t)}{2\sqrt{\lambda_i(t)}} \to 1$$

in essential supremum.

□

*Proof of Lemma 4.2.* We set $a_i(t) = \min\{t^2, \sigma_i\}$ and $\mathbf{E} = \mathbf{W} - \mathbf{U}_r\mathbf{\Lambda}\mathbf{V}_r^\top$. We use $\mathbf{u}_i, \mathbf{v}_i$ for the corresponding columns of $\mathbf{U}_r, \mathbf{V}_r$. The values $a_i$ are nonnegative and nonincreasing, so Weyl's inequality and $d_i - a_i = \mathbf{u}_i^\top \mathbf{E}\mathbf{v}_i$ give

$$|\lambda_i - d_i| \le |\lambda_i - a_i| + |d_i - a_i| \le 2\|\mathbf{E}\|_2 \le 2\|\mathbf{E}\|_{\mathrm{F}}.$$

On a compact interval $I \subset (\tau_{k+1}, \tau_k)$, Equation (43) gives $\dot{d}_i = \mathbf{u}_i^\top \dot{\mathbf{W}}\mathbf{v}_i \to 2t$ uniformly for $i \le k$. The convergence $d_i \to t^2 > 0$ then gives $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i} = \frac{\dot{d}_i}{2\sqrt{d_i}} \to 1$ uniformly on $I$. □

*Proofs of Corollaries 4.3 and 4.4.* We set $t_i = \sqrt{\sigma_i - \varepsilon}$. For sufficiently small fixed $a > 0$, the limiting singular value is below $\sigma_i - \varepsilon$ on $[0, t_i - a]$ and above the threshold at $t_i + a < \tau_i$. Uniform product convergence places $T_i(\varepsilon)$ between $t_i - a$ and $t_i + a$. Letting $a \downarrow 0$ gives the crossing time limit and the strict ordering for distinct targets.

If Corollary 4.4 were false, then for each $j$ we could choose $0 < \gamma_j \le \frac{1}{j}$ and $0 < \beta_j \le \frac{\gamma_j^2}{j}$ with failure probability greater than $\delta$. We couple the resulting flows to a single Gaussian $\mathbf{\Xi}$. The rows of $\mathbf{\Xi}\mathbf{V}_r$ are independent $\mathcal{N}(0, \mathbf{I}_r)$ vectors, so $\mathbf{\Xi}\mathbf{V}_r$ is invertible almost surely. We cover $I$ by finitely many compact intervals contained in individual phases. For almost every $\mathbf{\Xi}$, the product and velocity bounds follow from Theorem 4.1 for all sufficiently large $j$. The crossing-time limit gives the required ordering. Positivity also follows because $t^2 \ge (\inf I)^2 > 0$. The failure indicators therefore converge to zero almost surely. Dominated convergence contradicts the lower bound on the failure probabilities. □

We now keep $\gamma$ and $\beta$ fixed and let $t \to \infty$. Once the trajectory satisfies Equation (46), the reconstruction error bounds the remaining path length of $(\mathbf{A}, \mathbf{B})$.

**Corollary B.4.** *We assume $\mathbf{\Xi}\mathbf{V}_r$ is invertible. There is a $c_0 > 0$ such that the factors converge whenever $0 < \gamma \le c_0$ and $0 < \beta \le c_0\gamma^2$ are fixed. Their limits satisfy $\mathbf{A}_\infty\mathbf{B}_\infty = \mathbf{Y}$. For each such parameter pair, there are $T, C, c > 0$ for which*

$$\|\mathbf{A}(t) - \mathbf{A}_\infty\|_{\mathrm{F}} + \|\mathbf{B}(t) - \mathbf{B}_\infty\|_{\mathrm{F}} + \|\mathbf{W}(t) - \mathbf{Y}\|_{\mathrm{F}} \le Ce^{-c(t-T)}, \qquad t \ge T.$$

*Proof of Corollary B.4.* Our goal is to bound the remaining path length by the reconstruction error $e = \|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}$. We compare the factor speed $v = \|(\dot{\mathbf{A}}, \dot{\mathbf{B}})\|$ with the gradient norm $g = \|(\mathbf{G}_A, \mathbf{G}_B)\|$. We list the singular values of both gradients as $s_1, \ldots, s_{2r}$ with zeros added as needed. For each $i$, we set $v_i = s_i/\sqrt{s_i^2 + \beta}$. Then $g^2 = \sum_i s_i^2$ and $v^2 = \sum_i v_i^2$. The same index maximizes $s_i$ and $v_i$ because the normalization is increasing, so

$$-\dot{\mathcal{L}} = \sum_i s_i v_i \ge (\max_i s_i)(\max_i v_i) \ge \frac{gv}{2r}. \tag{45}$$

We first prove convergence from any time $T$ satisfying

$$\mu = \sigma_{\min}(\mathbf{X}(T)) > 0, \qquad e(T) < \frac{\mu^2}{8r}. \tag{46}$$

As long as $\|\mathbf{X}(t) - \mathbf{X}(T)\|_{\mathrm{F}} < \frac{\mu}{2}$, the smallest singular value satisfies $\sigma_{\min}(\mathbf{X}(t)) \geq \frac{\mu}{2}$. Writing the residual as $\mathbf{AB} - \mathbf{Y} = \mathbf{U}_r\mathbf{R}$ with $\|\mathbf{R}\|_{\mathrm{F}} = e$ gives

$$g \geq \|\mathbf{A}^\top(\mathbf{AB} - \mathbf{Y})\|_{\mathrm{F}} = \|\mathbf{X}^\top\mathbf{R}\|_{\mathrm{F}} \geq \frac{\mu}{2}e.$$

Using $\mathcal{L} = \frac{e^2}{2}$ in Equation (45) gives, for $e > 0$,

$$-\dot{e} \geq \frac{\mu}{4r}v, \qquad \int_T^t v(s)\,\mathrm{d}s \leq \frac{4r}{\mu}e(T) < \frac{\mu}{2}. \tag{47}$$

The length estimate rules out a first time with $\|\mathbf{X}(t) - \mathbf{X}(T)\|_{\mathrm{F}} = \frac{\mu}{2}$. If $e = 0$, the flow is stationary. The factors remain bounded and have finite path length, which implies convergence. Boundedness of the factors gives a uniform upper bound $G$ on the gradient singular values $s_i$ for $t \geq T$. In Equation (45), each term satisfies $s_i v_i \geq s_i^2/\sqrt{G^2 + \beta}$. Thus $g \geq \mu e/2$ and $\mathcal{L} = e^2/2$ imply $-\dot{\mathcal{L}} \geq 2c\mathcal{L}$ with $c = \mu^2/(4\sqrt{G^2 + \beta})$. Integration gives $e(t) \leq e(T)\exp(-c(t - T))$. By Equation (47),

$$\|(\mathbf{A}(t) - \mathbf{A}_\infty, \mathbf{B}(t) - \mathbf{B}_\infty)\| \leq \int_t^\infty v(s)\,\mathrm{d}s \leq \frac{4r}{\mu}e(t).$$

The limits satisfy $\mathbf{A}_\infty\mathbf{B}_\infty = \mathbf{Y}$ and the factor convergence is exponential.

To verify Equation (46), we fix $T > \tau_1$. For $\gamma \leq 1$, the factor bound gives $\|\mathbf{B}(T)\|_2 \leq M := 1 + \|\mathbf{\Xi}\|_{\mathrm{F}} + \sqrt{r}\,T$. Theorem 4.1 gives $e(T) \to 0$. Once $e(T) < \frac{\sigma_r}{2}$,

$$\frac{\sigma_r}{2} \leq \sigma_r(\mathbf{W}(T)) \leq \sigma_{\min}(\mathbf{X}(T))\|\mathbf{B}(T)\|_2, \qquad \sigma_{\min}(\mathbf{X}(T)) \geq \mu_0 := \frac{\sigma_r}{2M} > 0.$$

The singular value inequality follows from $\mathbf{XBB}^\top\mathbf{X}^\top \preceq \|\mathbf{B}\|_2^2\mathbf{XX}^\top$. By convergence as $\gamma \to 0$ and $\beta/\gamma^2 \to 0$, there is $c_0 > 0$ such that $e(T) < \min\{\frac{\sigma_r}{2}, \frac{\mu_0^2}{8r}\}$ whenever $0 < \gamma \leq c_0$ and $0 < \beta \leq c_0\gamma^2$. These inequalities establish Equation (46). □

We first take the limit as $\beta \downarrow 0$ with $\gamma$ fixed. We then study the resulting trajectories as $\gamma \downarrow 0$.

**Corollary B.5.** *For fixed $\gamma > 0$, every sequence $\beta \downarrow 0$ has a subsequence along which the factors converge uniformly on each bounded time interval.*

*When $\mathbf{\Xi}\mathbf{V}_r$ is invertible, we may choose any one of these limits for each $\gamma$. As $\gamma \to 0$, its products and active growth rates satisfy the limits in Theorem 4.1. Its crossing times converge as in Corollary 4.3.*

*Proof.* The speed bound and Arzelà–Ascoli give subsequential limits on every bounded interval. A diagonal subsequence converges locally uniformly on $[0, \infty)$. Along any $\gamma_j \to 0$, we approximate the factors of the chosen subsequential limit within $\frac{1}{j}$ on $[0, j]$ using $\beta_j \leq \frac{\gamma_j^2}{j}$. Theorem 4.1 and the approximation give product convergence on every finite interval. The loss remains nonincreasing under locally uniform convergence, so the product convergence extends to $[0, \infty)$. Uniform product convergence gives the crossing-time limits by the threshold argument used for Corollary 4.3.

For a compact interval $I \subset (\tau_{k+1}, \tau_k)$ with $k \geq 1$, we write $\mathbf{W}$ for the product at $\beta > 0$ and set

$$e(\gamma) = \limsup_{\beta \downarrow 0} \operatorname*{ess\,sup}_{t \in I} \|\mathbf{U}_k^\top\dot{\mathbf{W}}(t)\mathbf{V}_k - 2t\mathbf{I}_k\|_2.$$

If $e(\gamma)$ did not tend to zero, a sequence with $\beta_j \leq \frac{\gamma_j^2}{j}$ would contradict Equation (43). For fixed $\gamma$ and $\epsilon > 0$, the projected velocity error is at most $e(\gamma) + \epsilon$ for sufficiently small $\beta$. Integrating over $[s, t] \subset I$ gives a bound on $\mathbf{U}_k^\top(\mathbf{W}(t) - \mathbf{W}(s))\mathbf{V}_k$. Along the chosen subsequence, the two endpoint values converge to $\mathbf{W}^0(t)$ and $\mathbf{W}^0(s)$. Since $\epsilon > 0$ is arbitrary, we obtain

$$\|\mathbf{U}_k^\top(\mathbf{W}^0(t) - \mathbf{W}^0(s))\mathbf{V}_k - (t^2 - s^2)\mathbf{I}_k\|_2 \leq e(\gamma)(t - s).$$

The inequality bounds the derivative of $\mathbf{U}_k^\top\mathbf{W}^0(t)\mathbf{V}_k - t^2\mathbf{I}_k$ by $e(\gamma)$ almost everywhere. Uniform product convergence preserves the gap between the active singular values and the other singular values on $I$. Since $\mathbf{W}^0$ is Lipschitz, the spectral projector argument applies wherever $\mathbf{W}^0$ and its ordered singular values are differentiable. The projected velocity estimate then gives both speed limits. □

## C Targets of Rank Greater than the Factor Rank

We study SpecGF with $\mathcal{T}_\beta$ under the initialization in Equation (6). When $1 \le r < r^\star$, the product has rank at most $r$ and cannot recover all of $\mathbf{Y}$. We write $\mathbf{Y}_r$ for its best rank-$r$ approximation in Equation (30). The transition times for the first $r$ target values are $\tau_i = \sqrt{\sigma_i}$ with $\tau_{r+1} = 0$. We first ask which left and right singular subspaces the factors approach.

**Theorem C.1.** *We assume $\sigma_r > \sigma_{r+1}$. For Lebesgue-almost every $\mathbf{\Xi} \in \mathbb{R}^{r\times n}$, the limit in Equation* (7) *gives*

$$\sup_{t\ge 0}\Big\|\mathbf{W}(t) - \mathbf{U}_r \mathrm{Diag}\big(\min\{t^2,\sigma_i\}\big)_{i=1}^{r}\mathbf{V}_r^\top\Big\|_{\mathrm{F}} \longrightarrow 0. \tag{48}$$

*The active singular values satisfy Equation* (9)*. For each such $\mathbf{\Xi}$, there exists $c_0 > 0$ depending on $\mathbf{\Xi}$ such that for all $\gamma$ and $\beta$ satisfying*

$$0 < \gamma \le c_0, \qquad 0 < \beta \le c_0\gamma^2,$$

*the factors converge exponentially to finite limits satisfying $\mathbf{A}_\infty\mathbf{B}_\infty = \mathbf{Y}_r$.*

The crossing-time result in Corollary 4.3 applies to these $r$ directions. For Gaussian initialization, the probability bound in Corollary 4.4 also holds. For fixed $\gamma$, every subsequential limit as $\beta \downarrow 0$ satisfies Corollary B.5 with target $\mathbf{Y}_r$.

When $\sigma_r = \sigma_{r+1}$, the limiting rank-$r$ approximation can depend on the initialization. For example, we can take a rank-one factorization of $\mathbf{Y} = \sigma\mathbf{I}_2$. With $\mathbf{\Xi} = \|\mathbf{\Xi}\|\mathbf{v}^\top \neq 0$, the flow preserves $\mathbf{A} = a\mathbf{v}$ and $\mathbf{B} = b\mathbf{v}^\top$. The product thus grows in the direction $\mathbf{v}\mathbf{v}^\top$.

To identify these subspaces, we study the flow on the time scale $s = t/\gamma$. In target coordinates, the two blocks of interest are $\mathbf{U}_{r^\star}^\top\mathbf{A}$ and $\mathbf{V}_{r^\star}^\top\mathbf{B}^\top$. Both have size $r^\star \times r$. At time $t = \gamma s$, we divide both blocks by $\gamma$. For fixed $s$, their limits under Equation (7) are the matrices $\mathbf{X}(s)$ and $\mathbf{Z}(s)$ in the next lemma with $q = r^\star$. Since $\mathbf{X}(0) = 0$, we use the integral form of the equations at the origin. The limits of $\mathbf{X}(s)/s$ and $\mathbf{Z}(s)/s$ identify the singular subspaces in which the factors begin to grow.

**Lemma C.2.** *We write $q$ for the number of diagonal entries of $\mathbf{D} = \mathrm{Diag}(d_1,\dots,d_q)$, ordered as $d_1 \ge \cdots \ge d_q > 0$. We assume $1 \le r < q$ and $d_r > d_{r+1}$. For every full-column-rank $\mathbf{M} \in \mathbb{R}^{q\times r}$, the integral equations*

$$\mathbf{X}' = \mathcal{T}(\mathbf{D}\mathbf{Z}), \qquad \mathbf{Z}' = \mathcal{T}(\mathbf{D}\mathbf{X}), \qquad \mathbf{X}(0) = 0, \quad \mathbf{Z}(0) = \mathbf{M} \tag{49}$$

*have a unique global solution. Both factors are smooth and have full column rank for $s > 0$. For Lebesgue-almost every $\mathbf{M}$, there is an orthogonal $\mathbf{Q}$ such that*

$$\frac{\mathbf{X}(s)}{s} \longrightarrow \mathbf{E}\mathbf{Q}, \qquad \frac{\mathbf{Z}(s)}{s} \longrightarrow \mathbf{E}\mathbf{Q}, \qquad \mathbf{E} = [\mathbf{e}_1,\dots,\mathbf{e}_r], \qquad s \to \infty. \tag{50}$$

*Proof.* We set $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{X}$, $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{Z}$ and $F(\mathbf{P}) = \|\mathbf{D}^{1/2}\mathbf{P}\|_*$. On full column rank matrices,

$$\nabla F(\mathbf{P}) = \mathbf{D}\mathbf{P}(\mathbf{P}^\top\mathbf{D}\mathbf{P})^{-\frac{1}{2}}, \qquad \mathbf{P}' = \nabla F(\mathbf{R}), \quad \mathbf{R}' = \nabla F(\mathbf{P}).$$

The function $F$ is convex and analytic on matrices of full column rank, with a bounded gradient that is homogeneous of degree zero. With $\mathbf{R}_0 = \mathbf{D}^{1/2}\mathbf{M}$, the equations are equivalent near zero to

$$\begin{aligned}\mathbf{R}(s) &= \mathbf{R}_0 + \int_0^s \nabla F\left(\frac{1}{v}\int_0^v \nabla F(\mathbf{R}(w))\,\mathrm{d}w\right)\mathrm{d}v,\\ \mathbf{P}(s) &= \int_0^s \nabla F(\mathbf{R}(w))\,\mathrm{d}w.\end{aligned} \tag{51}$$

The inner average at $v = 0$ is $\nabla F(\mathbf{R}_0)$, which has full column rank. On small neighborhoods of $\mathbf{R}_0$ and $\nabla F(\mathbf{R}_0)$, the integral map preserves a ball of continuous curves and has Lipschitz constant at most $Cs$. For small $s$, the contraction mapping theorem gives a unique solution with $\mathbf{P}(s) = s\nabla F(\mathbf{R}_0) + O(s^2)$ and $\mathbf{R}(s) = \mathbf{R}_0 + O(s)$. The linearized integral operator also has norm at most $Cs < 1$, so its Neumann series gives $C^1$ dependence on $\mathbf{R}_0$. Every continuous solution satisfies Equation (51), since $\mathbf{P}(s)/s \to \nabla F(\mathbf{R}_0)$.

While both factors have full rank, differentiation and convexity give

$$(\mathbf{P}^\top\mathbf{R})' = (\mathbf{R}^\top\mathbf{D}\mathbf{R})^{\frac{1}{2}} + (\mathbf{P}^\top\mathbf{D}\mathbf{P})^{\frac{1}{2}} \succ 0, \qquad \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\|\mathbf{P}-\mathbf{R}\|_{\mathrm{F}}^2 \le 0.$$

For $s > 0$, $\mathbf{P}^\top\mathbf{R}$ is positive definite and its smallest eigenvalue is increasing. On every finite interval away from zero, both factors are bounded and their smallest singular values stay positive. The solution therefore extends to every $s > 0$ and depends $C^1$ on $\mathbf{R}_0$ at each fixed positive time.

We write $m = \lambda_{\min}(\mathbf{P}^\top\mathbf{R})$. Since $\|\mathbf{P}\|_2 + \|\mathbf{R}\|_2 \le C(1+s)$, a unit right singular vector of $\mathbf{P}$ gives $m \le \sigma_{\min}(\mathbf{P})\|\mathbf{R}\|_2$, and the same argument applies with the factors interchanged. Hence, almost everywhere for $s \ge 1$,

$$m' \ge \sqrt{d_q}\big(\sigma_{\min}(\mathbf{P}) + \sigma_{\min}(\mathbf{R})\big) \ge \frac{c}{1+s}m.$$

Consequently $m \to \infty$. For $\mathbf{S} = (\mathbf{P}+\mathbf{R})/2$ and $\mathbf{H} = (\mathbf{P}-\mathbf{R})/2$, $\mathbf{S}^\top\mathbf{S} = \mathbf{P}^\top\mathbf{R} + \mathbf{H}^\top\mathbf{H}$. Thus $\sigma_{\min}(\mathbf{S}) \to \infty$ while $\mathbf{H}$ remains bounded. Lemma B.1 applied to $\nabla F(\mathbf{P}) = \mathbf{D}^{1/2}\mathcal{T}(\mathbf{D}^{1/2}\mathbf{P})$ gives

$$\mathbf{S}' = \nabla F(\mathbf{S}) + \mathbf{E}(s), \qquad \|\mathbf{E}(s)\|_{\mathrm{F}} \le C\frac{\|\mathbf{H}\|_{\mathrm{F}}}{\sigma_{\min}(\mathbf{S})} = o(1).$$

Moreover, $\mathbf{S}^\top\nabla F(\mathbf{S}) = (\mathbf{S}^\top\mathbf{D}\mathbf{S})^{1/2}$ has smallest eigenvalue at least $\sqrt{d_q}\sigma_{\min}(\mathbf{S})$. Differentiating the smallest singular value of $\mathbf{S}$ therefore gives $\frac{\mathrm{d}}{\mathrm{d}s}\sigma_{\min}(\mathbf{S}) \ge \sqrt{d_q} - \|\mathbf{E}(s)\|_2 \ge c$ almost everywhere for large $s$. Since $\mathbf{H}$ is bounded,

$$\min\{\sigma_{\min}(\mathbf{P}), \sigma_{\min}(\mathbf{R})\} \ge cs, \qquad \max\{\|\mathbf{P}\|_2, \|\mathbf{R}\|_2\} \le Cs \tag{52}$$

eventually.

For $u = \log s$, we set $\widehat{\mathbf{S}} = (\mathbf{P}+\mathbf{R})/(2s)$ and $\widehat{\mathbf{H}} = (\mathbf{P}-\mathbf{R})/(2s)$. The normalized system is

$$\begin{aligned}\widehat{\mathbf{S}}_u &= \frac{1}{2}\big(\nabla F(\widehat{\mathbf{S}}+\widehat{\mathbf{H}}) + \nabla F(\widehat{\mathbf{S}}-\widehat{\mathbf{H}})\big) - \widehat{\mathbf{S}},\\ \widehat{\mathbf{H}}_u &= \frac{1}{2}\big(\nabla F(\widehat{\mathbf{S}}-\widehat{\mathbf{H}}) - \nabla F(\widehat{\mathbf{S}}+\widehat{\mathbf{H}})\big) - \widehat{\mathbf{H}}.\end{aligned} \tag{53}$$

By Equation (52), the normalized trajectory eventually stays in a compact set where $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{S}} \pm \widehat{\mathbf{H}}$ have full column rank. Since $\mathbf{H}$ is bounded, $\widehat{\mathbf{H}} = O(e^{-u})$. We set $\mathbf{G} = \nabla F(\widehat{\mathbf{S}}) - \widehat{\mathbf{S}}$. Bounded Hessians and convexity imply

$$\|\widehat{\mathbf{S}}_u - \mathbf{G}\|_{\mathrm{F}} \le C\|\widehat{\mathbf{H}}\|_{\mathrm{F}}, \qquad \|\widehat{\mathbf{H}}_u\|_{\mathrm{F}} \le C\|\widehat{\mathbf{H}}\|_{\mathrm{F}}, \qquad \langle\widehat{\mathbf{H}}, \widehat{\mathbf{H}}_u\rangle \le -\|\widehat{\mathbf{H}}\|_{\mathrm{F}}^2.$$

For a sufficiently large fixed $a$, the analytic function $\Psi = F(\widehat{\mathbf{S}}) - \frac{1}{2}\|\widehat{\mathbf{S}}\|_{\mathrm{F}}^2 - a\|\widehat{\mathbf{H}}\|_{\mathrm{F}}^2$ therefore satisfies

$$\Psi_u \ge c(\|\mathbf{G}\|_{\mathrm{F}}^2 + \|\widehat{\mathbf{H}}\|_{\mathrm{F}}^2), \qquad \|(\widehat{\mathbf{S}}_u, \widehat{\mathbf{H}}_u)\| \le C(\|\mathbf{G}\|_{\mathrm{F}} + \|\widehat{\mathbf{H}}\|_{\mathrm{F}}). \tag{54}$$

The first bound follows by absorbing $C\|\mathbf{G}\|_{\mathrm{F}}\|\widehat{\mathbf{H}}\|_{\mathrm{F}}$ into the two squares. Boundedness of $\Psi$ gives integrability of both squares. Uniform continuity then gives $\mathbf{G}, \widehat{\mathbf{H}} \to 0$. Every accumulation point is critical and has the same value $\Psi_\infty$. Applying Lemma D.6 on a finite cover of the compact set of accumulation points gives $0 < \alpha < 1$ such that $(\Psi_\infty - \Psi)^\alpha \le C\|\nabla\Psi\|$ eventually. Since $\nabla\Psi = (\mathbf{G}, -2a\widehat{\mathbf{H}})$, Equation (54) gives

$$\Psi_u \ge c\|\nabla\Psi\|\|(\widehat{\mathbf{S}}_u, \widehat{\mathbf{H}}_u)\|, \qquad -\frac{\mathrm{d}}{\mathrm{d}u}(\Psi_\infty - \Psi)^{1-\alpha} \ge c\|(\widehat{\mathbf{S}}_u, \widehat{\mathbf{H}}_u)\|.$$

If $\Psi = \Psi_\infty$, the trajectory is stationary. Otherwise the last inequality gives finite length. Hence $(\widehat{\mathbf{S}}, \widehat{\mathbf{H}}) \to (\mathbf{S}_\infty, 0)$ with $\nabla F(\mathbf{S}_\infty) = \mathbf{S}_\infty$. For $\mathbf{R}_\infty = (\mathbf{S}_\infty^\top\mathbf{D}\mathbf{S}_\infty)^{1/2}$,

$$\mathbf{D}\mathbf{S}_\infty = \mathbf{S}_\infty\mathbf{R}_\infty, \qquad \mathbf{R}_\infty = \mathbf{S}_\infty^\top\mathbf{S}_\infty.$$

Thus the column space of $\mathbf{S}_\infty$ is invariant under $\mathbf{D}$, and an orthonormal eigenbasis gives $\mathbf{S}_\infty = \mathbf{E}_J\mathrm{Diag}(\sqrt{d_i} : i \in J)\mathbf{Q}$ with $|J| = r$ and $\mathbf{Q}$ orthogonal. Within a repeated eigenspace, any orthonormal basis is allowed.

We vary $\mathbf{R}_0$ while keeping $\mathbf{P}(0) = 0$. The variations satisfy

$$\frac{\mathrm{d}}{\mathrm{d}s}\langle \delta\mathbf{P}, \delta\mathbf{R}\rangle = \langle \delta\mathbf{R}, \nabla^2 F(\mathbf{R})[\delta\mathbf{R}]\rangle + \langle \delta\mathbf{P}, \nabla^2 F(\mathbf{P})[\delta\mathbf{P}]\rangle \geq 0.$$

The $C^1$ dependence on the initial matrix makes the pairing tend to zero at $s = 0$. Therefore

$$\|\delta\widehat{\mathbf{S}}\|_{\mathrm{F}}^2 - \|\delta\widehat{\mathbf{H}}\|_{\mathrm{F}}^2 = \frac{\langle \delta\mathbf{P}, \delta\mathbf{R}\rangle}{s^2} \geq 0. \tag{55}$$

If $\delta\widehat{\mathbf{S}} = 0$ at a fixed $s > 0$, then $\delta\widehat{\mathbf{H}} = 0$ and backward uniqueness of the linear variation equations gives $\delta\mathbf{R}_0 = 0$. Thus $\mathbf{R}_0 \mapsto \widehat{\mathbf{S}}$ is a local $C^1$ diffeomorphism, and the image of the initial values is locally a graph $\widehat{\mathbf{H}} = \psi(\widehat{\mathbf{S}})$ with $\|\mathrm{D}\psi\| \leq 1$.

At $(\mathbf{S}_\infty, 0)$, the Hessian $\mathcal{A} = \nabla^2 F(\mathbf{S}_\infty)$ is positive semidefinite. The linearization is

$$\delta\widehat{\mathbf{S}}_u = \mathcal{A}[\delta\widehat{\mathbf{S}}] - \delta\widehat{\mathbf{S}}, \qquad \delta\widehat{\mathbf{H}}_u = -\mathcal{A}[\delta\widehat{\mathbf{H}}] - \delta\widehat{\mathbf{H}}.$$

If the limiting subspace is non-leading, some omitted eigenvalue $d_j$ exceeds an eigenvalue $d_i$ in that subspace. We choose an eigenbasis of $\mathbf{D}$ and a right orthogonal basis so that column $\ell$ of $\mathbf{S}_\infty$ is $\sqrt{d_i}\mathbf{e}_i$. For $\mathbf{H} = \mathbf{e}_j\mathbf{e}_\ell^\top$, $\mathbf{S}_\infty^\top\mathbf{D}\mathbf{H} = 0$ and $\mathcal{A}[\mathbf{H}] = (d_j/d_i)\mathbf{H}$. Hence at least one $\widehat{\mathbf{S}}$ direction is unstable, while every $\widehat{\mathbf{H}}$ direction is strictly stable.

We apply the local center-stable manifold theorem to the map that advances Equation (53) by one unit of $u$ [38, Theorem 7]. The map is a local diffeomorphism near the equilibrium. We write $\widehat{\mathbf{S}} - \mathbf{S}_\infty = \mathbf{S}_+ + \mathbf{S}_-$. Here $\mathbf{S}_+$ is the projection onto the eigenspaces of $\mathcal{A}$ with eigenvalues greater than one. The remaining component is $\mathbf{S}_-$. On a sufficiently small neighborhood, every orbit that remains in the neighborhood satisfies

$$\mathbf{S}_+ = \phi(\mathbf{S}_-, \widehat{\mathbf{H}}), \qquad \|\mathrm{D}\phi\| < \frac{1}{2},$$

because the manifold is tangent to the stable and neutral subspace. On the graph $\widehat{\mathbf{H}} = \psi(\widehat{\mathbf{S}})$, the derivative of $\mathbf{S}_+ - \phi(\mathbf{S}_-, \psi(\widehat{\mathbf{S}}))$ with respect to $\mathbf{S}_+$ is $I - \mathrm{D}_{\widehat{\mathbf{H}}}\phi\,\mathrm{D}_{\mathbf{S}_+}\psi$, which is invertible by Equation (55). The intersection therefore has positive codimension and Lebesgue measure zero in the initial coordinates.

We cover the non-leading equilibria by countably many smaller neighborhoods whose closures lie inside the neighborhoods where the manifold theorem applies. A trajectory with a non-leading limit stays in the corresponding larger neighborhood after some integer $u = j$. For each neighborhood and integer $j$, we cover the initial domain by countably many local diffeomorphism charts at $s = e^j$. The preimages have measure zero, and the countable union contains every initial value with a non-leading limit. For almost every $\mathbf{M}$, the columns of $\mathbf{S}_\infty$ therefore span the leading $r$-dimensional invariant subspace of $\mathbf{D}$. Multiplication by $\mathbf{D}^{-\frac{1}{2}}$ then gives Equation (50). □

*Proof of Theorem C.1.* The speed and descent bounds in Equations (20) and (22) hold for every target rank, and the columns of $\mathbf{A}$ remain in $\mathrm{col}(\mathbf{U}_{r^\star})$. We set $\widehat{\mathbf{A}}(s) = \mathbf{A}(\gamma s)/\gamma$, $\widehat{\mathbf{B}}(s) = \mathbf{B}(\gamma s)/\gamma$ and $\eta = \beta/\gamma^2$. Then

$$\widehat{\mathbf{A}}' = \mathcal{T}_\eta\big((\mathbf{Y} - \gamma^2\widehat{\mathbf{A}}\widehat{\mathbf{B}})\widehat{\mathbf{B}}^\top\big), \qquad (\widehat{\mathbf{B}}^\top)' = \mathcal{T}_\eta\big((\mathbf{Y}^\top - \gamma^2\widehat{\mathbf{B}}^\top\widehat{\mathbf{A}}^\top)\widehat{\mathbf{A}}\big). \tag{56}$$

We first consider directions $\mathbf{\Xi}$ for which $\mathbf{M} = (\mathbf{\Xi}\mathbf{V}_{r^\star})^\top$ has full column rank. Uniform velocity bounds give locally uniform subsequential limits. Near zero, the first limiting gradient has full rank because $\mathbf{Y}\mathbf{\Xi}^\top$ does. Hence $\widehat{\mathbf{A}}(s)/s \to \mathcal{T}(\mathbf{Y}\mathbf{\Xi}^\top)$. The matrix $\mathbf{Y}^\top\mathcal{T}(\mathbf{Y}\mathbf{\Xi}^\top)$ also has full column rank, since $\mathbf{Y}^\top$ is injective on $\mathrm{col}(\mathbf{Y})$. The second integral equation therefore passes to the limit for small $s > 0$, and bounded velocities allow its lower endpoint to tend to zero.

In target coordinates, the limit is Equation (49) with $\mathbf{D} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_{r^\star})$ and initial matrix $\mathbf{M}$. The component of $\widehat{\mathbf{B}}^\top$ orthogonal to $\mathbf{V}_{r^\star}$ is constant in the limit. Uniqueness and full rank in Lemma C.2 identify the limit on every bounded $s$ interval. The map $\mathbf{\Xi} \mapsto (\mathbf{\Xi}\mathbf{V}_{r^\star})^\top$ is a coordinate

projection after an orthogonal change of basis. By Fubini's theorem, the preimage of the exceptional set in Lemma C.2 has Lebesgue measure zero. For every remaining $\boldsymbol{\Xi}$,

$$\lim_{s_0\to\infty} \limsup_{\gamma\to 0,\ \frac{\beta}{\gamma^2}\to 0} \operatorname{dist}\left(\left(\frac{\mathbf{A}(\gamma s_0)}{\gamma s_0}, \frac{\mathbf{B}(\gamma s_0)^\top}{\gamma s_0}\right), \mathcal{G}\right) = 0, \qquad \mathcal{G} = \{(\mathbf{U}_r\mathbf{Q}, \mathbf{V}_r\mathbf{Q}) : \mathbf{Q} \in \mathrm{O}(r)\}. \tag{57}$$

Before $\tau_r$, we normalize the factors as $\widehat{\mathbf{A}} = \mathbf{A}/t$ and $\widehat{\mathbf{B}} = \mathbf{B}/t$. In logarithmic time $u = \log t$, the smoothing parameter becomes $\eta = \beta/t^2$. The normalized vector field is the right side of Equation (56) with $\gamma^2$ replaced by $t^2$ and with $-\widehat{\mathbf{A}}$ and $-\widehat{\mathbf{B}}^\top$ added. Near $\mathcal{G}$, the nearest orthogonal matrix and the errors are

$$\mathbf{Q} = \mathcal{T}(\mathbf{U}_r^\top\widehat{\mathbf{A}} + \mathbf{V}_r^\top\widehat{\mathbf{B}}^\top), \qquad \mathbf{H} = \widehat{\mathbf{A}}\mathbf{Q}^\top - \mathbf{U}_r, \quad \mathbf{K} = \widehat{\mathbf{B}}^\top\mathbf{Q}^\top - \mathbf{V}_r,$$
$$\mathcal{E} = \operatorname{dist}((\widehat{\mathbf{A}}, \widehat{\mathbf{B}}^\top), \mathcal{G})^2 = \|\mathbf{H}\|_{\mathrm{F}}^2 + \|\mathbf{K}\|_{\mathrm{F}}^2, \qquad \mathbf{H}_0 = \mathbf{U}_r^\top\mathbf{H}, \quad \mathbf{K}_0 = \mathbf{V}_r^\top\mathbf{K}.$$

The polar decomposition makes $\mathbf{H}_0 + \mathbf{K}_0$ symmetric. The leading blocks obey Equation (39) after replacing its $\mathbf{H}, \mathbf{K}$ by $\mathbf{H}_0, \mathbf{K}_0^\top$. Under this linearized flow, the squared norm of $(\mathbf{H}_0, \mathbf{K}_0)$ decays exponentially in $u$.

For $1 \le i \le r < j \le r^\star$, we write $h_{ji} = (\mathbf{u}_j^\top\mathbf{H})_i$ and $k_{ji} = (\mathbf{v}_j^\top\mathbf{K})_i$. The linearized gradient entries are $\sigma_j k_{ji} - t^2 h_{ji}$ and $\sigma_j h_{ji} - t^2 k_{ji}$. The derivative of the polar factor in an extra row divides column $i$ by $\sigma_i - t^2$, giving

$$\frac{\mathrm{d}}{\mathrm{d}u}\begin{bmatrix} h_{ji} \\ k_{ji} \end{bmatrix} = \frac{1}{\sigma_i - t^2}\begin{bmatrix} -\sigma_i & \sigma_j \\ \sigma_j & -\sigma_i \end{bmatrix}\begin{bmatrix} h_{ji} \\ k_{ji} \end{bmatrix}.$$

Both eigenvalues are negative and $(\sigma_i - \sigma_j)/(\sigma_i - t^2) \ge \kappa$. Components outside the full target spaces have coefficient $-\sigma_i/(\sigma_i - t^2)$. The motion of $\mathbf{Q}$ contributes $-\langle \mathbf{H}_0 + \mathbf{K}_0, \mathbf{Q}_u\mathbf{Q}^\top\rangle = 0$ to $\mathcal{E}_u/2$. For $0 \le t \le T < \tau_r$, the reference gradient singular values are at least $\sigma_r - T^2 > 0$. The polar maps therefore have uniformly bounded second derivatives near the compact set $\mathcal{G}$. Taylor expansion and Lemma B.1 give

$$\mathcal{E}_u \le -c\mathcal{E} + C\mathcal{E}^{\frac{3}{2}} + C\frac{\beta}{t^2}\sqrt{\mathcal{E}} \le -\frac{c}{2}\mathcal{E} + C\frac{\beta^2}{t^4}$$

on a fixed smaller neighborhood. Integration gives Equation (40) for the present $\mathcal{E}$. We take $t_0 = \gamma s_0$ and use Equation (57), first taking the parameter limit and then letting $s_0 \to \infty$. A first exit argument proves convergence to $\mathcal{M}_r(t)$ uniformly on compact subintervals of $(0, \tau_r)$.

Every transition time $\tau$ for the first $r$ directions satisfies $\tau \ge \tau_r$, so Lemma B.3 applies on $[\tau, T]$. The comparison path used in Equation (41) then gives convergence to $\mathcal{M}_k(t)$ on every subsequent compact phase interval, including $k = 0$. The uniform velocity bounds and Equation (42) extend product convergence to every finite interval. For fixed $T > \tau_1$, Equation (32) and loss decrease yield

$$\sup_{t\ge T}\|\mathbf{W}(t) - \mathbf{Y}_r\|_{\mathrm{F}}^2 \le \frac{2}{\kappa}\big(\mathcal{L}(\mathbf{A}(T), \mathbf{B}(T)) - \mathcal{L}_\star\big) \longrightarrow 0,$$

which proves Equation (48).

We fix a compact active interval $I \subset (\tau_{k+1}, \tau_k)$. For each $t \in I$, we choose a factor pair in $\mathcal{M}_k(t)$ nearest to $(\mathbf{A}(t), \mathbf{B}(t))$. We write $\mathbf{Q}_a$ for the first $k$ columns of its orthogonal factor $\mathbf{Q}$ in Equation (34). The factor gradients at the reference point are

$$-t\mathbf{U}_k\mathrm{Diag}(\sigma_i - t^2)_{i\le k}\mathbf{Q}_a^\top, \qquad -t\mathbf{Q}_a\mathrm{Diag}(\sigma_i - t^2)_{i\le k}\mathbf{V}_k^\top,$$

because the discarded target annihilates the reference factors. The nonzero singular values of both reference gradients are uniformly bounded below on $I$. The actual factor gradients approach the gradients at these nearest pairs uniformly on $I$. The projected perturbation bound in Lemma B.1 therefore gives

$$\|\mathbf{U}_k^\top\dot{\mathbf{A}} - \mathbf{Q}_a^\top\|_{\mathrm{F}} + \|\dot{\mathbf{B}}\mathbf{V}_k - \mathbf{Q}_a\|_{\mathrm{F}} \longrightarrow 0$$

uniformly on $I$. Combining the projected gradient bounds with $\mathbf{B}\mathbf{V}_k = t\mathbf{Q}_a + o(1)$ and $\mathbf{U}_k^\top\mathbf{A} = t\mathbf{Q}_a^\top + o(1)$ proves Equation (43). The spectral projector argument applies because the active singular

values are separated from the remaining spectrum on $I$, the product converges uniformly and the product velocity is bounded. Hence Equation (9) holds, including at repeated singular values almost everywhere.

To prove convergence at fixed parameters, we fix $T > \tau_1$. Convergence in the final phase and the factor bound place the state at $T$ arbitrarily close to a fixed compact subset of $\mathcal{M}_0$. We write $e = \sqrt{2(\mathcal{L} - \mathcal{L}_\star)}$, $g = \|\nabla\mathcal{L}\|$ and $v = \|(\dot{\mathbf{A}}, \dot{\mathbf{B}})\|$. Lemma B.2 gives $g \geq ce$ on a uniform neighborhood. The descent estimate Equation (45) implies

$$-\dot{e} \geq cv, \qquad \int_T^t v(s)\,\mathrm{d}s \leq Ce(T).$$

For sufficiently small parameters, the ball of radius $Ce(T)$ around the state at time $T$ lies inside this neighborhood. The length bound excludes a first exit and gives bounded factors of finite path length. If $G$ bounds the gradient singular values there, then

$$-\dot{\mathcal{L}} \geq \frac{g^2}{\sqrt{G^2 + \beta}} \geq c(\mathcal{L} - \mathcal{L}_\star).$$

Thus $e(t) \leq Ce^{-c(t-T)}$, and integrating the length bound from $t$ to infinity gives the same exponential rate for the factors. Equation (32) gives the product limit $\mathbf{Y}_r$ at the same rate, and the loss tends to $\mathcal{L}_\star = \frac{1}{2}\sum_{i>r}\sigma_i^2$. By joint convergence, the entry conditions hold for all $0 < \gamma \leq c_0$ and $0 < \beta \leq c_0\gamma^2$ once $c_0$ is small enough.

The crossing-time argument in Corollary 4.3 applies to the first $r$ values. For Gaussian initialization, we couple all flows to the same $\boldsymbol{\Xi}$. The exceptional set has probability zero, so dominated convergence gives Corollary 4.4. As $\beta \downarrow 0$, the product and velocity bounds pass to each subsequential limit by the argument in Corollary B.5. The excess-loss bound in Equation (32) controls the product for all $t \geq T$. □

The columns $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_{r^\star}$ are the right singular vectors omitted from $\mathbf{Y}_r$. If each row of $\mathbf{B}(0)$ is orthogonal to these vectors, the residual $\mathbf{Y} - \mathbf{Y}_r$ contributes zero to both factor gradients at initialization. The proof below shows that the two gradients continue to agree with those for $\mathbf{Y}_r$ at every later time. We can then apply the rank-matched result directly to $\mathbf{Y}_r$.

**Proposition C.3.** *We write $\mathbf{V}_{>r}$ for columns $r+1, \ldots, r^\star$ of $\mathbf{V}_{r^\star}$. We assume $\boldsymbol{\Xi}\mathbf{V}_r$ is invertible and $\boldsymbol{\Xi}\mathbf{V}_{>r} = 0$. Then Theorem 4.1 and Corollaries B.4 and 4.3 hold with target $\mathbf{Y}_r$ even when $\sigma_r = \sigma_{r+1}$.*

*Proof.* On $\{\mathbf{A} = \mathbf{U}_r\mathbf{X},\ \mathbf{B}\mathbf{V}_{>r} = 0\}$,

$$(\mathbf{Y} - \mathbf{Y}_r)\mathbf{B}^\top = 0, \qquad \mathbf{A}^\top(\mathbf{Y} - \mathbf{Y}_r) = 0.$$

The factor gradients for $\mathbf{Y}$ and $\mathbf{Y}_r$ therefore agree. Preservation of column spaces and right null spaces by $\mathcal{T}_\beta$ makes both constraints invariant. Uniqueness identifies the trajectory with the trajectory for $\mathbf{Y}_r$, to which the results for target rank $r$ apply. □

# D Convergence to Global Minima of SpecGF

## D.1 Analyticity of SpecGF

We first state that the inverse square root of symmetric positive definite (SPD) matrices is analytic. This directly follows from Bauer et al. [5, Lemma 4.4 (b)].

**Lemma D.1.** *At any real SPD $\mathbf{S}$, the function*

$$g(\mathbf{S}) = \mathbf{S}^{-\frac{1}{2}}$$

*is analytic. That is, $g$ is analytic on the (open) set of the symmetric positive-definite matrices.*

*Proof.* Any SPD matrix $\mathbf{S}$ is a closed, invertible, (densely) defined linear operator on a Banach space (also a Hilbert space for our case), denoted as $X$, and has real eigenvalues. Throughout this proof, we abuse the notation $\sigma(\mathbf{S})$ to denote the spectrum of $\mathbf{S}$.

We first show $\mathbf{S}$ is sectorial of any $\omega \in [0, \pi)$; fix any $\omega$. $\sigma(\mathbf{S})$ is exactly the set of eigenvalues of $\mathbf{S}$, and therefore is contained in $\overline{S_\omega}$. Next, for each $\omega' \in (\omega, \pi)$, take any $\lambda \notin \overline{S_{\omega'}}$. Then $|\mathrm{Arg}(\lambda)| > \omega'$ and thus, for any eigenvalue $s$ of $\mathbf{S}$,

$$|s - \lambda| \geq \mathrm{dist}(\lambda, (0, +\infty)) = \begin{cases} |\lambda||\sin(\mathrm{Arg}(\lambda))| \geq |\lambda| \sin \omega' & \text{if } |\mathrm{Arg}(\lambda)| \leq \dfrac{\pi}{2}, \\ |\lambda| \geq |\lambda| \sin \omega' & \text{if } |\mathrm{Arg}(\lambda)| > \dfrac{\pi}{2}. \end{cases}$$

Hence,

$$\|\lambda(\mathbf{S} - \lambda)^{-1}\| = \max_{s \in \sigma(\mathbf{S})} \frac{|\lambda|}{|s - \lambda|} \leq \frac{1}{\sin \omega'}$$

is uniformly bounded. On a Hilbert space, R-boundedness is equivalent to boundedness [5]. Thus, $\mathbf{S}$ is R-sectorial of any $\omega \in [0, \pi)$. We also see that $\mathbf{S}$ admits a bounded $\mathcal{H}^\infty(S_\omega)$ calculus for any $\omega \in (0, \pi)$, because for $f \in \mathcal{H}^\infty(S_\omega) \setminus \{0\}$,

$$\|f(\mathbf{S})\|_{L(X)} = \max_{s \in \sigma(\mathbf{S})} |f(s)| \leq \|f\|_{\mathcal{H}^\infty(S_\omega)}$$

As $\mathbf{S}$ is SPD, there exists a closed centered ball not in $\sigma(\mathbf{S})$. Now take $r = \frac{1}{2}$ in the statement of Bauer et al. [5, Lemma 4.4 (b)]. Then $g(\mathbf{S})$ is holomorphic; thus, it is real-analytic when restricted to the set of real matrices. As the dimension is finite, the fractional domain spaces $\dot{X}_p$ coincide with $X$ (viewed as sets), and thus $L(\dot{X}_p, \dot{X}_q) = L(X)$ for any real $p$ and $q$ (so are $L(\dot{X}_0, \dot{X}_{<r})$ and $L(\dot{X}_{>1-r}, \dot{X}_r)$). □

**Corollary D.2.** *Consider a matrix $\mathbf{X} \in \mathbb{R}^{m \times r}$ with $m \geq r$.*

(i) *$\mathbf{X} \mapsto (\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}}$ is analytic on the (open) set of full-rank matrices.*

(ii) *For every $\beta > 0$, $\mathbf{X} \mapsto (\mathbf{X}^\top \mathbf{X} + \beta \mathbf{I}_r)^{-\frac{1}{2}}$ is analytic everywhere.*

*The above statements analogously hold for $\mathbf{X} \in \mathbb{R}^{r \times n}$ with $n \geq r$ by switching $\mathbf{X}^\top \mathbf{X}$ to $\mathbf{X}\mathbf{X}^\top$.*

*Proof.* This is apparent from Lemma D.1 and the facts that polynomials are analytic and that compositions of analytic functions are analytic. □

Recall that, for a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$,

$$\mathcal{T}(\mathbf{M}) = \left( \left( \mathbf{M}\mathbf{M}^\top \right)^\dagger \right)^{1/2} \mathbf{M},$$
$$\mathcal{T}_\beta(\mathbf{M}) = \left( \mathbf{M}\mathbf{M}^\top + \beta \mathbf{I}_m \right)^{-1/2} \mathbf{M},$$

Let the compact SVD of $\mathbf{X} \in \mathbb{R}^{m \times n}$ be $\mathbf{U}_r \mathbf{S} \mathbf{V}_r^\top$ with $\mathbf{S} \triangleq \mathrm{Diag}\{\sigma_1, \ldots, \sigma_r\}$. $\sigma_i > 0$ are the positive singular values of $\mathbf{X}$, $1 \leq i \leq r \leq \min\{m, n\}$. Then,

$$\mathcal{T}(\mathbf{X}) = \mathbf{U}_r \mathbf{V}_r^\top,$$
$$\mathcal{T}_\beta(\mathbf{X}) = \mathbf{U}_r \mathbf{S}_\beta \mathbf{V}_r^\top,$$

where

$$\mathbf{S}_\beta \triangleq \mathrm{Diag}\left\{ \frac{\sigma_i}{\sqrt{\sigma_i^2 + \beta}} \right\}_{i \in [r]}.$$

**Proposition D.3** (Analyticity of $\mathcal{T}$ and $\mathcal{T}_\beta$)**.** *SpecGF with $\mathcal{T}_\beta$ is analytic everywhere and its solution uniquely exists. SpecGF with $\mathcal{T}$ is analytic and its solution uniquely exists when both factor gradients are full-rank.*

*Proof.* Analyticity is straightforward from Corollary D.2, so we prove the existence and uniqueness part. For both $\mathcal{T}$ and $\mathcal{T}_\beta$, $\|\dot{\mathbf{A}}\|_{\mathrm{F}}^2$ and $\|\dot{\mathbf{B}}\|_{\mathrm{F}}^2$ are both bounded by $r$. Hence, the trajectory should satisfy

$$\|(\mathbf{A}(t), \mathbf{B}(t))\|_{\mathrm{F}} \leq \|(\mathbf{A}(0), \mathbf{B}(0))\|_{\mathrm{F}} + \sqrt{2r}t.$$

If the maximal solution does not exist for all $t \geq 0$, then the iterate should leave every compact set of the domain. Therefore, the solution must globally exist for $\mathcal{T}_\beta$. Uniqueness follows from the local Lipschitzness (by the Picard-Lindelöf theorem). For $\mathcal{T}$, the solution exists uniquely as long as both factor gradients have full rank. □

### D.2 Convergence Analysis of SpecGF

We restate the setup of SpecGF.

- The target matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ yields a compact SVD $\mathbf{Y} = \mathbf{U}_{r^\star}\mathbf{\Sigma}\mathbf{V}_{r^\star}^\top$ where $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \sigma_2, \cdots, \sigma_{r^\star})$ with $\sigma_1 \geq \cdots \geq \sigma_{r^\star} > 0$.
- The loss function $\mathcal{L} : \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n} \to \mathbb{R}$ is defined as

$$\mathcal{L}(\mathbf{A}, \mathbf{B}) \triangleq \frac{1}{2}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}^2. \tag{58}$$

  We may abuse the notation and simply state $\mathcal{L}(t)$.
- The time derivatives of $\mathbf{A}$ and $\mathbf{B}$ are governed by either $\mathcal{T}$ or $\mathcal{T}_\beta$ applied to the gradient:

$$\dot{\mathbf{A}}(t) = -\mathcal{T}\left(\nabla_{\mathbf{A}}\mathcal{L}(t)\right), \quad \dot{\mathbf{B}}(t) = -\mathcal{T}\left(\nabla_{\mathbf{B}}\mathcal{L}(t)\right).$$

**Lemma D.4.** *On SpecGF with $\mathcal{T}$, the loss is monotonically non-increasing:*

$$\frac{d\mathcal{L}(t)}{\mathrm{d}t} \leq 0.$$

*The equality holds if and only if $\nabla_{\mathbf{A}}\mathcal{L}(t) = \nabla_{\mathbf{B}}\mathcal{L}(t) = 0$ holds.*

*Proof.* We first show that for any matrix $\mathbf{X}$, $\langle \mathcal{T}(\mathbf{X}), \mathbf{X}\rangle = \|\mathbf{X}\|_*$. Let the compact SVD of $\mathbf{X}$ be $\mathbf{U}'\mathbf{S}'\mathbf{V}'^\top$.

$$\langle \mathcal{T}(\mathbf{X}), \mathbf{X}\rangle = \mathrm{tr}\left((\mathbf{U}'\mathbf{V}'^\top)^\top\mathbf{U}'\mathbf{S}'\mathbf{V}'^\top\right) = \mathrm{tr}\left(\mathbf{V}'\mathbf{S}'\mathbf{V}'^\top\right) = \mathrm{tr}\left(\mathbf{S}'\right) = \|\mathbf{X}\|_*.$$

We drop the argument $t$ for simplicity.

$$\begin{aligned}\frac{d\mathcal{L}}{\mathrm{d}t} &= \left\langle \nabla_{\mathbf{A}}\mathcal{L}, \dot{\mathbf{A}}\right\rangle + \left\langle \nabla_{\mathbf{B}}\mathcal{L}, \dot{\mathbf{B}}\right\rangle \\ &= \langle \nabla_{\mathbf{A}}\mathcal{L}, -\mathcal{T}\left(\nabla_{\mathbf{A}}\mathcal{L}\right)\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}, -\mathcal{T}\left(\nabla_{\mathbf{B}}\mathcal{L}\right)\rangle \\ &= -\|\nabla_{\mathbf{A}}\mathcal{L}\|_* - \|\nabla_{\mathbf{B}}\mathcal{L}\|_* \leq 0.\end{aligned}$$

□

**Lemma D.5.** *On SpecGF with $\mathcal{T}_\beta$, the loss is monotonically non-increasing:*

$$\frac{d\mathcal{L}(t)}{\mathrm{d}t} \leq 0.$$

*The equality holds if and only if $\nabla_{\mathbf{A}}\mathcal{L}(t) = \nabla_{\mathbf{B}}\mathcal{L}(t) = 0$ holds.*

*Proof.* We first simplify $\langle \mathcal{T}_\beta(\mathbf{X}), \mathbf{X}\rangle$ for a matrix $\mathbf{X}$. Let the SVD of $\mathbf{X}$ be $\mathbf{U}'\mathbf{S}'\mathbf{V}'^\top$ with singular values $\sigma_1, \ldots, \sigma_r \geq 0$.

$$\langle \mathcal{T}_\beta(\mathbf{X}), \mathbf{X}\rangle = \mathrm{tr}\left((\mathbf{U}'\mathbf{S}'_\beta\mathbf{V}'^\top)^\top\mathbf{U}'\mathbf{S}'\mathbf{V}'^\top\right) = \mathrm{tr}\left(\mathbf{V}'\mathbf{S}'_\beta\mathbf{S}'\mathbf{V}'^\top\right) = \mathrm{tr}\left(\mathbf{S}'_\beta\mathbf{S}'\right) = \sum_{i\in[r]}\frac{\sigma_i^2}{\sqrt{\sigma_i^2+\beta}}.$$

We drop the argument $t$ for simplicity. Let $\sigma_i(\nabla_{\mathbf{A}}\mathcal{L}), \sigma_i(\nabla_{\mathbf{B}}\mathcal{L})$ be the singular values of $\nabla_{\mathbf{A}}\mathcal{L}$ and $\nabla_{\mathbf{B}}\mathcal{L}$, respectively. Then:

$$\begin{aligned}\frac{d\mathcal{L}}{\mathrm{d}t} &= \left\langle \nabla_{\mathbf{A}}\mathcal{L}, \dot{\mathbf{A}}\right\rangle + \left\langle \nabla_{\mathbf{B}}\mathcal{L}, \dot{\mathbf{B}}\right\rangle \\ &= \langle \nabla_{\mathbf{A}}\mathcal{L}, -\mathcal{T}_\beta\left(\nabla_{\mathbf{A}}\mathcal{L}\right)\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}, -\mathcal{T}_\beta\left(\nabla_{\mathbf{B}}\mathcal{L}\right)\rangle \\ &= -\sum_{i\in[r]}\frac{\sigma_i(\nabla_{\mathbf{A}}\mathcal{L})^2}{\sqrt{\sigma_i(\nabla_{\mathbf{A}}\mathcal{L})^2+\beta}} - \sum_{i\in[r]}\frac{\sigma_i(\nabla_{\mathbf{B}}\mathcal{L})^2}{\sqrt{\sigma_i(\nabla_{\mathbf{B}}\mathcal{L})^2+\beta}} \leq 0.\end{aligned}$$

□

Thus, $\mathcal{L}(t)$ is monotonically non-increasing and lower bounded; hence, $\mathcal{L}(\infty) \le \mathcal{L}(0)$ should exist. Moreover,

$$\|\mathbf{A}(t)\mathbf{B}(t)\|_{\mathrm{F}} \le \|\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}\|_{\mathrm{F}} + \|\mathbf{Y}\|_{\mathrm{F}} = \sqrt{2\mathcal{L}(t)} + \|\mathbf{Y}\|_{\mathrm{F}} \le \sqrt{2\mathcal{L}(0)} + \|\mathbf{Y}\|_{\mathrm{F}}. \tag{59}$$

We now show that the SpecGF iterate should either converge or diverge to infinity.

**Lemma D.6** (Łojasiewicz gradient inequality, [42, 18])**.** *Let $F : U \to \mathbb{R}$ be a real analytic function on an open set $U \subset \mathbb{R}^d$. Then, for every point $\mathbf{p} \in U$, there exist a (possibly small) neighborhood $W$ of $\mathbf{p}$ and constants $C > 0$ and $b \in [0, 1)$ such that for all $\mathbf{x} \in W$,*

$$|F(\mathbf{x}) - F(\mathbf{p})|^b \le C\|\nabla F(\mathbf{x})\|. \tag{60}$$

*In addition, if $\mathbf{p}$ is a critical point of $F$, then Equation* (60) *holds for $b \in [\frac{1}{2}, 1)$.*

**Proposition D.7.** *Let $\mathbf{x}(t)$ be a differentiable and absolutely continuous curve on $\mathbb{R}^d$ and $F$ be a real analytic function on $\mathbb{R}^d$. Assume that $F(\mathbf{x}(t))$ is well-defined for all $t \ge 0$ and the following holds:*

$$\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t} = 0 \quad \Rightarrow \quad \dot{\mathbf{x}}(t) = 0. \tag{61}$$

*If there exists a constant $c > 0$ such that*

$$\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t} \le -c\|\nabla F(\mathbf{x}(t))\|\|\dot{\mathbf{x}}(t)\| \tag{62}$$

*for all $t \ge 0$, then either $\mathbf{x}(t)$ converges in $\mathbb{R}^d$ or it diverges to infinity (in norm).*

*Proof.* Assume that $\|\mathbf{x}(t)\| \not\to \infty$ as $t \to \infty$. Then $\mathbf{x}(t)$ has an accumulation point in $\mathbb{R}^d$; it suffices to prove that the accumulation point is the limit of $\mathbf{x}(t)$. Denote the accumulation point as $\bar{\mathbf{x}}$.

By continuity of $F$ and monotonicity of $F$ from Equation (62), $F(\mathbf{x}(t)) \downarrow F(\bar{\mathbf{x}})$. If there exists a finite time $t' \ge 0$ such that $F(\mathbf{x}(t')) = F(\bar{\mathbf{x}})$, then $F(\mathbf{x}(t)) = F(\bar{\mathbf{x}})$ for $t \ge t'$; hence, $\mathbf{x}(t) = \bar{\mathbf{x}}$ by Equation (61). From here we assume that $F(\mathbf{x}(t)) > F(\bar{\mathbf{x}})$ for all $t$.

From Lemma D.6 and Equation (62), we see that

$$\begin{aligned}\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t} &\le -c\|\nabla F(\mathbf{x}(t))\|\|\dot{\mathbf{x}}(t)\| \\ &\le -\frac{c}{C}|F(\mathbf{x}(t)) - F(\bar{\mathbf{x}})|^b\|\dot{\mathbf{x}}(t)\| = -\frac{c}{C}\left(F(\mathbf{x}(t)) - F(\bar{\mathbf{x}})\right)^b\|\dot{\mathbf{x}}(t)\|\end{aligned}$$

on some neighborhood $W$ of $\bar{\mathbf{x}}$ with $C > 0$ and $b \in [0, 1)$. Hence,

$$-\|\dot{\mathbf{x}}(t)\| \ge \frac{C}{c}\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t}\left(F(\mathbf{x}(t)) - F(\bar{\mathbf{x}})\right)^{-b} = \frac{C}{c(1-b)}\frac{\mathrm{d}}{\mathrm{d}t}\left(F(\mathbf{x}(t)) - F(\bar{\mathbf{x}})\right)^{1-b}.$$

For brevity, let $G(t) \triangleq F(\mathbf{x}(t)) - F(\bar{\mathbf{x}}) > 0$. Then $G$ is absolutely continuous and $G(t) \downarrow 0$. We also see that

$$-\|\dot{\mathbf{x}}(t)\| \ge \frac{C}{c(1-b)}\frac{\mathrm{d}}{\mathrm{d}t}\left(G(t)^{1-b}\right). \tag{63}$$

We now show that $\mathbf{x}(t)$ eventually moves to the vicinity of $\bar{\mathbf{x}}$ and remains there forever. As $W$ is open, there exists $r > 0$ where

$$B(r) \triangleq \{z \in \mathbb{R}^d : \|z - \bar{\mathbf{x}}\| < r\} \subset W.$$

Fix any such $r > 0$. Because $\bar{\mathbf{x}}$ is the accumulation point of $\mathbf{x}(t)$ and $G$ continuously decreases to zero, there exists a time $t_1 \ge 0$ such that

$$\|x(t_1) - \bar{\mathbf{x}}\| < \frac{r}{2}, \quad G(t_1) < \left(\frac{rc(1-b)}{2C}\right)^{\frac{1}{1-b}}.$$

Notice the former implies $x(t_1) \in B(r)$ and the latter implies $\frac{C}{c(1-b)} G(t_1)^{1-b} < \frac{r}{2}$. Suppose the claim is false, so there exists $t_2 > t_1$ where $x(t_2) \notin B(r)$; choose the smallest such $t_2$. Then $\mathbf{x}(t) \in B(r)$ for $t \in [t_1, t_2)$, hence

$$\begin{aligned}\|x(t_2) - x(t_1)\| &\le \int_{t_1}^{t_2} \|\dot{\mathbf{x}}(t)\| \mathrm{d}t \\ &\le \int_{t_1}^{t_2} -\frac{C}{c(1-b)} \frac{\mathrm{d}}{\mathrm{d}t} \left(G(t)^{1-b}\right) \mathrm{d}t \\ &= \frac{C}{c(1-b)} \left[G(t_1)^{1-b} - G(t_2)^{1-b}\right] < \frac{r}{2}.\end{aligned}$$

The first inequality and the equality follow from the absolute continuity of $\mathbf{x}(t)$ and $G(t)$ on $[t_1, t_2]$, respectively. The second inequality stems from Equation (63). The above yields a contradiction:

$$\|x(t_2) - \bar{\mathbf{x}}\| \le \|x(t_2) - x(t_1)\| + \|x(t_1) - \bar{\mathbf{x}}\| < r \quad \Rightarrow \quad x(t_2) \in B(r).$$

Therefore, for all $t \ge t_1$, $\mathbf{x}(t) \in B(r)$. As $r$ can be arbitrarily small, $\mathbf{x}(t) \to \bar{\mathbf{x}}$ should hold. □

It is evident that $\mathcal{L}(\mathbf{A}, \mathbf{B})$ is analytic everywhere. For $\mathcal{T}_\beta$, $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are continuously differentiable as their time derivatives $-\mathcal{T}_\beta(\nabla_{\mathbf{A}} \mathcal{L}(t))$ and $-\mathcal{T}_\beta(\nabla_{\mathbf{B}} \mathcal{L}(t))$ are analytic. In addition, $\mathbf{M} = 0 \Leftrightarrow \mathcal{T}_\beta(\mathbf{M}) = 0$; hence, Equation (61) is satisfied. Moreover, let the ranks of $\nabla_{\mathbf{A}} \mathcal{L}(t)$ and $\nabla_{\mathbf{B}} \mathcal{L}(t)$ be $r_A$ and $r_B$, respectively. Sort the singular values $\sigma(t)$ of the gradients with respect to $\mathbf{A}$ and $\mathbf{B}$ in non-increasing order. Then:

$$\begin{aligned}\frac{d\mathcal{L}(t)}{\mathrm{d}t} &= -\sum_{i \in [r_A + r_B]} \sigma_i(t) \frac{\sigma_i(t)}{\sqrt{\sigma_i(t)^2 + \beta}} \\ &\le -\frac{1}{r_A + r_B} \sum_{i \in [r_A + r_B]} \sigma_i(t) \sum_{i \in [r_A + r_B]} \frac{\sigma_i(t)}{\sqrt{\sigma_i(t)^2 + \beta}} \\ &\le -\frac{1}{r_A + r_B} \|\nabla \mathcal{L}(t)\|_{\mathrm{F}} \left\| (\dot{\mathbf{A}}, \dot{\mathbf{B}}) \right\|_{\mathrm{F}} \\ &\le -\frac{1}{2r} \|\nabla \mathcal{L}(t)\|_{\mathrm{F}} \left\| (\dot{\mathbf{A}}, \dot{\mathbf{B}}) \right\|_{\mathrm{F}}.\end{aligned}$$

The first inequality follows from Chebyshev's sum inequality. Hence, Equation (62) is also met. Therefore, the SpecGF with $\mathcal{T}_\beta$ can only either converge to a critical point or diverge to infinity.

For $\mathcal{T}$, assuming that iteration is always full rank (for the analyticity),

$$\begin{aligned}\frac{d\mathcal{L}(t)}{\mathrm{d}t} &= -\|\nabla_{\mathbf{A}} \mathcal{L}(t)\|_* - \|\nabla_{\mathbf{B}} \mathcal{L}(t)\|_* \\ &\le -\|\nabla_{\mathbf{A}} \mathcal{L}(t)\|_{\mathrm{F}} - \|\nabla_{\mathbf{B}} \mathcal{L}(t)\|_{\mathrm{F}} \le -\|\nabla \mathcal{L}(t)\|_{\mathrm{F}} \\ &\le -\frac{1}{\sqrt{2r}} \|\nabla \mathcal{L}(t)\|_{\mathrm{F}} \left\| (\dot{\mathbf{A}}, \dot{\mathbf{B}}) \right\|_{\mathrm{F}}.\end{aligned}$$

The last inequality follows from the fact that $\|\mathcal{T}(\mathbf{M})\|_{\mathrm{F}}^2 \le \mathrm{rank}(\mathbf{M})$ for any matrix $\mathbf{M}$.

**Invariance.** We conjecture that SpecGF may not have a nontrivial invariant term with respect to $\mathbf{A}(t)$ and $\mathbf{B}(t)$, at least for polynomials, while plain GF has one. Here, an "invariant" is a function $I : \mathbb{R}^d \to \mathbb{R}$ such that for all solutions $\mathbf{x}(t)$ of the differential equation $\dot{\mathbf{x}}(t) = f(\mathbf{x})$, $I(\mathbf{x}(t))$ is constant (possibly depending on $\mathbf{x}(0)$). Then, $\frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{x}(t)) \equiv 0$, or

$$\langle \nabla_{\mathbf{x}} I, f(\mathbf{x}) \rangle \equiv 0$$

globally holds on the set of initial points where the solution is well-defined. As evidence for this conjecture, we show that every invariant polynomial is constant in the scalar case. Let $m = n = 1 = r$. Then $\mathbf{Y} = \sigma \neq 0$, $a(t), b(t) \in \mathbb{R}$, and thus the loss becomes

$$\mathcal{L}(a, b) = \frac{1}{2}(ab - \sigma)^2.$$

We state the partial derivatives for convenience:

$$\partial_a \mathcal{L} = b(ab - \sigma), \quad \partial_b \mathcal{L} = a(ab - \sigma).$$

**Proposition D.8** (Absence of polynomial invariants for $\mathcal{T}_\beta$)**.** *For SpecGF with $\mathcal{T}_\beta$, there does not exist a non-constant invariant polynomial with real coefficients in $a$ and $b$ on all of $\mathbb{R} \times \mathbb{R}$.*

*Proof.* Note that, as $\mathbb{R}$ is a field, $\mathbb{R}[a][b] \cong \mathbb{R}[a, b] \cong \mathbb{R}[b][a]$. In addition, $\mathbb{R}[x]$ is a unique factorization domain (UFD) for an indeterminate $x$.

Let $P \in \mathbb{R}[a, b]$ be a polynomial (of finite degree) with real coefficients. Then

$$\partial_a P(a(t), b(t))\dot{a}(t) + \partial_b P(a(t), b(t))\dot{b}(t) = 0.$$

Substitute $\dot{a}(t) = -\mathcal{T}_\beta(\partial_a \mathcal{L})$ and $\dot{b}(t) = -\mathcal{T}_\beta(\partial_b \mathcal{L})$. Rearranging the equations gives that, for any $a$ and $b$,

$$\partial_a P(a, b) \times b(ab - \sigma)\sqrt{a^2(ab - \sigma)^2 + \beta} + \partial_b P(a, b) \times a(ab - \sigma)\sqrt{b^2(ab - \sigma)^2 + \beta} = 0.$$

For simplicity, denote $a^2(ab - \sigma)^2 + \beta$ and $b^2(ab - \sigma)^2 + \beta$ as $Q_{(a)}$ and $Q_{(b)}$ in $\mathbb{R}[a, b]$, respectively. Then

$$\left(\partial_a P(a, b)\right)^2 \times b^2 Q_{(a)} = \left(\partial_b P(a, b)\right)^2 \times a^2 Q_{(b)}.$$

Plugging $a = 0$ gives $\left(\partial_a P(0, b)\right)^2 \times b^2\beta = 0$, which implies that $a \mid \partial_a P(a, b)$. Plugging $b = 0$ implies $b \mid \partial_b P(a, b)$ by analogous reasoning. We see that there exist polynomials $P'_{(a)}$ and $P'_{(b)}$ such that

$$\partial_a P(a, b) = aP'_{(a)}, \quad \partial_b P(a, b) = bP'_{(b)}.$$

Thus, for any $a$ and $b$,

$$P'^2_{(a)} Q_{(a)} = P'^2_{(b)} Q_{(b)}.$$

As $\beta > 0$, the discriminant of $Q_{(a)}$, viewed as a quadratic in $b$, is

$$(-2a^3\sigma)^2 - 4(a^4)(a^2\sigma^2 + \beta) = -4a^4\beta,$$

which is not a perfect square in $\mathbb{R}[a]$. A quadratic whose discriminant is not a perfect square has no roots in $\mathbb{R}[a]$. $\mathbb{R}[a]$ being a UFD allows us to use Gauss' lemma; hence, $Q_{(a)}$ is (irreducible, thus) a prime in $\mathbb{R}[a, b]$. Similarly, $Q_{(b)}$ is also a prime in $\mathbb{R}[a, b]$. We now show that $Q_{(a)}$ and $Q_{(b)}$ are coprime. Indeed, for every common factor $D$ of $Q_{(a)}$ and $Q_{(b)}$, $D$ should also divide

$$b^2 Q_{(a)} - a^2 Q_{(b)} = \beta(b^2 - a^2) = \beta(b - a)(b + a).$$

Since $b - a$ and $b + a$ are coprime, either $D \mid (b - a)$ or $D \mid (b + a)$. However, $Q_{(a)}(a, a) = a^2(a^2 - \sigma)^2 + \beta \not\equiv 0$ and $Q_{(a)}(a, -a) = a^2(-a^2 - \sigma)^2 + \beta \not\equiv 0$; so $D = 1$.

Therefore, we have

$$\begin{aligned} Q_{(a)} \mid P'^2_{(b)} \quad &\Rightarrow \quad Q_{(a)} \mid P'_{(b)}, \\ Q_{(b)} \mid P'^2_{(a)} \quad &\Rightarrow \quad Q_{(b)} \mid P'_{(a)}. \end{aligned}$$

Then, there exist polynomials $P''_{(a)}$ and $P''_{(b)}$ such that

$$P'_{(a)} = Q_{(b)} P''_{(a)}, \quad P'_{(b)} = Q_{(a)} P''_{(b)}.$$

Thus,

$$P''^2_{(a)} Q_{(b)} = P''^2_{(b)} Q_{(a)}.$$

Then, by a similar argument, there exist polynomials $P'''_{(a)}$ and $P'''_{(b)}$ such that

$$P''_{(a)} = Q_{(a)} P'''_{(a)}, \quad P''_{(b)} = Q_{(b)} P'''_{(b)}.$$

Thus,

$$P'''^2_{(a)} Q_{(a)} = P'''^2_{(b)} Q_{(b)}.$$

The above reasoning implies that $P'_{(a)}$ is divisible by $Q^k_{(a)}$ for all positive integers $k \geq 1$, which can only happen when $P'_{(a)} \equiv 0$; we also have $P'_{(b)} \equiv 0$. Then $\partial_a P(a, b) \equiv 0$ and $\partial_b P(a, b) \equiv 0$. Therefore, such an invariant $P$ has to be constant. □

A similar result is obtained for $\mathcal{T}$ on the set where SpecGF with $\mathcal{T}$ is analytic.

**Proposition D.9** (Absence of polynomial invariants for $\mathcal{T}$)**.** *For SpecGF with $\mathcal{T}$, there does not exist a non-constant invariant polynomial with real coefficients in $a$ and $b$ on*

$$O \triangleq \{(a,b) : a \neq 0, b \neq 0, ab - \sigma \neq 0\}.$$

*Proof.* Note that on $\mathbb{R}$,

$$\mathcal{T}(x) = \operatorname{sgn}(x) \quad (x \neq 0),$$

where $\operatorname{sgn}$ is the sign function.

Let $P$ be a polynomial (of finite degree) with real coefficients. Then

$$\partial_a P(a(t), b(t))\dot{a}(t) + \partial_b P(a(t), b(t))\dot{b}(t) = 0.$$

Substitute $\dot{a}(t) = -\mathcal{T}(\partial_a \mathcal{L})$ and $\dot{b}(t) = -\mathcal{T}(\partial_b \mathcal{L})$. Rearranging the equations gives that, for any $a$ and $b$ in $O$,

$$\partial_a P(a,b) \times \operatorname{sgn}(b(ab-\sigma)) + \partial_b P(a,b) \times \operatorname{sgn}(a(ab-\sigma)) = 0.$$

If $\operatorname{sgn}(a) = \operatorname{sgn}(b)$, we have

$$\partial_a P(a,b) + \partial_b P(a,b) = 0.$$

As the equality holds on the open subset, we actually have

$$\partial_a P(a,b) + \partial_b P(a,b) \equiv 0$$

on all of $\mathbb{R} \times \mathbb{R}$. Analogously, if $\operatorname{sgn}(a) = -\operatorname{sgn}(b)$, we have

$$\partial_a P(a,b) - \partial_b P(a,b) = 0,$$

so in fact, on all of $\mathbb{R} \times \mathbb{R}$,

$$\partial_a P(a,b) - \partial_b P(a,b) \equiv 0.$$

Therefore $\partial_a P(a,b)$ and $\partial_b P(a,b)$ are identically zero everywhere, implying that such an invariant $P$ has to be constant. □

In addition, we show that $\mathbf{A}(t)^\top \mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top$ is not an invariant for SpecGF with both $\mathcal{T}$ and $\mathcal{T}_\beta$.

**Proposition D.10.** *For both SpecGF with $\mathcal{T}_\beta$ and $\mathcal{T}$, there exists $\mathbf{Y}$ such that $\mathbf{A}(t)^\top \mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top$ cannot be an invariant.*

*Proof.* Fix $\sigma > 0$ and $a, b \neq 0$ with $ab \neq \sigma$, $|a| \neq |b|$. We show that with

$$\mathbf{Y} = \begin{bmatrix} \sigma \mathbf{I}_{r^\star} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

for any $r^\star \geq r$, SpecGF with $\mathcal{T}_\beta$ cannot have $\mathbf{A}(t)^\top \mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top$ as its invariant, due to its non-zero time derivative at

$$(\mathbf{A}, \mathbf{B}) = \left( \begin{bmatrix} a\mathbf{I}_r \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} b\mathbf{I}_r & \mathbf{0} \end{bmatrix} \right)$$

Compute the gradient at $(\mathbf{A}, \mathbf{B})$:

$$\nabla_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} b(ab-\sigma)\mathbf{I}_r \\ \mathbf{0} \end{bmatrix} \triangleq \begin{bmatrix} a'\mathbf{I}_r \\ \mathbf{0} \end{bmatrix}$$
$$\nabla_{\mathbf{B}} \mathcal{L}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} a(ab-\sigma)\mathbf{I}_r & \mathbf{0} \end{bmatrix} \triangleq \begin{bmatrix} b'\mathbf{I}_r & \mathbf{0} \end{bmatrix}.$$

Then, for SpecGF with $\mathcal{T}_\beta$

$$\dot{\mathbf{A}} = -\mathcal{T}_\beta\left(\nabla_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \mathbf{B})\right) = \begin{bmatrix} -\frac{a'}{\sqrt{a'^2+\beta}}\mathbf{I}_r \\ \mathbf{0} \end{bmatrix}$$
$$\dot{\mathbf{B}} = -\mathcal{T}_\beta\left(\nabla_{\mathbf{B}} \mathcal{L}(\mathbf{A}, \mathbf{B})\right) = \begin{bmatrix} -\frac{b'}{\sqrt{b'^2+\beta}}\mathbf{I}_r & \mathbf{0} \end{bmatrix}.$$

Hence, the time derivatives of $\mathbf{A}$ and $\mathbf{B}$ translate to those of $a$ and $b$, respectively. We obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{A}^\top\mathbf{A}-\mathbf{B}\mathbf{B}^\top) &= \frac{\mathrm{d}}{\mathrm{d}t}(a^2-b^2)\mathbf{I}_r \\ &= -2\left(a\frac{a'}{\sqrt{a'^2+\beta}}-b\frac{b'}{\sqrt{b'^2+\beta}}\right)\mathbf{I}_r \\ &= -2ab(ab-\sigma)\left(\frac{1}{\sqrt{b^2(ab-\sigma)^2+\beta}}-\frac{1}{\sqrt{a^2(ab-\sigma)^2+\beta}}\right)\mathbf{I}_r \neq \mathbf{0}.\end{aligned}$$

The same setting results in the same result for SpecGF with $\mathcal{T}$. For SpecGF with $\mathcal{T}$

$$\begin{aligned}\dot{\mathbf{A}} &= -\mathcal{T}\left(\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{A},\mathbf{B})\right) = \begin{bmatrix}-\mathrm{sgn}(a')\mathbf{I}_r \\ \mathbf{0}\end{bmatrix} \\ \dot{\mathbf{B}} &= -\mathcal{T}\left(\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{A},\mathbf{B})\right) = \begin{bmatrix}-\mathrm{sgn}(b')\mathbf{I}_r & \mathbf{0}\end{bmatrix}.\end{aligned}$$

Again, the time derivatives of $\mathbf{A}$ and $\mathbf{B}$ translate to those of $a$ and $b$, respectively. We obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{A}^\top\mathbf{A}-\mathbf{B}\mathbf{B}^\top) &= \frac{\mathrm{d}}{\mathrm{d}t}(a^2-b^2)\mathbf{I}_r \\ &= -2\left(a\cdot\mathrm{sgn}(a')-b\cdot\mathrm{sgn}(b')\right)\mathbf{I}_r \\ &= -2\mathrm{sgn}(ab-\sigma)\left(a\cdot\mathrm{sgn}(b)-b\cdot\mathrm{sgn}(a)\right)\mathbf{I}_r \neq \mathbf{0}.\end{aligned}$$

□

### D.3 Saddle Point Avoidance

Bah et al. [4] characterize the critical points of the loss

$$L^N(\mathbf{W}_1,\ldots,\mathbf{W}_N) = \frac{1}{2}\|\mathbf{Y}-\mathbf{W}_N\cdots\mathbf{W}_1\mathbf{X}\|_{\mathrm{F}}^2, \tag{64}$$

where $\mathbf{W}_j \in \mathbb{R}^{d_j\times d_{j-1}}$ for $d_0 = d_x$, $d_N = d_y$ and $\mathbf{X} \in \mathbb{R}^{d_x\times m}$ and $\mathbf{Y} \in \mathbb{R}^{d_y\times m}$. They analyzed whether the critical points are the minima or strict saddle points through a closely related loss

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

where $\mathbf{W} \in \mathbb{R}^{d_y\times d_x}$. We summarize their result, especially for the case $N = 2$ and $\mathbf{X} = \mathbf{I}$.

**Proposition D.11** ([4], Propositions 6.6, 6.9, and 6.11)**.** *Let $r \triangleq \min\{d_0, d_1, d_2\}$ and $q = \mathrm{rank}(\mathbf{Y})$.*

(i) *The minimizers of $L^1$ over matrices of rank at most $k$ are precisely the best rank-$k$ approximations of $\mathbf{Y}$.*

(ii) *Let $K = \mathrm{rank}(\mathbf{W}_2\mathbf{W}_1)$. If $K = \min\{r, q\}$, every critical point $(\mathbf{W}_1, \mathbf{W}_2)$ of $L^2$ is a global minimizer if $\mathbf{W}_2\mathbf{W}_1$ is a best rank-$K$ approximation of $\mathbf{Y}$ and a strict saddle point if it is not.*

(iii) *If $K < \min\{r, q\}$, every critical point $(\mathbf{W}_1, \mathbf{W}_2)$ is a strict saddle point.*

We say that a critical point $x_c$ of a twice continuously differentiable function $f$ is a *strict saddle point* if the Hessian of $f$ at $x_c$ has a negative eigenvalue. Proposition D.11 implies that every critical point of $L^2$ is exactly either a global minimum or a strict saddle point. Notice that if $\mathbf{X}$ is the identity matrix and $N = 2$, then the setting of Bah et al. [4] exactly matches with ours. They also proved that the set of initial points such that the corresponding plain gradient flow for Equation (64) converges to a strict saddle point has measure zero [4, Theorem 6.3]. We also show that this holds for SpecGF with $\mathcal{T}_\beta$. Then, if SpecGF converges, it almost surely converges to the global minima, or the best low-rank approximation of $\mathbf{Y}$.

To this end, we employ a proposition from Cheridito et al. [11] that claims a similar result.

**Proposition D.12** ([11], Proposition 2.5)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$. Suppose that there exist open sets $V \subseteq U \subseteq \mathbb{R}^d$ such that the complement of $V$ has Lebesgue measure zero and $f$ is continuously differentiable on $U$ with a locally Lipschitz continuous Jacobian, which is non-degenerate on $V$. Let $\mathcal{S} \subseteq \{x \in U : f(x) = x\}$ and assume for all $x \in \mathcal{S}$ that $f'(x)$ is symmetric and has an eigenvalue of absolute value strictly greater than 1. Then, the set $\{x \in \mathbb{R}^d : \lim_{k\to\infty} f^k(x) \in \mathcal{S}\}$ has Lebesgue measure 0.*

There, $f^k(x)$ for $k \in \mathbb{N}_0$ is inductively defined as $f^0 \equiv \mathrm{id}$ and $f^{k+1} = f \circ f^k$.

**Theorem D.13.** *SpecGF with $\mathcal{T}_\beta$ almost surely avoids strict saddle points. Specifically, if $r \leq \mathrm{rank}(\mathbf{Y})$, then SpecGF with $\mathcal{T}_\beta$ almost surely converges to the best rank-$r$ approximation of $\mathbf{Y}$ if convergence is guaranteed.*

*Proof.* We first verify that the conditions of Proposition D.12 are met for SpecGF with $\mathcal{T}_\beta$ and then extend the statement from the discrete to the continuous version.

For simplicity, let $\mathbf{x}(t) \triangleq (\mathbf{A}(t), \mathbf{B}(t))$ and $f_\beta(\mathbf{x}) \triangleq (-\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{x})), -\mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{x})))$. Denote the solution of

$$\dot{\mathbf{x}}(t) = f_\beta(\mathbf{x}), \quad \mathbf{x}(0) \triangleq \mathbf{x}_0.$$

as $\phi(t)$; when the initial point matters, we use

$$\varphi_t(\mathbf{x}_0) = \phi(t).$$

We note that $\varphi_t(\mathbf{x}_0)$, and also $\phi(t)$, satisfy the semigroup property:

$$\varphi_{t'+t}(\mathbf{x}_0) = \varphi_{t'}(\varphi_t(\mathbf{x}_0)).$$

The Jacobian matrix of $\varphi_t(\mathbf{x}_0)$, $D\varphi_t(\mathbf{x}_0)$, satisfies the following with the Jacobian matrix of $f_\beta$, $Df_\beta$:

$$\frac{\mathrm{d}}{\mathrm{d}t} D\varphi_t(\mathbf{x}_0) = Df_\beta\left(\varphi_t(\mathbf{x}_0)\right) D\varphi_t(\mathbf{x}_0), \quad D\varphi_0(\mathbf{x}_0) = \mathbf{I}.$$

Liouville's theorem yields:

$$|D\varphi_t(\mathbf{x}_0)| = |D\varphi_0(\mathbf{x}_0)| \exp\left(\int_0^t \mathrm{tr} Df_\beta\left(\varphi_\tau(\mathbf{x}_0)\right) \mathrm{d}\tau\right).$$

Analyticity of $f_\beta$ makes $|D\varphi_t(\mathbf{x}_0)|$ nonzero for all $\mathbf{x}_0$ and all $t \geq 0$; hence, the conditions are met with $V = U = \mathbb{R}^d$. Now take $\mathcal{S}$ as the set of strict saddle points. It is straightforward to see that $\mathcal{T}_\beta$ is (Fréchet) differentiable at 0 with

$$D\mathcal{T}_\beta(0)[\mathbf{\Delta}] = \beta^{-1/2}\mathbf{\Delta}, \tag{65}$$

for every matrix $\mathbf{\Delta}$. It suffices to prove that

$$\lim_{h \to 0} \frac{\left\| \mathcal{T}_\beta(h\mathbf{\Delta}) - \mathcal{T}_\beta(0) - h\beta^{-\frac{1}{2}}\mathbf{\Delta} \right\|_{\mathrm{F}}}{|h|} = 0,$$

or

$$\lim_{h \to 0} \left\| \frac{\mathcal{T}_\beta(h\mathbf{\Delta})}{h} - \beta^{-\frac{1}{2}}\mathbf{\Delta} \right\|_{\mathrm{F}} = 0.$$

Notice that, for $s(x) = x^{-\frac{1}{2}}$ on $[\beta, \infty)$ and $c > 0$,

$$|s(x+c) - s(x)| = |s'(x+rc)|c \leq \frac{c}{2\beta^{\frac{3}{2}}} \tag{66}$$

for some $r \in (0,1)$. Applying Equation (66) to the eigenvalues of $\mathbf{\Delta}^\top\mathbf{\Delta}$ yields

$$\begin{aligned}
\left\| \frac{\mathcal{T}_\beta(h\mathbf{\Delta})}{h} - \beta^{-\frac{1}{2}}\mathbf{\Delta} \right\|_{\mathrm{F}} &= \left\| \mathbf{\Delta}\left(h^2\mathbf{\Delta}^\top\mathbf{\Delta} + \beta\mathbf{I}\right)^{-\frac{1}{2}} - \beta^{-\frac{1}{2}}\mathbf{\Delta} \right\|_{\mathrm{F}} \\
&\leq \|\mathbf{\Delta}\|_{\mathrm{F}} \|\left(h^2\mathbf{\Delta}^\top\mathbf{\Delta} + \beta\mathbf{I}\right)^{-\frac{1}{2}} - \beta^{-\frac{1}{2}}\mathbf{I}\|_2 \\
&\leq \|\mathbf{\Delta}\|_{\mathrm{F}} \frac{h^2\|\mathbf{\Delta}\|_2^2}{2\beta^{\frac{3}{2}}}.
\end{aligned}$$

Equation (65) directly yields that

$$Df_\beta(\bar{\mathbf{x}}) = -\beta^{-1/2}\nabla^2\mathcal{L}(\bar{\mathbf{x}}) \tag{67}$$

and it is symmetric for all $\bar{\mathbf{x}} \in \mathcal{S}$ (in fact, this holds at any critical point). As $\bar{\mathbf{x}}$ is an equilibrium point,

$$\frac{\mathrm{d}}{\mathrm{d}t} D\varphi_t(\bar{\mathbf{x}}) = Df_\beta\left(\varphi_t(\bar{\mathbf{x}})\right) D\varphi_t(\bar{\mathbf{x}}) = Df_\beta\left(\bar{\mathbf{x}}\right) D\varphi_t(\bar{\mathbf{x}}).$$

Thus

$$D\varphi_t(\bar{\mathbf{x}}) = e^{tDf_\beta(\bar{\mathbf{x}})}.$$

Since $\bar{\mathbf{x}}$ is a strict saddle, the Hessian of $\mathcal{L}$ at $\bar{\mathbf{x}}$ has a negative eigenvalue. By Equation (67), $Df_\beta(\bar{\mathbf{x}})$ has a positive eigenvalue, and hence $D\varphi_t(\bar{\mathbf{x}})$ has an eigenvalue whose (absolute) value is strictly greater than 1 at any $t > 0$.

The remaining part, the extension from discrete to continuous, is simple. Fix any time $T > 0$. By the semigroup property,

$$\begin{aligned}\left\{\mathbf{x}_0 : \lim_{t\to\infty} \varphi_t(\mathbf{x}_0) \in \mathcal{S}\right\} &= \bigcup_{\bar{\mathbf{x}}\in\mathcal{S}} \left\{\mathbf{x}_0 : \lim_{t\to\infty} \varphi_t(\mathbf{x}_0) = \bar{\mathbf{x}}\right\} \\ &\subseteq \bigcup_{\bar{\mathbf{x}}\in\mathcal{S}} \left\{\mathbf{x}_0 : \lim_{k\to\infty} \varphi_{kT}(\mathbf{x}_0) = \bar{\mathbf{x}}\right\} \\ &= \left\{\mathbf{x}_0 : \lim_{k\to\infty} (\varphi_T)^k(\mathbf{x}_0) \in \mathcal{S}\right\}.\end{aligned}$$

The proof is concluded since the subset of a Lebesgue measure zero set is again of Lebesgue measure zero. □

### D.4 Convergence Rate of SpecGF

Here, we study the convergence rate of SpecGF with $\mathcal{T}_\beta$, assuming the convergence itself. This analysis closely follows that of Proposition D.7; the SpecGF iterate will get close to some critical point of $\mathcal{L}$ after a finite time, where the Łojasiewicz gradient inequality holds. In the vicinity of that critical point, the rate of convergence is explicitly derived.

Let $(\mathbf{A}(t), \mathbf{B}(t)) \to (\bar{\mathbf{A}}, \bar{\mathbf{B}})$. Then $\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(t)), \mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}(t)) \to 0$; thus, $\nabla_{\mathbf{A}}\mathcal{L}(t), \nabla_{\mathbf{B}}\mathcal{L}(t) \to 0$. This implies that there exists some time $t' > 0$ such that for all $t \geq t'$,

$$\|\nabla_{\mathbf{A}}\mathcal{L}(t)\|_2 \leq \sqrt{\beta}, \quad \|\nabla_{\mathbf{B}}\mathcal{L}(t)\|_2 \leq \sqrt{\beta}.$$

In addition, let $W$ be the neighborhood of $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ where the Łojasiewicz gradient inequality holds: there also exist constants $C > 0$ and $b \in [\frac{1}{2}, 1)$ such that, for all $(\mathbf{A}, \mathbf{B}) \in W$,

$$|\mathcal{L}(\mathbf{A}, \mathbf{B}) - \mathcal{L}(\bar{\mathbf{A}}, \bar{\mathbf{B}})|^b \leq C\|\nabla\mathcal{L}(\mathbf{A}, \mathbf{B})\|.$$

We know that there exists some time $t'' > 0$ such that for all $t \geq t''$, $(\mathbf{A}(t), \mathbf{B}(t)) \in W$. Take $T \triangleq \max\{t', t''\} > 0$. We briefly state the helper lemma for deriving the convergence rate.

**Lemma D.14.** *For every matrix* $\mathbf{M}$ *such that* $\|\mathbf{M}\|_2 \leq \sqrt{\beta}$*,*

$$\langle \mathbf{M}, \mathcal{T}_\beta(\mathbf{M})\rangle \geq \frac{1}{\sqrt{2\beta}}\|\mathbf{M}\|_{\mathrm{F}}^2.$$

*Proof.*

$$\langle \mathbf{M}, \mathcal{T}_\beta(\mathbf{M})\rangle = \sum_i \frac{\sigma_i(\mathbf{M})^2}{\sqrt{\sigma_i(\mathbf{M})^2 + \beta}} \geq \sum_i \frac{\sigma_i(\mathbf{M})^2}{\sqrt{2\beta}} = \frac{1}{\sqrt{2\beta}}\|\mathbf{M}\|_{\mathrm{F}}^2.$$

□

Observe that, for $t \geq T$,

$$\begin{aligned}\frac{d}{\mathrm{d}t}\mathcal{L}(t) &= \langle \nabla_{\mathbf{A}}\mathcal{L}(t), -\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(t))\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}(t), -\mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}(t))\rangle \\ &\leq -\frac{1}{\sqrt{2\beta}}\left(\|\nabla_{\mathbf{A}}\mathcal{L}(t)\|_{\mathrm{F}}^2 + \|\nabla_{\mathbf{B}}\mathcal{L}(t)\|_{\mathrm{F}}^2\right) = -\frac{1}{\sqrt{2\beta}}\|\nabla\mathcal{L}(t)\|_{\mathrm{F}}^2 \\ &\leq -\frac{1}{C^2\sqrt{2\beta}}(\mathcal{L}(t) - \mathcal{L}(\bar{\mathbf{A}}, \bar{\mathbf{B}}))^{2b}.\end{aligned}$$

Let $\mathcal{E}(\mathbf{A}, \mathbf{B}) \triangleq \mathcal{L}(\mathbf{A}, \mathbf{B}) - \mathcal{L}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \geq 0$; we use simplified notation $\mathcal{E}(t) \triangleq \mathcal{L}(t) - \mathcal{L}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \geq 0$. Then for $t \geq T$, $\dot{\mathcal{E}}(t) \leq -C(\beta)\mathcal{E}(t)^{2b}$, where $C(\beta) = \frac{1}{C^2\sqrt{2\beta}}$. The convergence rate of $\mathcal{E}$ depends

on $b$. In our case, we established the almost sure convergence towards global minima (assuming convergence); hence, assume $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is the global minimum. Then, $\mathcal{E}$ is a Morse-Bott function, and thus the (optimal) exponent $b$ is exactly $\frac{1}{2}$ [18]. Then,

$$\dot{\mathcal{E}}(t) \leq -C(\beta)\mathcal{E}(t),$$

or

$$\mathcal{E}(t) \leq \mathcal{E}(T)e^{-C(\beta)(t-T)}.$$

Therefore, $\mathcal{E}(t)$ exponentially decreases to 0. It remains to show that $\mathcal{E}(\mathbf{A}, \mathbf{B})$ is a Morse-Bott function.

**Assumption D.15.** The nonzero singular values of $\mathbf{Y}$ are mutually distinct.

Recall that $\bar{\mathbf{A}}\bar{\mathbf{B}}$ is the best rank-$r$ approximation of $\mathbf{Y}$ [4]. We denote $\bar{\mathbf{Y}}_r \triangleq \bar{\mathbf{A}}\bar{\mathbf{B}}$ for simplicity. Furthermore, denote

$$\bar{\mathbf{Y}}_r = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top, \quad \mathbf{Y} - \bar{\mathbf{Y}}_r \triangleq \mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{V}_\perp^\top.$$

$[\mathbf{U}_r|\mathbf{U}_\perp]$ forms an $m \times m$ orthogonal matrix; similarly, $[\mathbf{V}_r|\mathbf{V}_\perp]$ forms an $n \times n$ orthogonal matrix. Note that zero singular values may appear in $\mathbf{\Sigma}_\perp$, but never appear in $\mathbf{\Sigma}_r$.

The solution set

$$S \triangleq \left\{\mathbf{M} \in \mathbb{R}^{m\times r}, \mathbf{N} \in \mathbb{R}^{r\times n} : \mathbf{MN} = \bar{\mathbf{Y}}_r\right\}$$

is easily identified with $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$:

$$S = \left\{(\bar{\mathbf{A}}\mathbf{Q}, \mathbf{Q}^{-1}\bar{\mathbf{B}}) : \mathbf{Q} \in \mathsf{GL}(r)\right\}.$$

Indeed, for $(\mathbf{M}, \mathbf{N}) \in S$, if the left inverse of $\bar{\mathbf{A}}$ is $\mathbf{L}_{\bar{\mathbf{A}}}$, then $\mathbf{Q} = \mathbf{L}_{\bar{\mathbf{A}}}\mathbf{M}$ gives $\bar{\mathbf{A}}\mathbf{Q} = \bar{\mathbf{A}}\mathbf{L}_{\bar{\mathbf{A}}}\mathbf{M} = \mathbf{M}$ and $\mathbf{QN} = \mathbf{L}_{\bar{\mathbf{A}}}\bar{\mathbf{Y}}_r = \bar{\mathbf{B}}$; definitely $\mathrm{rank}(\mathbf{Q}) = r$ from $\bar{\mathbf{A}}\mathbf{Q} = \mathbf{M}$.

The mapping

$$\Phi : \mathsf{GL}(r) \to \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}, \quad \Phi(\mathbf{Q}) = (\bar{\mathbf{A}}\mathbf{Q}, \mathbf{Q}^{-1}\bar{\mathbf{B}})$$

is real-analytic and satisfies

$$\Phi^{-1}(\mathbf{M}, \mathbf{N}) = \mathbf{L}_{\bar{\mathbf{A}}}\mathbf{M}$$

for a left inverse $\mathbf{L}_{\bar{\mathbf{A}}}$ of $\bar{\mathbf{A}}$. Thus $\Phi$ is a real-analytic diffeomorphism from an open subset (hence an analytic submanifold) $\mathsf{GL}(r)$ to $\Phi(\mathsf{GL}(r)) = S$, and $S$ is an analytic submanifold. As $\mathsf{GL}(r)$ has two connected components—matrices with positive and negative determinants—so does $S$. This issue is resolved by shrinking the domain of $S$ to an open neighborhood of $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$; then $S$ is connected. It remains to show that $T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S = \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ [18, Definition 1.5]. Other conditions for being a Morse-Bott function are automatically satisfied [18, Remark 1.6]; $\nabla^2\mathcal{E} : (\mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}) \to (\mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n})$ is a linear map and thus a Fredholm operator of index 0; hence, the range condition is automatically satisfied.

Let $\mathbf{R}(\mathbf{A}, \mathbf{B}) \triangleq \mathbf{AB} - \bar{\mathbf{Y}}_r$ be the residual matrix. The differential of $\mathbf{R}$ at $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is

$$D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}})[\Delta\mathbf{A}, \Delta\mathbf{B}] = \bar{\mathbf{A}}\Delta\mathbf{B} + \Delta\mathbf{A}\bar{\mathbf{B}}.$$

It is straightforward to see that

$$\ker D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \left\{(\bar{\mathbf{A}}\mathbf{K}, -\mathbf{K}\bar{\mathbf{B}}) : \mathbf{K} \in \mathbb{R}^{r\times r}\right\}.$$

Indeed, if $\bar{\mathbf{A}}\Delta\mathbf{B} + \Delta\mathbf{A}\bar{\mathbf{B}} = \mathbf{0}$, the right inverse of $\bar{\mathbf{B}}$, $\mathbf{R}_{\bar{\mathbf{B}}}$, gives $\bar{\mathbf{A}}\Delta\mathbf{B}\mathbf{R}_{\bar{\mathbf{B}}} + \Delta\mathbf{A} = \mathbf{0}$, which implies $\Delta\mathbf{A} = -\bar{\mathbf{A}}\Delta\mathbf{B}\mathbf{R}_{\bar{\mathbf{B}}}$ and $\bar{\mathbf{A}}(\Delta\mathbf{B} - \Delta\mathbf{B}\mathbf{R}_{\bar{\mathbf{B}}}\bar{\mathbf{B}}) = \mathbf{0} \Rightarrow \Delta\mathbf{B} = \Delta\mathbf{B}\mathbf{R}_{\bar{\mathbf{B}}}\bar{\mathbf{B}}$.

$T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S \subset \ker D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is trivial; we show the opposite inclusion. For $(\Delta\mathbf{A}, \Delta\mathbf{B}) \in \ker D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$, we have $(\Delta\mathbf{A}, \Delta\mathbf{B}) = (\bar{\mathbf{A}}\mathbf{K}, -\mathbf{K}\bar{\mathbf{B}})$ for some $\mathbf{K}$. Define a smooth curve

$$\gamma(t) \triangleq (\bar{\mathbf{A}}(\mathbf{I}_r + t\mathbf{K}), (\mathbf{I}_r + t\mathbf{K})^{-1}\bar{\mathbf{B}}).$$

For all $t$ sufficiently close to 0, $\mathbf{I}_r + t\mathbf{K}$ is invertible and $\gamma(t) \in S$. Moreover,

$$\gamma'(0) = (\bar{\mathbf{A}}\mathbf{K}, -\mathbf{K}\bar{\mathbf{B}}).$$

Hence, every $(\Delta\mathbf{A}, \Delta\mathbf{B}) \in \ker D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is realized as the velocity of a curve in $S$, or $(\Delta\mathbf{A}, \Delta\mathbf{B}) \in T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S$. We obtain

$$T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S = \ker D\mathbf{R}(\bar{\mathbf{A}}, \bar{\mathbf{B}}).$$

The Hessian of $\mathcal{E}$ at $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is computed:

$$\begin{aligned}&\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})[(\Delta\mathbf{A}_0, \Delta\mathbf{B}_0), (\Delta\mathbf{A}_1, \Delta\mathbf{B}_1)]\\ &= \left\langle \Delta\mathbf{A}_0\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_0, \Delta\mathbf{A}_1\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_1 \right\rangle + \left\langle \bar{\mathbf{A}}\bar{\mathbf{B}} - \mathbf{Y}, \Delta\mathbf{A}_0\Delta\mathbf{B}_1 + \Delta\mathbf{A}_1\Delta\mathbf{B}_0 \right\rangle.\end{aligned}$$

Without loss of generality, take $\bar{\mathbf{A}} = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2}$ and $\bar{\mathbf{B}} = \mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^\top$. Note that any other $(\hat{\mathbf{A}}, \hat{\mathbf{B}}) \in S$ must satisfy

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = (\bar{\mathbf{A}}\mathbf{Q}, \mathbf{Q}^{-1}\bar{\mathbf{B}}) \triangleq \phi_{\mathbf{Q}}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$$

for $\mathbf{Q} \in \mathsf{GL}(r)$. From $\bar{\mathbf{A}}\Delta\mathbf{B} + \Delta\mathbf{A}\bar{\mathbf{B}} = \mathbf{0}$, we can check that

$$(\Delta\mathbf{A}, \Delta\mathbf{B}) \in T_{(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S \quad \Leftrightarrow \quad \phi_{\mathbf{Q}}(\Delta\mathbf{A}, \Delta\mathbf{B}) \in T_{\phi_{\mathbf{Q}}(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S.$$

This implies

$$\begin{aligned}&\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})[(\Delta\mathbf{A}_0, \Delta\mathbf{B}_0), (\Delta\mathbf{A}_1, \Delta\mathbf{B}_1)]\\ =&\nabla^2\mathcal{E}(\phi_{\mathbf{Q}}(\bar{\mathbf{A}}, \bar{\mathbf{B}}))[\phi_{\mathbf{Q}}(\Delta\mathbf{A}_0, \Delta\mathbf{B}_0), \phi_{\mathbf{Q}}(\Delta\mathbf{A}_1, \Delta\mathbf{B}_1)],\end{aligned}$$

and thus

$$(\Delta\mathbf{A}, \Delta\mathbf{B}) \in \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \quad \Leftrightarrow \quad \phi_{\mathbf{Q}}(\Delta\mathbf{A}, \Delta\mathbf{B}) \in \ker\nabla^2\mathcal{E}(\hat{\mathbf{A}}, \hat{\mathbf{B}}).$$

Hence, it suffices to show $\ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = T_{(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S$ at a single point $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \in S$ to show that $\ker\nabla^2\mathcal{E}(\mathbf{M}, \mathbf{N}) = T_{(\mathbf{M}, \mathbf{N})}S$ for any $(\mathbf{M}, \mathbf{N}) \in S$. Moreover, as the critical set $S$ is an analytic manifold of the analytic $\mathcal{L}$, we have $T_{(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S \subset \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ [18, Remark 2.9]; hence, what remains is to show $\ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \subset T_{(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S$.

Now return to the specific point $\bar{\mathbf{A}} = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2}$ and $\bar{\mathbf{B}} = \mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^\top$. If

$$(\Delta\mathbf{A}_0, \Delta\mathbf{B}_0) \in \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}}),$$

then for all $(\Delta\mathbf{A}_1, \Delta\mathbf{B}_1)$, $\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})[(\Delta\mathbf{A}_0, \Delta\mathbf{B}_0), (\Delta\mathbf{A}_1, \Delta\mathbf{B}_1)] = \mathbf{0}$. Specifically,

$$\Delta\mathbf{A}_1 = \mathbf{0} \Rightarrow \bar{\mathbf{A}}^\top(\Delta\mathbf{A}_0\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_0) + \Delta\mathbf{A}_0^\top(\bar{\mathbf{A}}\bar{\mathbf{B}} - \mathbf{Y}) = \mathbf{0}, \tag{68}$$

$$\Delta\mathbf{B}_1 = \mathbf{0} \Rightarrow (\Delta\mathbf{A}_0\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_0)\bar{\mathbf{B}}^\top + (\bar{\mathbf{A}}\bar{\mathbf{B}} - \mathbf{Y})\Delta\mathbf{B}_0^\top = \mathbf{0}. \tag{69}$$

There exist $\mathbf{C}_r, \mathbf{D}_r \in \mathbb{R}^{r\times r}$, $\mathbf{C}_\perp \in \mathbb{R}^{(m-r)\times r}$, and $\mathbf{D}_\perp \in \mathbb{R}^{r\times(n-r)}$ such that

$$\Delta\mathbf{A}_0 = \mathbf{U}_r\mathbf{C}_r + \mathbf{U}_\perp\mathbf{C}_\perp, \quad \Delta\mathbf{B}_0 = \mathbf{D}_r\mathbf{V}_r^\top + \mathbf{D}_\perp\mathbf{V}_\perp^\top.$$

Plugging into Equations (68) and (69), we obtain

$$\begin{aligned}\mathbf{\Sigma}_r^{1/2}(\mathbf{C}_r\mathbf{\Sigma}_r^{1/2} + \mathbf{\Sigma}_r^{1/2}\mathbf{D}_r)\mathbf{V}_r^\top + \mathbf{\Sigma}_r\mathbf{D}_\perp\mathbf{V}_\perp^\top - \mathbf{C}_\perp^\top\mathbf{\Sigma}_\perp\mathbf{V}_\perp^\top = \mathbf{0},\\ \mathbf{U}_r(\mathbf{C}_r\mathbf{\Sigma}_r^{1/2} + \mathbf{\Sigma}_r^{1/2}\mathbf{D}_r)\mathbf{\Sigma}_r^{1/2} + \mathbf{U}_\perp\mathbf{C}_\perp\mathbf{\Sigma}_r - \mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{D}_\perp^\top = \mathbf{0},\end{aligned}$$

which reduces to

$$\mathbf{C}_r\mathbf{\Sigma}_r^{1/2} + \mathbf{\Sigma}_r^{1/2}\mathbf{D}_r = \mathbf{0}, \tag{70}$$

$$\mathbf{\Sigma}_r\mathbf{D}_\perp - \mathbf{C}_\perp^\top\mathbf{\Sigma}_\perp = \mathbf{0}, \tag{71}$$

$$\mathbf{C}_\perp\mathbf{\Sigma}_r - \mathbf{\Sigma}_\perp\mathbf{D}_\perp^\top = \mathbf{0}. \tag{72}$$

Since no singular value of $\mathbf{\Sigma}_r$ equals any singular value of $\mathbf{\Sigma}_\perp$, Equations (71) and (72) imply that $\mathbf{C}_\perp\mathbf{\Sigma}_r^2 = \mathbf{\Sigma}_\perp\mathbf{\Sigma}_\perp^\top\mathbf{C}_\perp$, and thus $\mathbf{C}_\perp = \mathbf{0}$; $\mathbf{D}_\perp = \mathbf{0}$ also holds. Moreover, take $\mathbf{K} = \mathbf{\Sigma}_r^{-1/2}\mathbf{C}_r$; we see that

$$\begin{aligned}&\bar{\mathbf{A}}\mathbf{K} = \mathbf{U}_r\mathbf{C}_r = \Delta\mathbf{A}_0,\\ &\mathbf{K}\bar{\mathbf{B}} = \mathbf{\Sigma}_r^{-1/2}\mathbf{C}_r\mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^\top = \mathbf{\Sigma}_r^{-1/2}(-\mathbf{\Sigma}_r^{1/2}\mathbf{D}_r)\mathbf{V}_r^\top = -\Delta\mathbf{B}_0.\end{aligned}$$

Therefore, $\ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is a subset of $T_{(\bar{\mathbf{A}}, \bar{\mathbf{B}})}S$. This argument analogously holds at any zero of $\mathbf{R}$, or any $(\mathbf{M}, \mathbf{N}) \in S$, so $\mathcal{E}$ is Morse-Bott.

**Finite-time convergence of $\mathcal{T}$.** We further show that if SpecGF with $\mathcal{T}$ is well-defined, has an absolutely continuous trajectory, and converges, then the convergence should happen at a finite time.

**Proposition D.16.** *Assume that SpecGF with $\mathcal{T}$ admits a locally absolutely continuous solution on $t \geq 0$ satisfying Equation* (3) *almost everywhere. If the solution converges, then it converges to a critical point in finite time.*

*Proof.* Let

$$\Delta(t) \triangleq \mathcal{L}(\mathbf{A}(t), \mathbf{B}(t)) - \mathcal{L}(\mathbf{A}(\infty), \mathbf{B}(\infty)) \triangleq \mathcal{L}(\mathbf{A}(t), \mathbf{B}(t)) - \mathcal{L}_\infty \geq 0.$$

For each (open) neighborhood $U$ of $(\mathbf{A}(\infty), \mathbf{B}(\infty))$, there exists a time $t_U \geq 0$ such that for all $t \geq t_U$, $(\mathbf{A}(t), \mathbf{B}(t)) \in U$. Apply Lemma D.6: there exist a neighborhood $W$ of $(\mathbf{A}(\infty), \mathbf{B}(\infty))$ and constants $C > 0$ and $b \in [\frac{1}{2}, 1)$ such that for all $x \in W$,

$$|\mathcal{L}(\mathbf{A}(t), \mathbf{B}(t)) - \mathcal{L}_\infty|^b \leq C \|\nabla \mathcal{L}(\mathbf{A}(t), \mathbf{B}(t))\|.$$

Observe that

$$\begin{aligned}
\Delta'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mathbf{A}(t), \mathbf{B}(t)) &= -\left(\|\nabla_{\mathbf{A}}\mathcal{L}\|_* + \|\nabla_{\mathbf{B}}\mathcal{L}\|_*\right) \leq -\left(\|\nabla_{\mathbf{A}}\mathcal{L}\|_{\mathrm{F}} + \|\nabla_{\mathbf{B}}\mathcal{L}\|_{\mathrm{F}}\right) \\
&\leq -\|\nabla \mathcal{L}(\mathbf{A}(t), \mathbf{B}(t))\|_{\mathrm{F}} \\
&\leq -\frac{1}{C}\left(\mathcal{L}(\mathbf{A}(t), \mathbf{B}(t)) - \mathcal{L}_\infty\right)^b = -\frac{1}{C}\Delta(t)^b.
\end{aligned}$$

At any time $t \geq t_W$, if $\Delta(t) = 0$, done. Otherwise,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\Delta(t)^{1-b} = (1-b)\Delta(t)^{-b}\Delta'(t) \leq -\frac{1-b}{C} \\
\Rightarrow &\Delta(t)^{1-b} - \Delta(t_W)^{1-b} \leq -\frac{1-b}{C}(t - t_W).
\end{aligned}$$

Taking $T = t_W + \frac{C}{1-b}\Delta(t_W)^{1-b}$ gives $\Delta(T) \leq 0$; hence, $\Delta(T) = 0$. □

### D.5 Stability Analysis of SpecGF

We first introduce the definition of Lyapunov stability.

**Definition D.17** (Lyapunov stability)**.** A critical point $\bar{\mathbf{x}} \in \mathbf{R}^d$ is *Lyapunov stable* if for every neighborhood $U_\epsilon \subset \mathbf{R}^d$ of $\bar{\mathbf{x}}$, there exists a neighborhood $U_\delta \subset \mathbf{R}^d$ of $\bar{\mathbf{x}}$ such that for all $\mathbf{x}(t)$ with $\mathbf{x}(0) \in U_\delta$, $\mathbf{x}(t) \in U_\epsilon$ for all $t \geq 0$.

Recall that, from Section D.1,

$$\mathcal{T}_\beta(\mathbf{X}) = \mathbf{U}\mathbf{S}_\beta\mathbf{V}^\top, \quad \mathbf{S}_\beta \triangleq \mathrm{Diag}\left\{\frac{\sigma_i}{\sqrt{\sigma_i^2 + \beta}}\right\}_{i \in [r]}.$$

Thus,

$$\langle \mathbf{X}, \mathcal{T}_\beta(\mathbf{X})\rangle \geq \sqrt{\beta}\|\mathcal{T}_\beta(\mathbf{X})\|_{\mathrm{F}}^2 \quad \Rightarrow \quad \|\mathcal{T}_\beta(\mathbf{X})\|_{\mathrm{F}} \leq \beta^{-\frac{1}{2}}\|\mathbf{X}\|_{\mathrm{F}}. \tag{73}$$

**Proposition D.18** ([7], Theorem 5.2)**.** *Consider the system of differential equations*

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)) \tag{74}$$

*with $f : \mathbb{R}^d \to \mathbb{R}^d$ that yields a unique $\mathcal{C}^1$ solution for every initial condition $\mathbf{x}(0)$. Let $\mathbf{x} \in \mathbb{R}^d$, and suppose there exists a continuous function $V : \mathcal{V} \to \mathbb{R}$ defined on an open neighborhood $\mathcal{V}$ of $\mathbf{x}$, and $c > 0$ such that*

$$c\dot{V}(\mathbf{z}) + \|f(\mathbf{z})\| \leq 0 \tag{75}$$

*is satisfied for every $\mathbf{z} \in \mathcal{V}$. If $\mathbf{x}$ is a local minimizer of V, then $\mathbf{x}$ is a Lyapunov stable equilibrium.*

We now prove that every (global) minimum of the loss is Lyapunov stable.

**Proposition D.19** (Lyapunov stability of minima)**.** *Every global minimum of $\mathcal{L}(\mathbf{A}, \mathbf{B})$ is Lyapunov stable for SpecGF with $\mathcal{T}_\beta$.*

*Proof.* Recall Lemma D.6: at each global minimum $\mathbf{x}_*$ of $\mathcal{L}$, there exist a neighborhood $W$ of $\mathbf{x}_*$ and constants $C > 0$ and $b \in [\frac{1}{2}, 1)$ such that for all $\mathbf{x} \in W$,

$$|\mathcal{L}(\mathbf{x}) - \mathcal{L}(\mathbf{x}_*)|^b \leq C\|\nabla\mathcal{L}(\mathbf{x})\|.$$

Note that $\mathcal{L}(\mathbf{x}_*) \triangleq \mathcal{L}_*$ for any global minimum $\mathbf{x}_*$. Without loss of generality, take $W$ as an open ball centered at $\mathbf{x}_*$ and of (finite) positive radius. On $W$, define

$$V(\mathbf{x}) \triangleq (\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)^{1-b}. \tag{76}$$

Then $V$ is continuous, nonnegative, minimized at $\mathbf{x}_*$ and

$$\dot{V}(\mathbf{x}) = (1-b)\,(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)^{-b}\,\dot{\mathcal{L}}(\mathbf{x}).$$

On $W$, $\nabla_{\mathbf{A}}\mathcal{L}$ and $\nabla_{\mathbf{B}}\mathcal{L}$ are both bounded; take the bound as $C_W$ so that

$$\|\nabla_{\mathbf{A}}\mathcal{L}\|_{\mathrm{F}} \leq C_W, \quad \|\nabla_{\mathbf{B}}\mathcal{L}\|_{\mathrm{F}} \leq C_W.$$

Observe

$$\begin{aligned}
-\dot{\mathcal{L}} &= \langle \nabla_{\mathbf{A}}\mathcal{L}, \mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L})\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}, \mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L})\rangle \\
&= \sum_i \frac{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L})}{\sqrt{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L}) + \beta}} + \sum_j \frac{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L})}{\sqrt{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L}) + \beta}} \\
&\geq \sum_i \frac{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L})}{\sqrt{C_W^2 + \beta}} + \sum_j \frac{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L})}{\sqrt{C_W^2 + \beta}} \\
&= \frac{1}{\sqrt{C_W^2 + \beta}}\|\nabla\mathcal{L}\|_{\mathrm{F}}^2.
\end{aligned}$$

In addition, for the same $f_\beta$ in the proof of Theorem D.13,

$$\begin{aligned}
\|f_\beta(\mathbf{x})\|_{\mathrm{F}} &= \sqrt{\|\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{x}))\|_{\mathrm{F}}^2 + \|\mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{x}))\|_{\mathrm{F}}^2} \\
&\leq \sqrt{\beta^{-1}\left(\|\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2 + \|\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2\right)} = \beta^{-\frac{1}{2}}\|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}.
\end{aligned}$$

Hence, on $W$,

$$\begin{aligned}
-\dot{V}(\mathbf{x}) &= -(1-b)\,(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)^{-b}\,\dot{\mathcal{L}}(\mathbf{x}) \\
&\geq (1-b)\,(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)^{-b}\,\frac{1}{\sqrt{C_W^2 + \beta}}\|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2 \\
&\geq \frac{1-b}{C\sqrt{C_W^2 + \beta}}\|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}} \\
&\geq \frac{1-b}{C\sqrt{C_W^2 + \beta}}\beta^{\frac{1}{2}}\|f_\beta(\mathbf{x})\|_{\mathrm{F}}.
\end{aligned}$$

$c = \frac{C\sqrt{C_W^2+\beta}}{(1-b)\sqrt{\beta}}$ gives the desired inequality. □

Therefore, the a.s. local convergence of SpecGF with $\mathcal{T}_\beta$ towards global minima is guaranteed.

**Conjecture: The basin of attraction is large.** Assume $r = r^\star$. For now, let $\rho_W$ be the radius of $W$. Fix $\epsilon > 0$ and take $\rho = \min\{\rho_W, \epsilon\} > 0$. We will find $\delta > 0$ for the Lyapunov stability of each minimizer $\mathbf{x}_* = (\mathbf{A}_*, \mathbf{B}_*)$ with $\mathbf{A}_*\mathbf{B}_* = \mathbf{Y}$.

Let $\mathbf{z} = (\mathbf{A}, \mathbf{B})$ and assume $\|\mathbf{z} - \mathbf{x}_*\|_{\mathrm{F}} \leq \delta$. We have

$$\|\Delta\mathbf{A}\|_{\mathrm{F}} \triangleq \|\mathbf{A} - \mathbf{A}_*\|_{\mathrm{F}} \leq \delta, \quad \|\Delta\mathbf{B}\|_{\mathrm{F}} \triangleq \|\mathbf{B} - \mathbf{B}_*\|_{\mathrm{F}} \leq \delta.$$

Then, for some constant $c' > 0$ (which may depend on $\mathbf{x}_*$)

$$\begin{aligned}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}} &= \|\Delta\mathbf{A}\mathbf{B}_* + \mathbf{A}_*\Delta\mathbf{B} + \Delta\mathbf{A}\Delta\mathbf{B}\|_{\mathrm{F}} \\ &\le \|\Delta\mathbf{A}\mathbf{B}_*\|_{\mathrm{F}} + \|\mathbf{A}_*\Delta\mathbf{B}\|_{\mathrm{F}} + \|\Delta\mathbf{A}\Delta\mathbf{B}\|_{\mathrm{F}} \\ &\le \delta\left(\|\mathbf{A}_*\|_2 + \|\mathbf{B}_*\|_2\right) + \delta^2 \\ &\le c' \max\{\delta, \delta^2\}.\end{aligned}$$

(e.g., take $c' = \|\mathbf{A}_*\|_2 + \|\mathbf{B}_*\|_2 + 1$.) Hence,

$$\mathcal{L}(\mathbf{z}) = \frac{1}{2}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}^2 \le \frac{1}{2}c'^2 \max\{\delta^2, \delta^4\} \quad \Rightarrow \quad V(\mathbf{z}) \le \left(\frac{1}{2}c'^2 \max\{\delta^2, \delta^4\}\right)^{1-b}.$$

In the proof of Bhat and Bernstein [7, Theorem 5.2], two conditions are required:

$$\|\mathbf{z} - \mathbf{x}_*\|_{\mathrm{F}} \le \frac{\rho}{4}, \quad |V(\mathbf{z}) - V(\mathbf{x}_*)| \le \frac{\rho}{4c},$$

where $c = \frac{C\sqrt{C_W^2 + \beta}}{(1-b)\sqrt{\beta}}$. It suffices that

$$\delta \le \frac{\rho}{4}, \quad \left(\frac{1}{2}c'^2 \max\{\delta^2, \delta^4\}\right)^{1-b} \le \frac{\rho}{4c},$$

which translates to

$$\delta(\epsilon, \mathbf{x}_*) = \min\left\{\frac{\rho}{4}, \sqrt{\frac{2}{c'^2}\left(\frac{\rho}{4c}\right)^{\frac{1}{1-b}}}, \sqrt[4]{\frac{2}{c'^2}\left(\frac{\rho}{4c}\right)^{\frac{1}{1-b}}}\right\}, \quad \rho = \min\{\rho_W, \epsilon\}. \tag{77}$$

We conjecture that $\rho_W$ is sufficiently large for each and every minimizer $\mathbf{x}_*$. If so, $\delta$ increases as $\epsilon$ increases, and there is a possibility that initial points near (but not exactly) the origin fall in some basin of attraction.

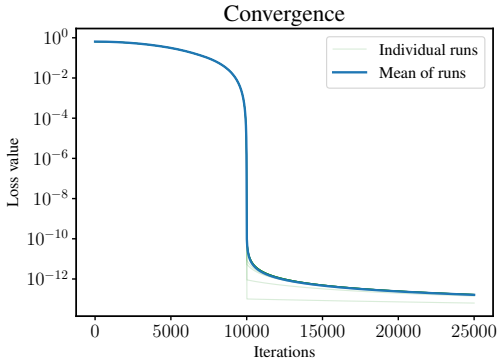



Figure 5: Loss trajectories started at $(\mathbf{A}_i(0), \mathbf{B}_i(0))$.

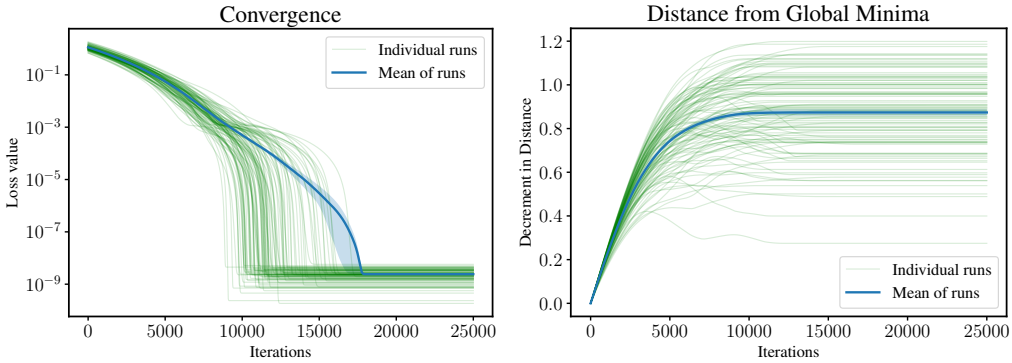



Figure 6: Lyapunov stability of global minima. **Left**: loss trajectories. **Right**: reduction in Euclidean distance to the corresponding reference minimizer.

We set $\mathbf{Y} \in \mathbb{R}^{9\times 9}$ with $r = 4$ and $\mathbf{\Sigma} = \mathrm{Diag}(1, 0.5, 0.2, 0.05)$, $\mathbf{A} \in \mathbb{R}^{9\times 4}$, and $\mathbf{B} \in \mathbb{R}^{4\times 9}$. We first sample 20 random initial points $\mathbf{B}_i(0)$ with i.i.d. Gaussian entries of scale $\gamma = 5 \times 10^{-4}$; $\mathbf{A}_i(0) = \mathbf{0}$. We use a fixed learning rate of $\eta = 10^{-4}$. We ran SpecGD with $\mathcal{T}_\beta$ starting from $(\mathbf{A}_i(0), \mathbf{B}_i(0))$ until convergence. For all $i \in [20]$, SpecGD converged to global minima; see Figure 5. Name those global minima as $(\mathbf{A}_i^*, \mathbf{B}_i^*)$. For each $i$, we sample 50 random directions $(\Delta\mathbf{A}_{ij}, \Delta\mathbf{B}_{ij})$ from i.i.d. Gaussian entries, with magnitude equal to the Euclidean distance between $(\mathbf{A}_i(0), \mathbf{B}_i(0))$ and $(\mathbf{A}_i^*, \mathbf{B}_i^*)$. Denote the initial points as $(\mathbf{A}_{ij}(0), \mathbf{B}_{ij}(0)) = (\mathbf{A}_i^*, \mathbf{B}_i^*) + (\Delta\mathbf{A}_{ij}, \Delta\mathbf{B}_{ij})$. We ran $20 \times 50 = 1000$ SpecGD experiments and recorded (i) the loss trajectory $\mathcal{L}(t)$ and (ii) the decrement of the distance from global minima, $\mathrm{dist}((\mathbf{A}_{ij}(0), \mathbf{B}_{ij}(0)), (\mathbf{A}_i^*, \mathbf{B}_i^*)) - \mathrm{dist}((\mathbf{A}_{ij}(t), \mathbf{B}_{ij}(t)), (\mathbf{A}_i^*, \mathbf{B}_i^*))$; $\mathrm{dist}(\mathbf{x}, \mathbf{y})$ stands for the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. The recorded values are plotted in Figure 6. These experiments support convergence to the set of global minima from a broad range of initializations.

### D.6 SpecGF with $\ell_2$ Regularization

With such regularization, the global convergence of SpecGF immediately follows. The setup of regularized SpecGF becomes

- The target matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ yields a compact SVD $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \sigma_2, \cdots, \sigma_{r^\star})$ with $\sigma_1 \geq \cdots \geq \sigma_{r^\star} > 0$.
- The loss function $\mathcal{L}_\lambda : \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n} \to \mathbb{R}$ for a fixed $\lambda > 0$ is defined as
$$\mathcal{L}_\lambda(\mathbf{A}, \mathbf{B}) \triangleq \frac{1}{2}\|\mathbf{A}\mathbf{B} - \mathbf{Y}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{B}\|_{\mathrm{F}}^2. \tag{78}$$
We may abuse the notation and simply state $\mathcal{L}_\lambda(t)$. $\mathcal{L}_\lambda$ is also analytic in $\mathbf{A}$ and $\mathbf{B}$.
- The time derivatives of $\mathbf{A}$ and $\mathbf{B}$ are governed by either $\mathcal{T}$ or $\mathcal{T}_\beta$ applied to the gradient:
$$\dot{\mathbf{A}}(t) = -\mathcal{T}\left(\nabla_{\mathbf{A}}\mathcal{L}_\lambda(t)\right), \quad \dot{\mathbf{B}}(t) = -\mathcal{T}\left(\nabla_{\mathbf{B}}\mathcal{L}_\lambda(t)\right),$$
or
$$\dot{\mathbf{A}}(t) = -\mathcal{T}_\beta\left(\nabla_{\mathbf{A}}\mathcal{L}_\lambda(t)\right), \quad \dot{\mathbf{B}}(t) = -\mathcal{T}_\beta\left(\nabla_{\mathbf{B}}\mathcal{L}_\lambda(t)\right).$$

The following statements are true, because the corresponding statements do not utilize the structure of the loss; only the analyticity of $\mathcal{T}_\beta$ and monotonicity of the loss matter, if any of those two is required.

1. Lemmas D.4 and D.5 readily follow for the regularized SpecGF with $\mathcal{T}_\beta$. Moreover,
$$\|\mathbf{A}(t)\|_{\mathrm{F}}^2 + \|\mathbf{B}(t)\|_{\mathrm{F}}^2 \leq \frac{2}{\lambda}\mathcal{L}_\lambda(t) \leq \frac{2}{\lambda}\mathcal{L}_\lambda(0).$$
Thus, $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are both bounded for all $t \geq 0$.
2. Equation (62) in Proposition D.7 also holds for $\mathcal{L}_\lambda$ (with the same constant), implying that regularized SpecGF with $\mathcal{T}_\beta$ should either converge or diverge to infinity. However, the latter cannot happen; hence, regularized SpecGF with $\mathcal{T}_\beta$ always converges.
3. Proposition D.12 is applicable for the regularized SpecGF with $\mathcal{T}_\beta$. That is, regularized SpecGF with $\mathcal{T}_\beta$ avoids strict saddle points almost surely as SpecGF with $\mathcal{T}_\beta$ does.
4. Proposition D.19 readily follows for the regularized SpecGF with $\mathcal{T}_\beta$. That is, every global minimum of regularized SpecGF with $\mathcal{T}_\beta$ is Lyapunov stable.

It remains to characterize the critical points of regularized SpecGF with $\mathcal{T}_\beta$, especially when the LoRA rank $r$ is smaller than the target rank $r^\star$. Although it is specified in Udell et al. [56], we reproduce their result. There, $\mathbf{M}_+ \triangleq \max\{\mathbf{M}, 0\}$ where $\max$ is applied coordinate-wise.

**Proposition D.20** ($\ell_2$-Regularized MF, [56])**.** *Set $r \leq r^\star$. Assume that all singular values of $\mathbf{Y}$ are distinct.*

*(i) Let $\Omega \subset [r^\star]$ with $|\Omega| \leq r$. Every stationary point of Equation* (78) *is of the form*
$$\mathbf{A} = \begin{bmatrix}\mathbf{U}_\Omega \left(\mathbf{\Sigma}_\Omega - \lambda\mathbf{I}_{|\Omega|}\right)_+^{\frac{1}{2}} & \mathbf{0}_{m\times(r-|\Omega|)}\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix}\left(\mathbf{\Sigma}_\Omega - \lambda\mathbf{I}_{|\Omega|}\right)_+^{\frac{1}{2}}\mathbf{V}_\Omega^\top \\ \mathbf{0}_{(r-|\Omega|)\times n}\end{bmatrix}.$$
*up to orthogonal transformation by $\mathbf{R} \in \mathsf{O}(r)$. We denote the submatrix of $\mathbf{U}$ with columns indexed by $\Omega$ as $\mathbf{U}_\Omega$, and similarly for $\mathbf{\Sigma}$ and $\mathbf{V}$.*

*(ii) Every critical point that is not a global minimum is a strict saddle point.*

Item D.20.(ii) implies that every local minimum is a *balanced* global minimum, and every saddle point is a strict saddle point. Therefore, the a.s. global convergence of regularized SpecGF with $\mathcal{T}_\beta$ towards global minima is guaranteed. The proof of Proposition D.20 is found in Udell et al. [56, Sections 2.3 and A.1] and Chen et al. [10, Proposition A.1].

# E Experiments

## E.1 Plots for Empirical Validation

In this section, we list additional plots relevant to Section 6.1. Figure 7 plots the direct loss evolution of SpecGF with $\beta = 0, 10^{-8}$. Figure 8 shows the singular value evolution under SpecGD with $\mathcal{T}$; it is highly similar to Figure 2a, extending our validation of uniform growth. Figure 9 plots the same evolution but under other orthogonalization-based optimizers. Both Shampoo and SOAP succeed in converging to global minima and result in nearly identical slopes of the curves $\sigma_i(\mathbf{A}(t)\mathbf{B}(t))$.

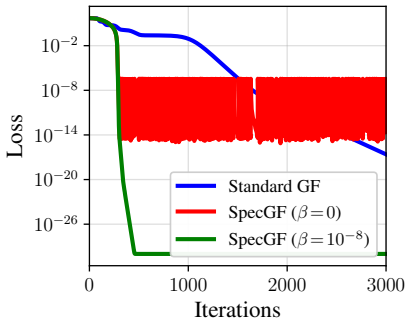



Figure 7: Loss comparison.

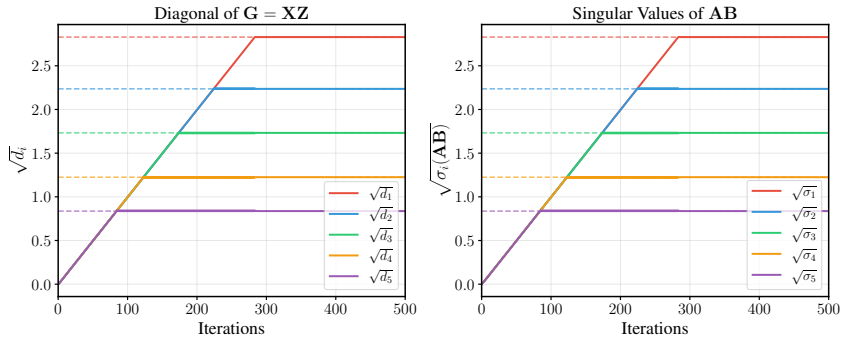
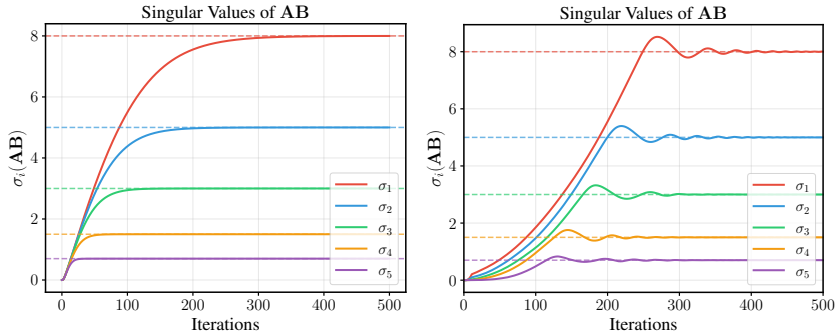



Figure 8: SpecGF ($\beta = 0$) on matrix factorization with a rank-5 target.




(a) Shampoo (b) SOAP

Figure 9: Comparison of orthogonalization optimizers on matrix factorization with a rank-5 target.

### E.2 Matrix Factorization Experiments Under Various Settings

In this section, we present additional experiments that vary the training setting of Figure 2. In Figure 10, we vary the momentum parameter $\mu$ in Equation (1), where $\mu = 0$ corresponds to SpecGD, while applying exact orthogonalization via $\mathcal{T}$ at every update. We observe that the singular values of the product matrix exhibit consistent uniform growth across all choices of $\mu$, indicating that this behavior is robust to the strength of momentum under exact orthogonalization.

In Figure 11, we consider the same range of momentum parameters but replace the exact orthogonalization operator with Newton–Schulz iterations; this setting is identical to the practical Muon optimizer used in real training scenarios. Despite the use of approximate orthogonalization, we again observe uniform growth of singular values, demonstrating that the phenomenon persists under practical implementations of Muon.

In Figure 12, we vary the LoRA rank from 1 to 4, which is strictly smaller than the target rank $5 = \mathrm{rank}(\mathbf{\Sigma})$. In all cases, the SpecGD iteration converges, reaching plateaus at the top-$r$ singular values of $\mathbf{\Sigma}$, and the resulting solution corresponds to the best rank-$r$ approximation of $\mathbf{Y}$.

Finally, we increase the number of LoRA factors (depth), by reparameterizing the weight $\mathbf{W}$ as $\mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1$. All intermediate factors $\mathbf{W}_l$ lie in $\mathbb{R}^{r \times r}$, with the first factor $\mathbf{W}_L$ in $\mathbb{R}^{m \times r}$, and the last factor $\mathbf{W}_1$ in $\mathbb{R}^{r \times n}$. Only the first factor $\mathbf{W}_L$ is initialized at zero, while the remaining factors are initialized as i.i.d. Gaussian matrices scaled by $\gamma$. In Figure 13, we observe that the uniform growth of singular values persists as depth increases, and Figure 14 further reveals that the singular values grow as $\Theta(t^L)$ for depth $L$.

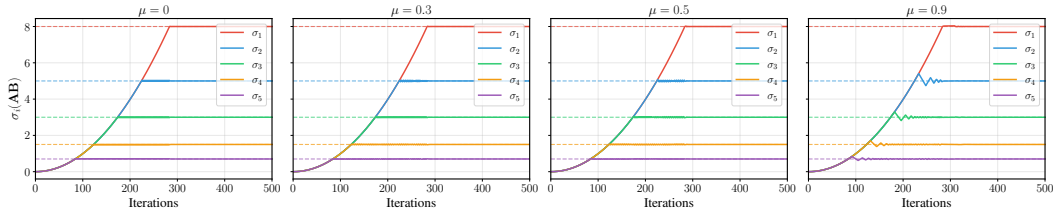



Figure 10: Comparison of singular value evolutions across $\mu \in \{0, 0.3, 0.5, 0.9\}$ from exact orthogonalization $\mathcal{T}$. Notice the consistent uniform growth.

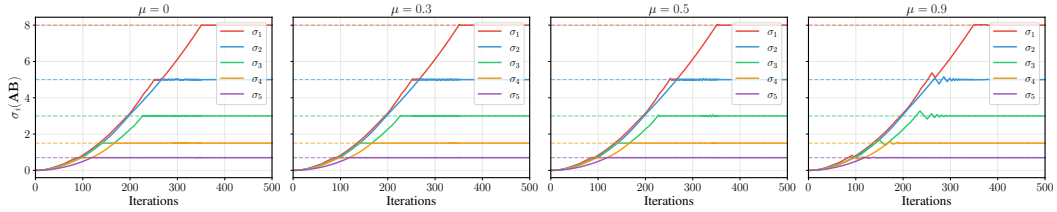



Figure 11: Comparison of singular value evolutions across $\mu \in \{0, 0.3, 0.5, 0.9\}$ from Newton-Schulz iterations. Notice the consistent uniform growth.

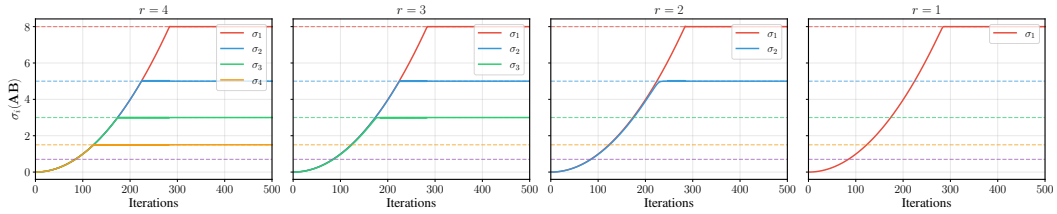



Figure 12: Comparison of singular value evolutions across LoRA rank $r \in \{1, 2, 3, 4\}$. Notice the convergence toward the best rank-$r$ approximation.

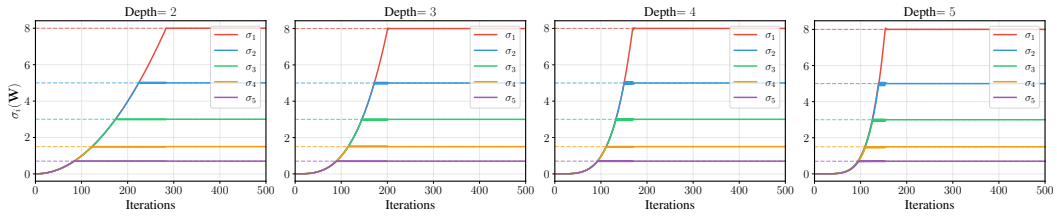



Figure 13: Comparison of singular value evolutions across the depth $L \in \{2, 3, 4, 5\}$. Notice the consistent uniform growth.

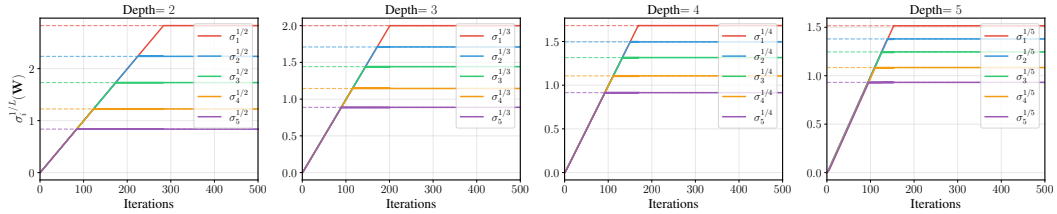



Figure 14: Comparison of the $L$-th root of the singular value evolutions across the depth $L \in \{2, 3, 4, 5\}$. Notice the consistent uniform *linear* growth.

### E.3 LLM Experiments

In this section, we demonstrate that near-uniform growth is consistently observed for LoRA fine-tuning with orthogonalization-based optimizers and nonzero LoRA initialization schemes.

#### E.3.1 Experimental Details

We fine-tune a pre-trained RoBERTa-base model (125M) [41] on the SST-2, RTE, and CoLA datasets [59] using LoRA adapters applied to the query and value projection matrices. We set the LoRA rank to $r = 8$ and the scaling factor to $\alpha = 16$. We trained the model for 10 epochs with a batch size of 16 using AdamW and Muon optimizers, with a weight decay of 0.01. The learning rate was set to $10^{-4}$ and scheduled using a cosine decay. We also fine-tune a ViT-B [14] on the CIFAR100 [37], Caltech101 [19], and EuroSAT [27] datasets using LoRA adapters applied to the query, key, and value projection matrices. We set the LoRA rank to $r = 8$ and the patch size to $16$. We trained the model for 100 epochs using AdamW and Muon optimizers, with a weight decay of $10^{-4}$. We set $\mu = 0.95$ for Muon, $\beta_1 = 0.9$ and $\beta_2 = 0.95$ for AdamW. The learning rate was set to $10^{-3}$. We used NVIDIA RTX A6000 GPUs for these training runs.

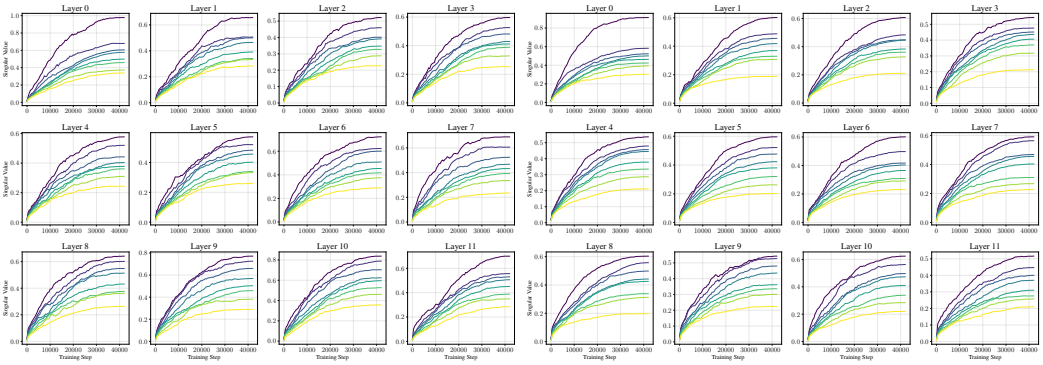
For larger-scale experiments, we train LLaMA-3.2-1B [23] on the Alpaca dataset [53] for supervised fine-tuning. We also train a larger model, LLaMA-3.2-3B, on the Alpaca, OpenHermes 2.5 [54], and UltraChat [13] datasets. We randomly sampled one million instances from OpenHermes 2.5. As in RoBERTa-base, we also use the same LoRA configuration, i.e., rank $r = 8$ and the scaling factor $\alpha = 16$. We train for $3$ epochs with a batch size of $128$ using AdamW and Muon optimizers and a weight decay of 0.01. We use cosine learning rate scheduling with warmup over the first $4\%$ of the training iterations. We used NVIDIA A100 GPUs for these training runs.

#### E.3.2 Singular Value Dynamics

We provide the full training dynamics of the singular values of the LoRA adapter $\mathbf{AB}$ associated with projection matrices. Each line in each subplot stands for a single singular value.

The singular values of all LoRA $\mathbf{AB}$ layers trained under Muon consistently and nearly uniformly evolve. On the other hand, those of LoRA $\mathbf{AB}$ layers trained under AdamW vary from matrix to matrix.

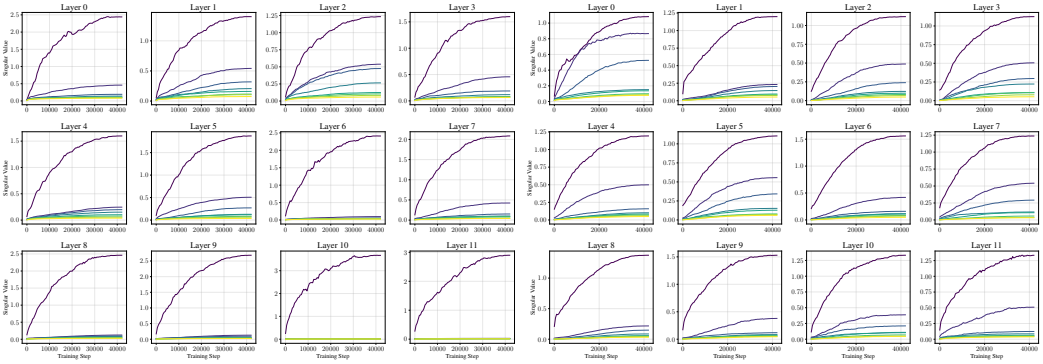
**RoBERTa-Base with SST-2/CoLA/RTE datasets.**



(a) Muon, $Q$ matrix



(b) Muon, $V$ matrix

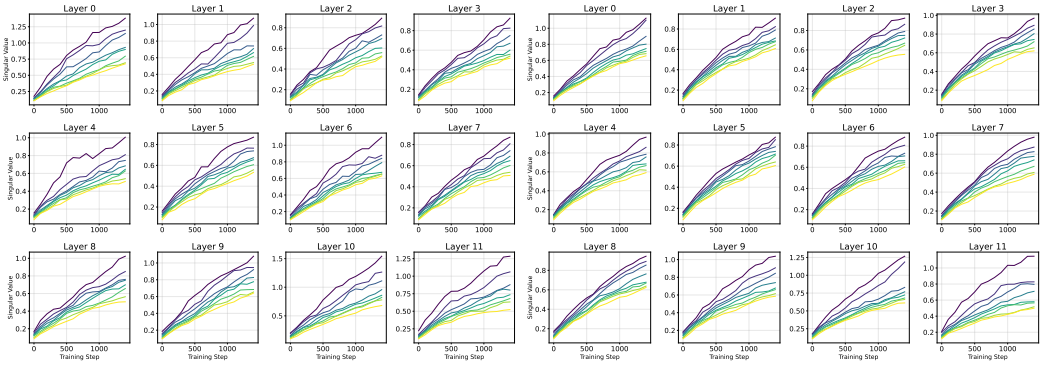


(c) AdamW, $Q$ matrix



(d) AdamW, $V$ matrix

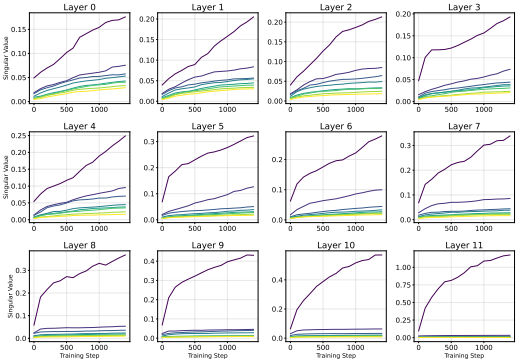
Figure 15: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained on the SST-2 dataset.



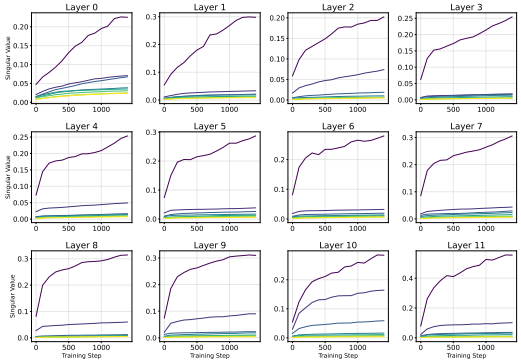

(a) Muon, $Q$ matrix




(b) Muon, $V$ matrix

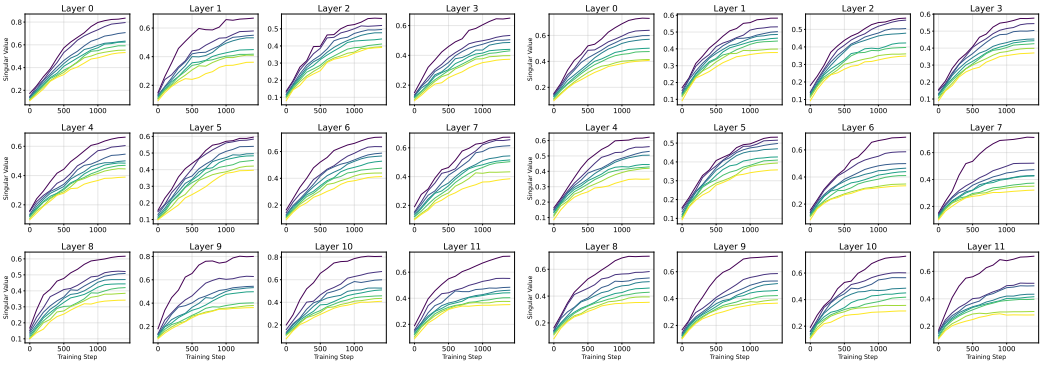



(c) AdamW, $Q$ matrix




(d) AdamW, $V$ matrix

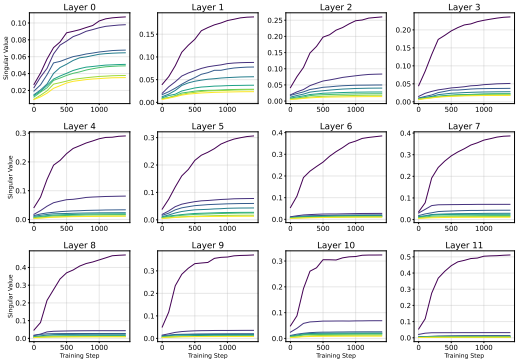
Figure 16: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained on the CoLA dataset.



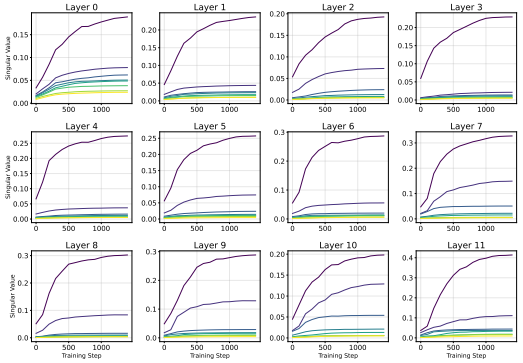

(a) Muon, $Q$ matrix




(b) Muon, $V$ matrix



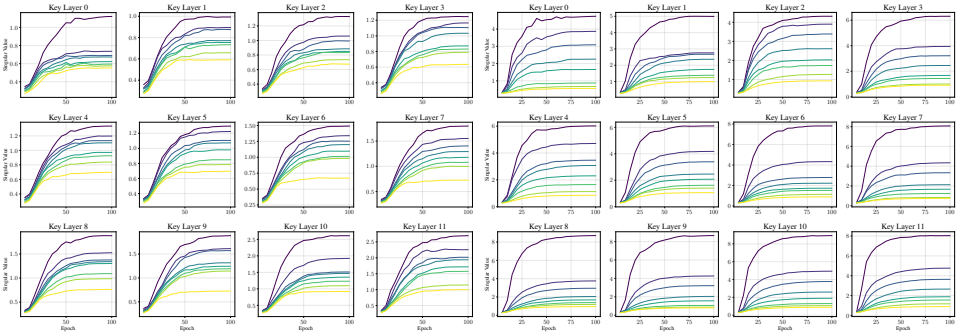

(c) AdamW, $Q$ matrix




(d) AdamW, $V$ matrix

Figure 17: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained on the RTE dataset.

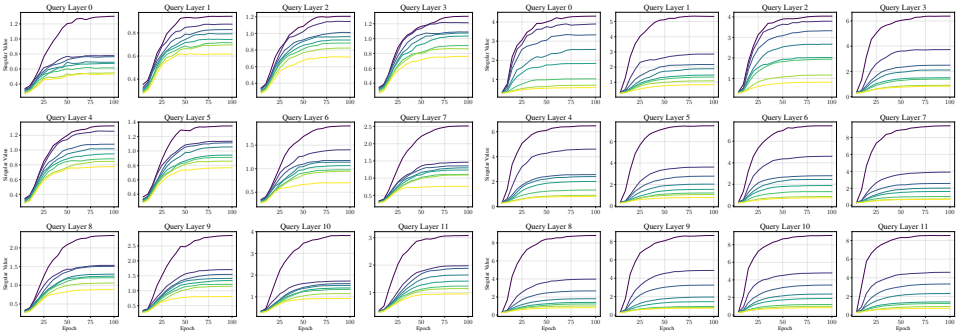
**ViT-B with CIFAR100/Caltech101/EuroSAT datasets.**




(a) Muon, $K$ matrix



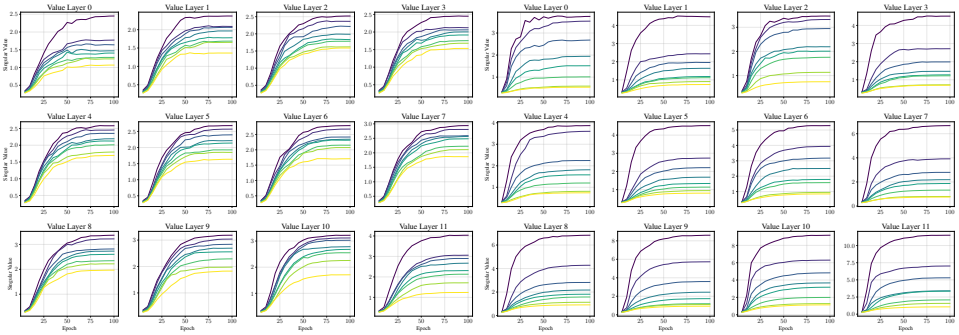

(b) AdamW, $K$ matrix




(c) Muon, $Q$ matrix




(d) AdamW, $Q$ matrix

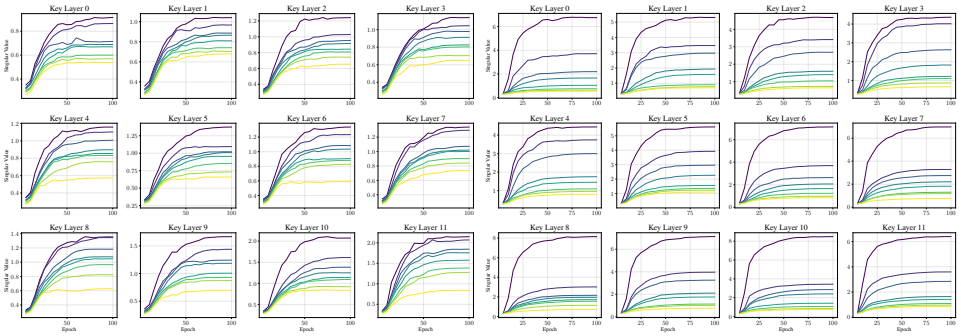



(e) Muon, $V$ matrix




(f) AdamW, $V$ matrix

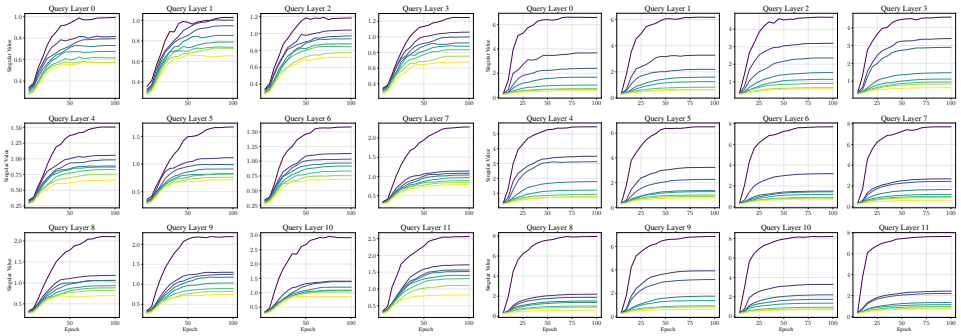
Figure 18: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained on the CIFAR100 dataset.




(a) Muon, $K$ matrix



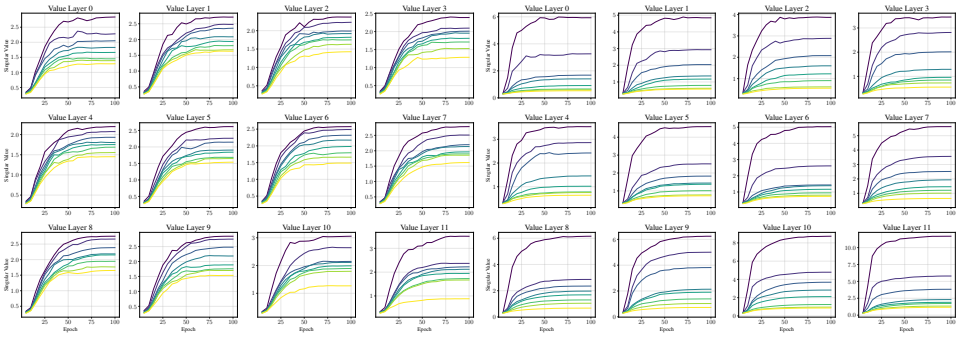

(b) AdamW, $K$ matrix




(c) Muon, $Q$ matrix




(d) AdamW, $Q$ matrix

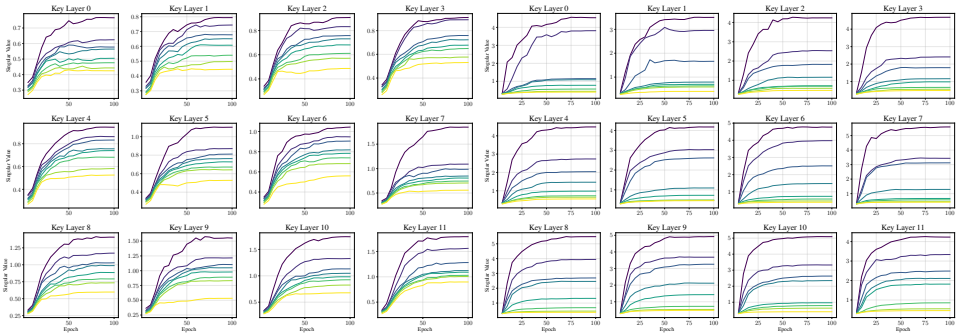



(e) Muon, $V$ matrix




(f) AdamW, $V$ matrix

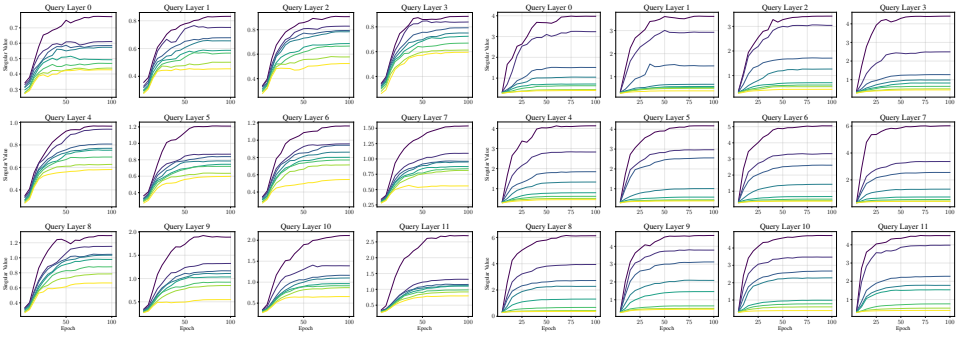
Figure 19: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained on the Caltech101 dataset.




(a) Muon, $K$ matrix



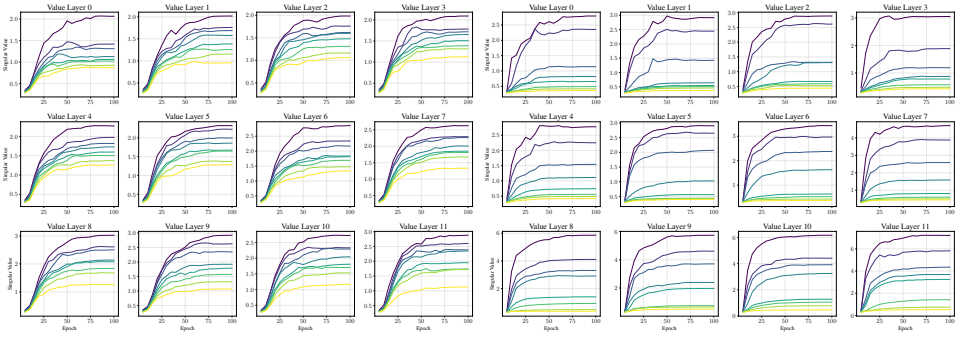

(b) AdamW, $K$ matrix




(c) Muon, $Q$ matrix




(d) AdamW, $Q$ matrix



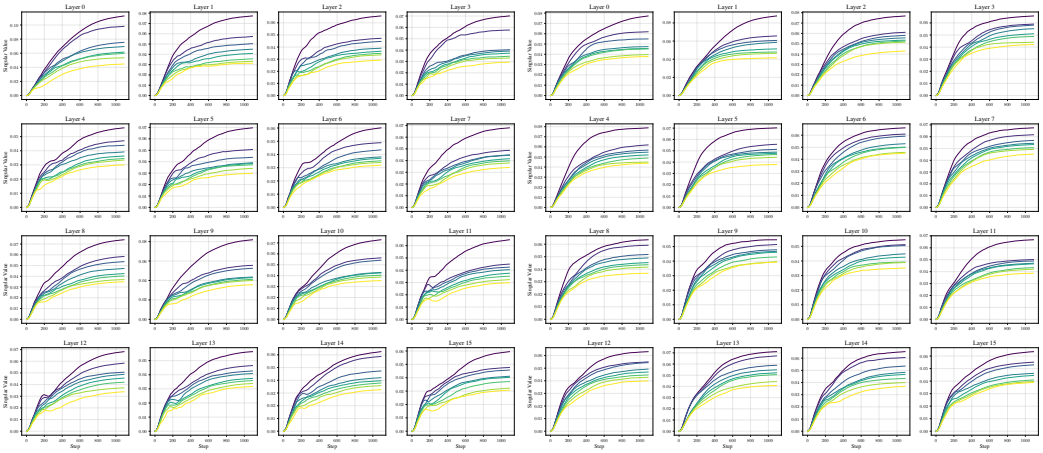

(e) Muon, $V$ matrix




(f) AdamW, $V$ matrix

Figure 20: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained on the EuroSAT dataset.

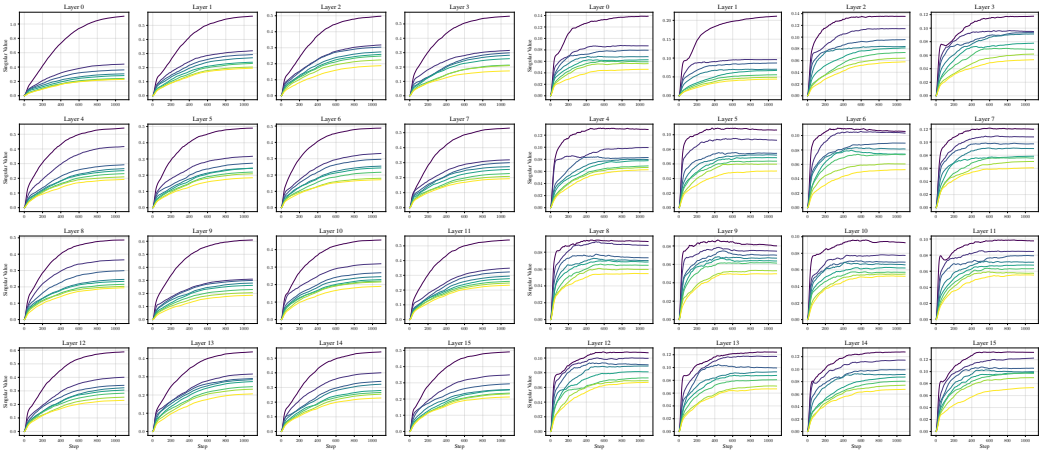
**LLaMA-3.2-1B with Alpaca dataset.**



(a) Muon, $Q$ matrix



(b) Muon, $V$ matrix



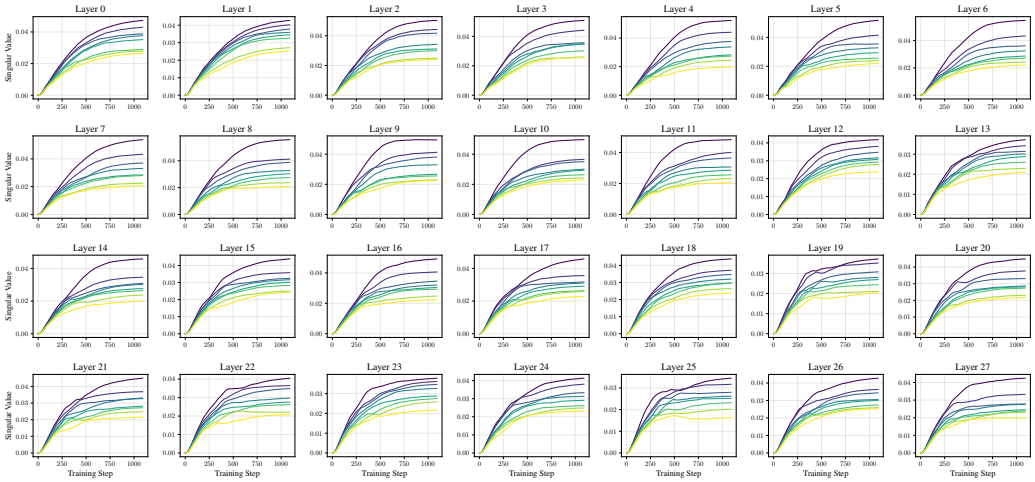
(c) AdamW, $Q$ matrix



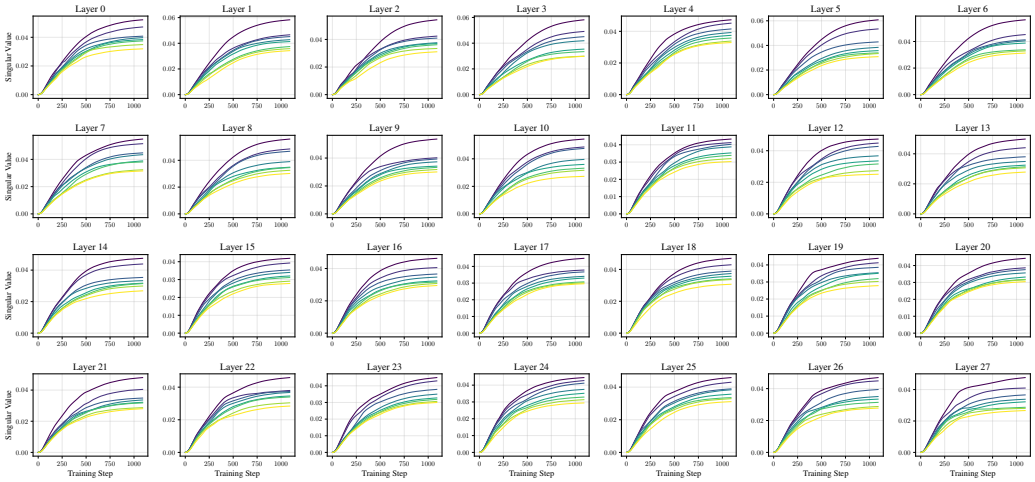
(d) AdamW, $V$ matrix

Figure 21: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-1B trained on the Alpaca dataset.

**LLaMA-3.2-3B with Alpaca/OpenHermes 2.5/UltraChat datasets.**

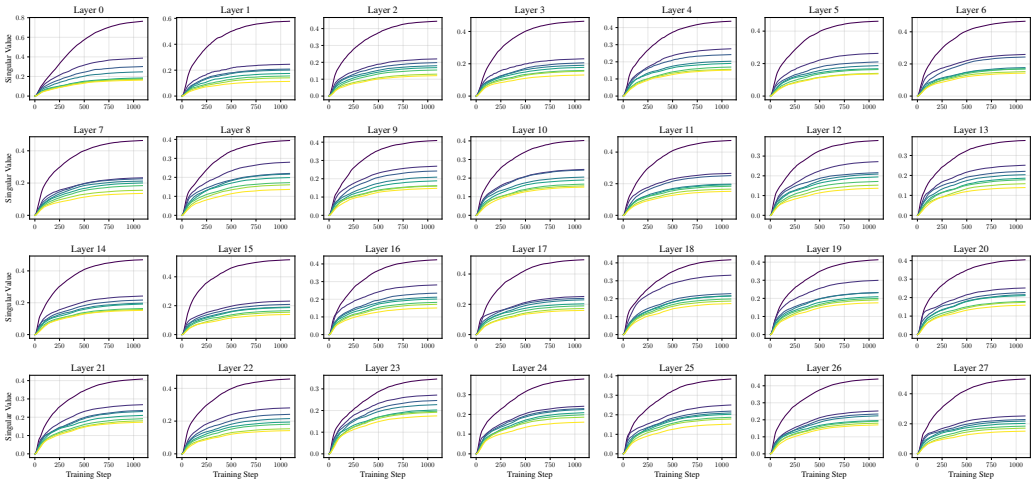



(a) $Q$ matrix

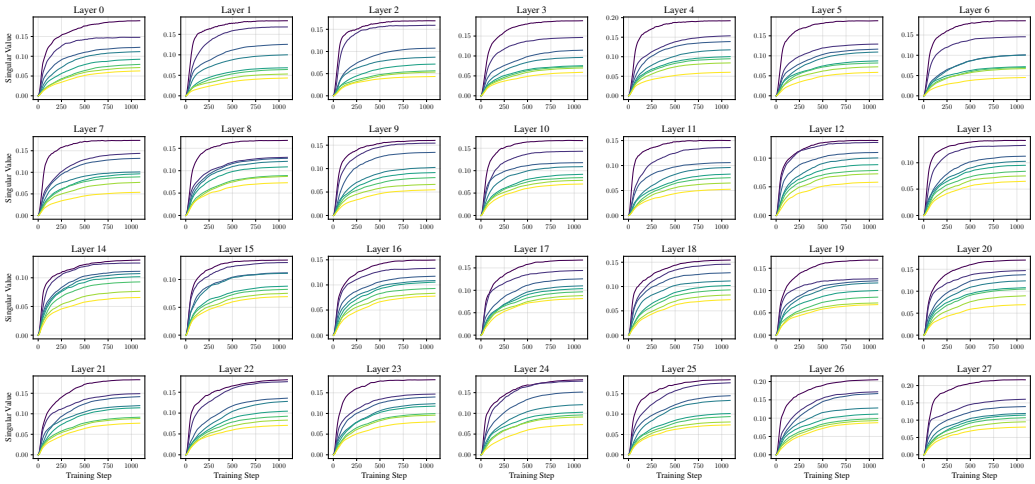



(b) $V$ matrix

Figure 22: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Muon on the Alpaca dataset.

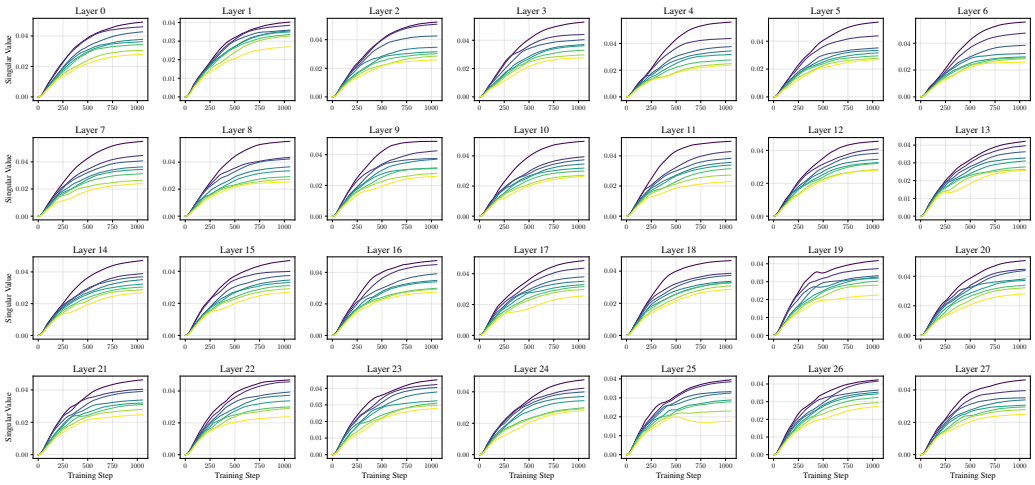



(a) $Q$ matrix

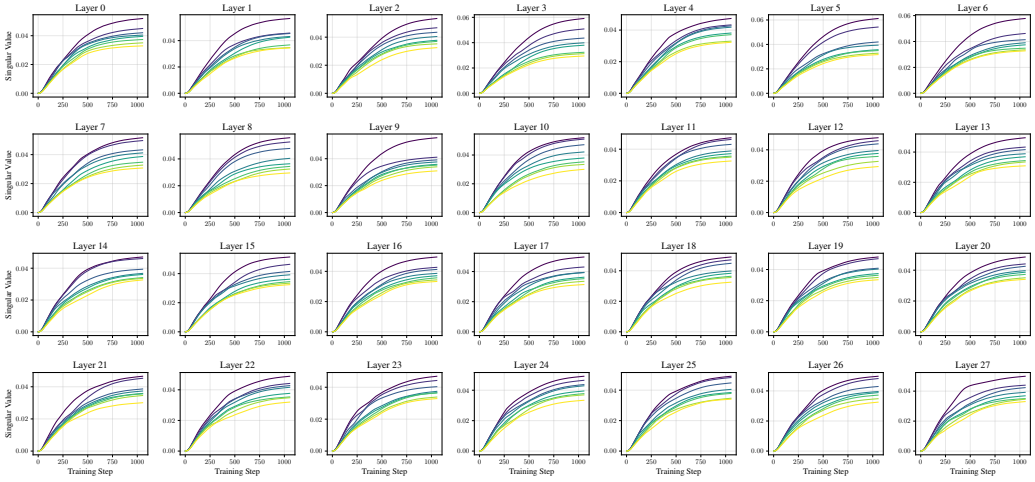



(b) $V$ matrix

Figure 23: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using AdamW on the Alpaca dataset.

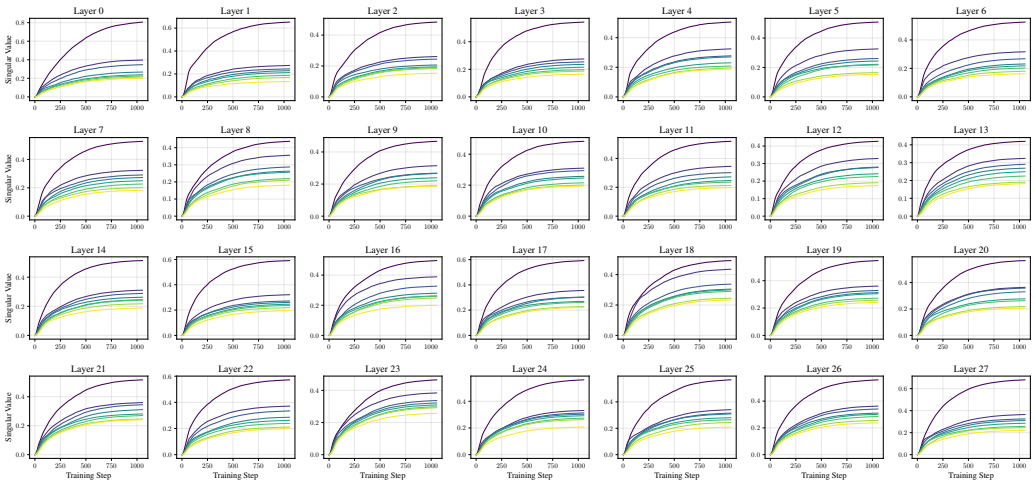



(a) $Q$ matrix

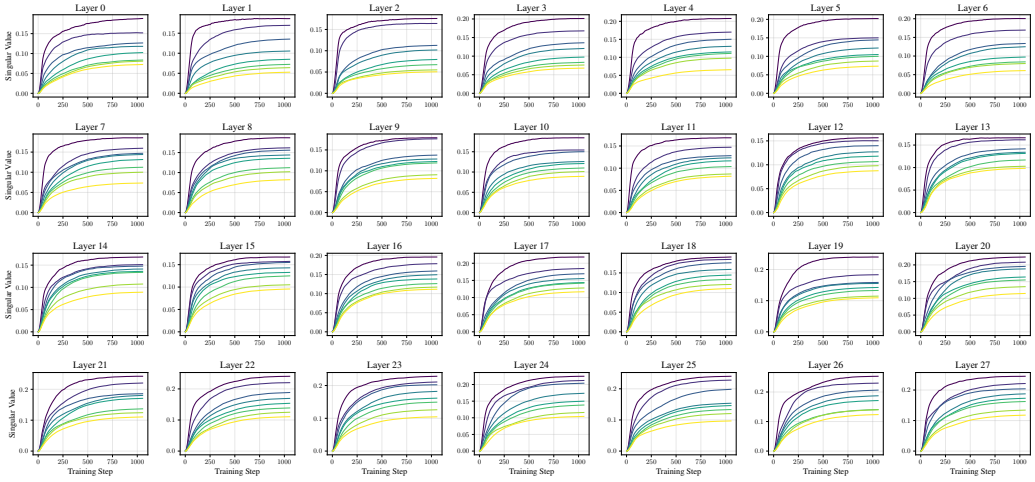



(b) $V$ matrix

Figure 24: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Muon on the OpenHermes 2.5 dataset.

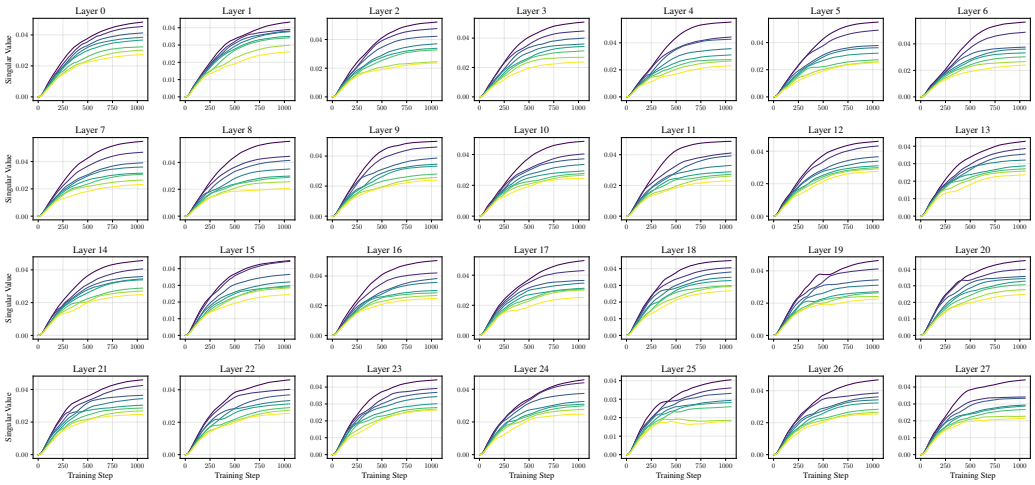



(a) $Q$ matrix

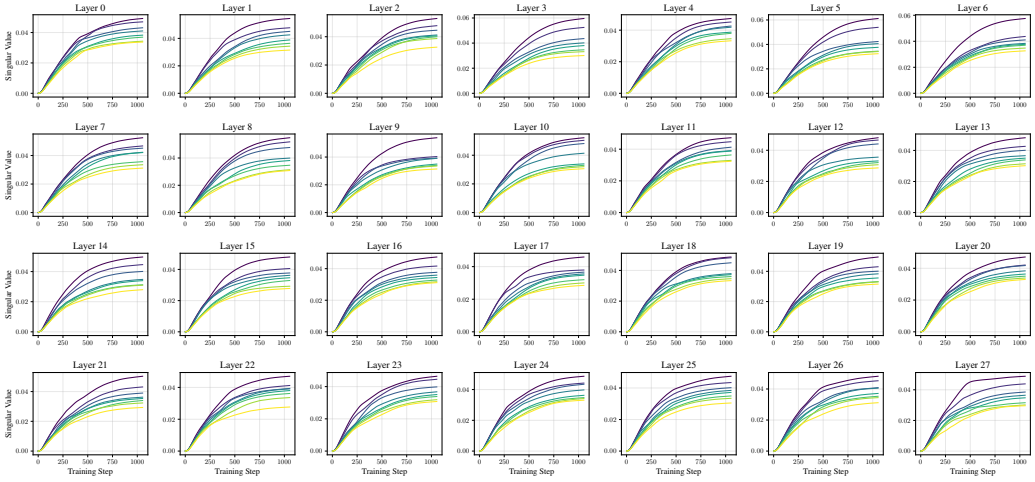



(b) $V$ matrix

Figure 25: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using AdamW on the OpenHermes 2.5 dataset.

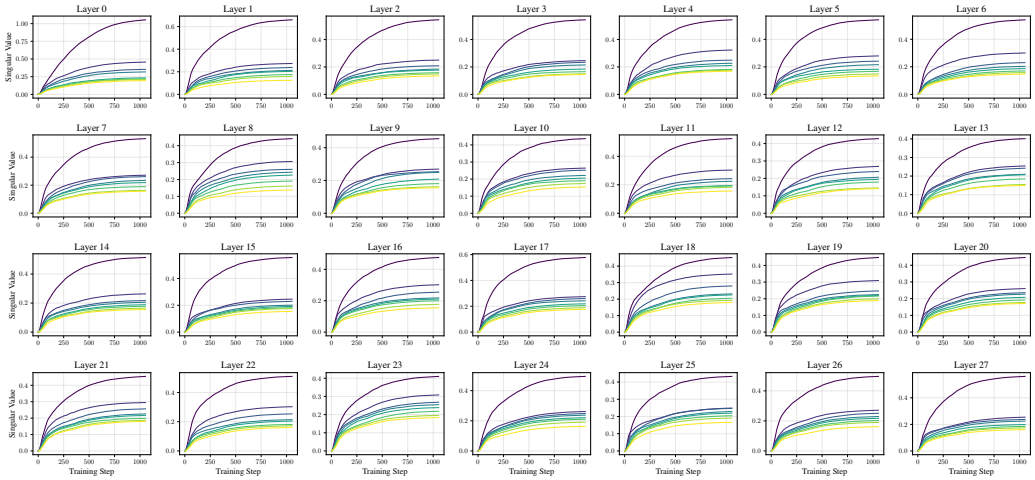



(a) $Q$ matrix

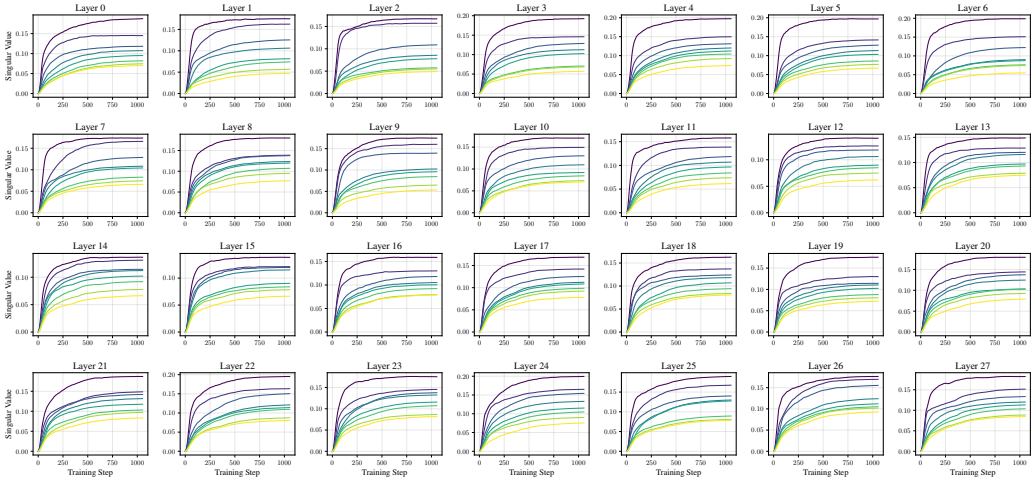



(b) $V$ matrix

Figure 26: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Muon on the UltraChat dataset.




(a) $Q$ matrix



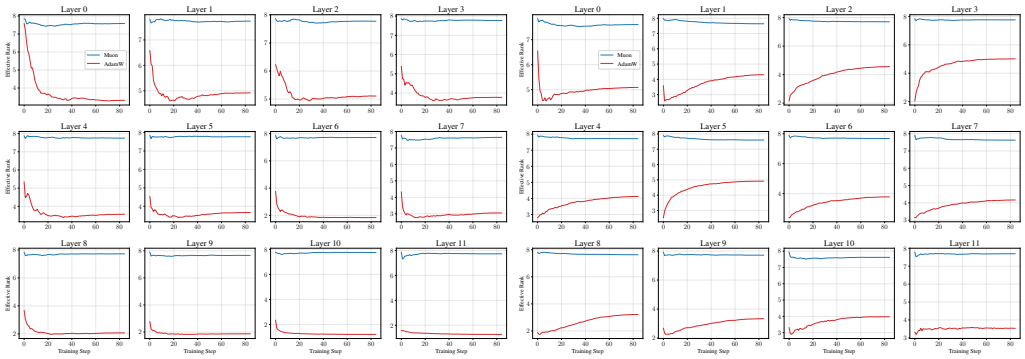

(b) $V$ matrix

Figure 27: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using AdamW on the UltraChat dataset.

### E.3.3 Effective Rank

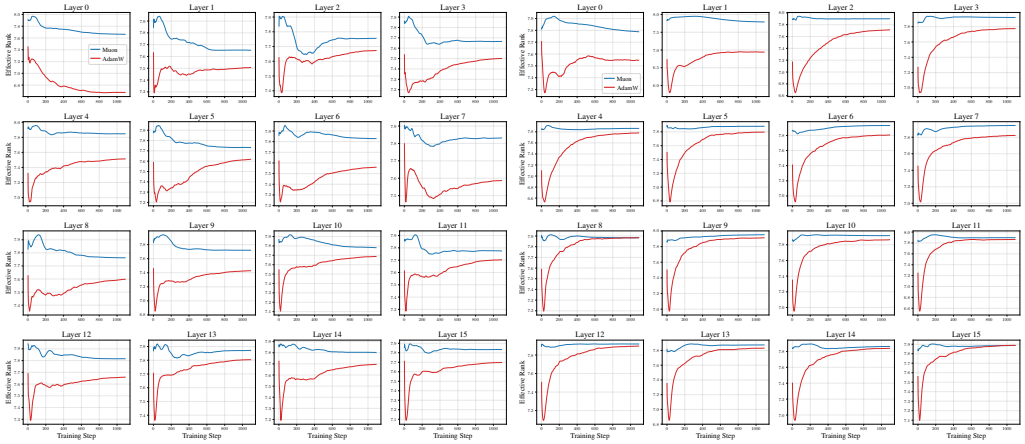
We provide the evolution of the effective rank of the LoRA adapter $\mathbf{AB}$ associated with query $Q$ and value $V$ projection matrices under the following settings: RoBERTa-Base with the SST-2 dataset and LLaMA-3.2-1B with the Alpaca dataset. Blue and red lines are the effective ranks of the matrix trained under Muon and AdamW, respectively.

Figure 28 shows the effective rank evolutions for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base; Figure 29 shows those in LLaMA-3.2-1B. The effective rank of matrices trained under Muon maintains a stable value, close to the LoRA rank $r = 8$. That of matrices trained under AdamW exhibits non-trivial evolution.




(a) $Q$ matrix (b) $V$ matrix

Figure 28: Effective rank of the LoRA $\mathbf{AB}$ layers for the $Q$ and $V$ matrices in RoBERTa-Base.




(a) $Q$ matrix (b) $V$ matrix

Figure 29: Effective rank of the LoRA $\mathbf{AB}$ matrices for the $Q$ and $V$ in LLaMA-3.2-1B.

**$VO$ products during Transformer pretraining.** To test whether a similar product-level spectral behavior can be observed in distinct settings where two trainable parameter matrices enter the model as a product and are optimized by orthogonalized updates, we adopted the $VO$-product spectral analysis [60]. Specifically, Wang et al. [60] report that Muon yields more isotropic spectra than Adam does for the $VO$-product, but they do not report the effective rank separately for each layer. We measure the effective rank of the $VO$ product in each of the 12 layers under the Transformer pretraining setup of Semenov et al. [50]. We use a constant learning rate of $10^{-3}$ after 2,000 warmup steps. Table 1 reports the mean and standard deviation of effective rank over recorded training steps for each layer.

Table 1: Effective rank of $VO$-products during Transformer pretraining. The first three rows for each optimizer report values at selected training steps. Mean and Std report the mean and standard deviation over all steps. Layer avg. averages the corresponding values across the 12 layers.

| Step\Layer | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Layer avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AdamW | | | | | | | | | | | | | |
| 2K | 446.8 | 303.2 | 321.1 | 336.7 | 356.9 | 349.4 | 329.3 | 326.8 | 305.1 | 338.7 | 328.7 | 318.1 | 338.4 |
| 9K | 236.8 | 204.0 | 276.1 | 322.4 | 385.6 | 369.9 | 382.8 | 355.1 | 350.1 | 379.4 | 356.0 | 347.2 | 330.5 |
| 16K | 249.0 | 237.0 | 326.8 | 353.2 | 411.2 | 394.6 | 409.5 | 390.6 | 383.1 | 409.6 | 391.6 | 391.9 | 362.3 |
| Mean | 308.7 | 256.2 | 312.0 | 343.6 | 391.8 | 378.7 | 385.1 | 367.1 | 356.8 | 384.9 | 366.3 | 357.7 | 350.7 |
| Std | 93.8 | 86.8 | 67.6 | 49.5 | 35.7 | 38.8 | 41.5 | 46.1 | 47.1 | 42.0 | 47.6 | 53.6 | 54.2 |
| Muon | | | | | | | | | | | | | |
| 2K | 414.4 | 427.8 | 421.6 | 423.4 | 427.7 | 423.9 | 416.6 | 413.0 | 416.6 | 428.3 | 426.7 | 426.7 | 422.2 |
| 9K | 367.3 | 472.2 | 457.9 | 460.7 | 457.4 | 457.2 | 435.3 | 443.8 | 437.3 | 449.7 | 428.7 | 462.9 | 444.2 |
| 16K | 327.9 | 483.0 | 466.0 | 475.1 | 470.3 | 469.8 | 442.1 | 452.1 | 446.2 | 459.4 | 437.5 | 477.6 | 450.6 |
| Mean | 380.5 | 466.3 | 454.3 | 457.4 | 455.8 | 455.7 | 437.2 | 442.6 | 439.3 | 451.0 | 435.3 | 461.1 | 444.7 |
| Std | 44.2 | 22.5 | 20.7 | 21.6 | 20.1 | 20.6 | 20.9 | 21.9 | 20.5 | 18.2 | 20.7 | 20.0 | 22.7 |

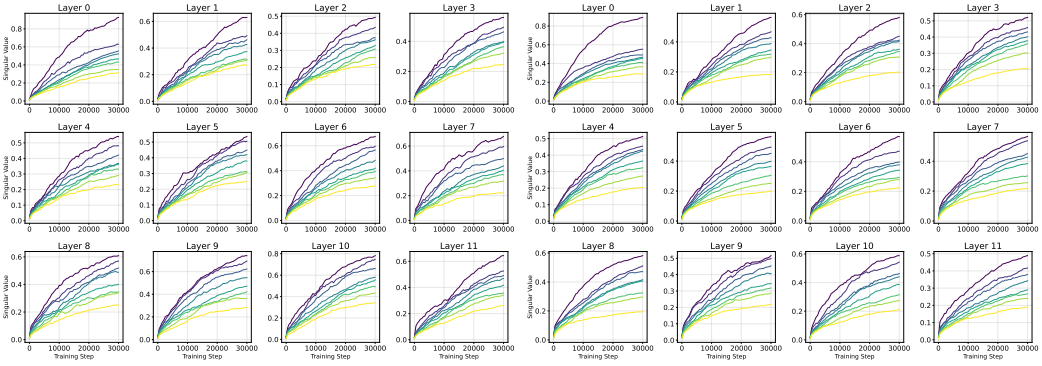
We observe that the effective rank of the VO product is consistently higher and more stable for Muon than AdamW. These results provide empirical evidence about product-level spectra during pretraining.

### E.3.4 Uniform Growth under LoRA Initialization Schemes

We fine-tune a pre-trained RoBERTa-base model on the SST-2 dataset using LoRA adapters applied to the query and value matrices, but with nonzero initialization schemes for LoRA factors. We applied two schemes: nonzero random initialization (NRI) where $\mathbf{A}$ and $\mathbf{B}$ are both randomly sampled from the Kaiming uniform distribution, and Locally Optimal Initialization (LOI) from Bogachev et al. [8]. LOI uses spectral information from the initial loss gradient to choose the adapter's initial subspaces. We used NVIDIA RTX A6000 GPUs for these training runs.

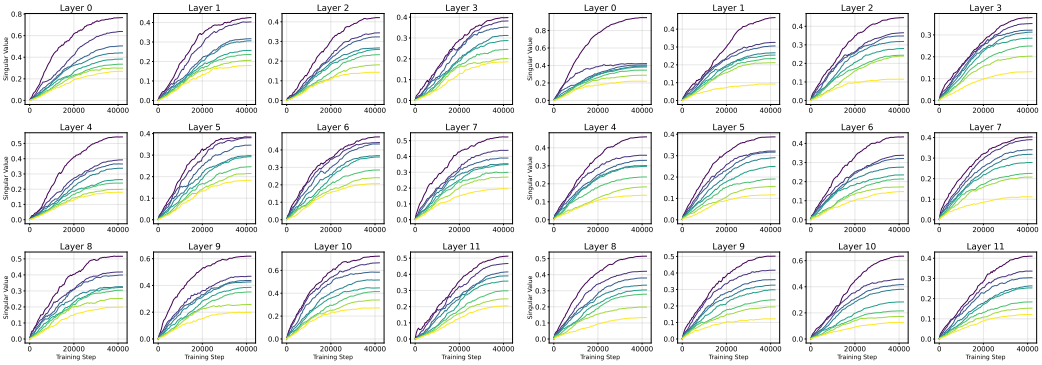
We provide the full training dynamics of the singular values of the LoRA adapter $\mathbf{AB}$ associated with query $Q$ and value $V$ projection matrices. Each line in each subplot stands for a single singular value.

Figure 30 shows the singular value evolutions for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under nonzero random initialization; Figure 31 shows those under LOI. Note that the uniform growth persists for these nonzero LoRA initialization schemes.




(a) $Q$ matrix (b) $V$ matrix

Figure 30: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under NRI.




(a) $Q$ matrix (b) $V$ matrix

Figure 31: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under LOI.

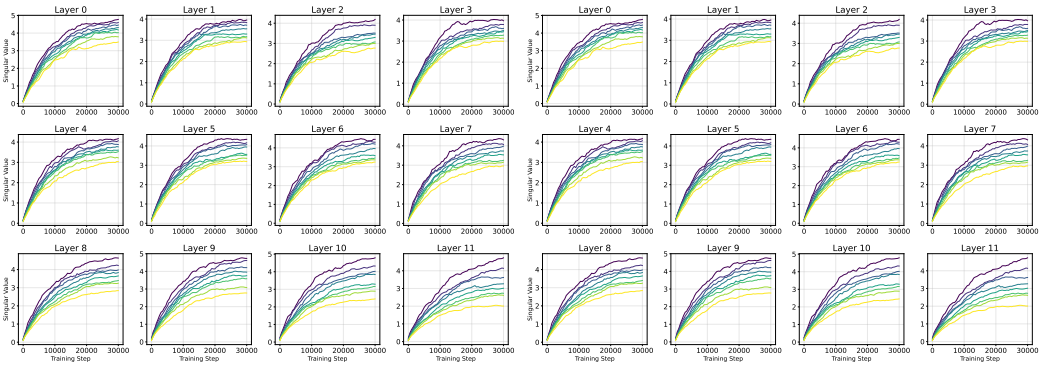
### E.3.5 Uniform Growth under Other Matrix Optimizers

We fine-tune LLaMA-3.2-3B with the Alpaca, OpenHermes, and UltraChat datasets under additional matrix optimizers: Shampoo [25], SOAP [58], and Dion [2] optimizers. We also fine-tune RoBERTa-Base with the SST-2 dataset under SOAP. We use the same LoRA configuration as in Section E.3.1. We used NVIDIA A100 GPUs for these training runs.

We provide the full training dynamics of the singular values of the LoRA adapter $\mathbf{AB}$ associated with query $Q$ and value $V$ matrices. Each line in each subplot stands for a single singular value.

For all orthogonalization optimizers, the singular values of all LoRA $\mathbf{AB}$ layers consistently and nearly uniformly evolve.

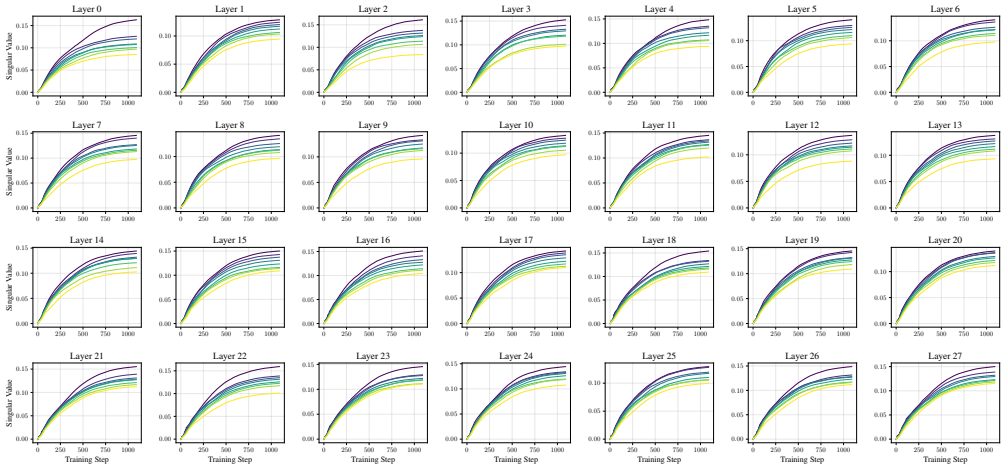
**RoBERTa-Base with SOAP optimizer.**



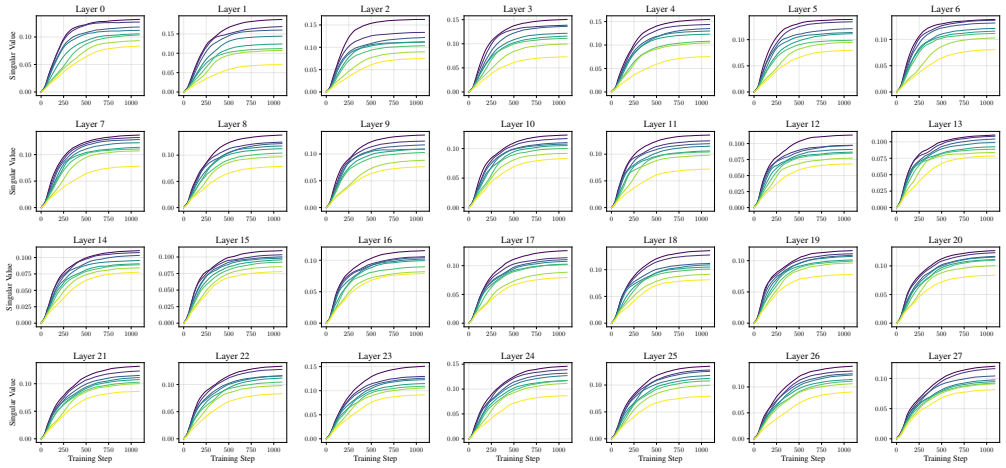

(a) $Q$ matrix (b) $V$ matrix

Figure 32: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained using SOAP on the SST-2 dataset.

**LLaMA-3.2-3B with SOAP optimizer.**

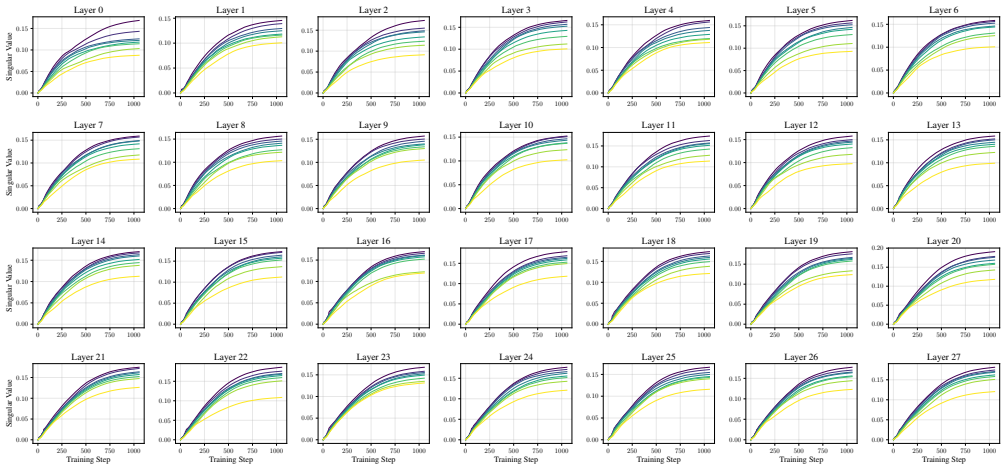



(a) $Q$ matrix

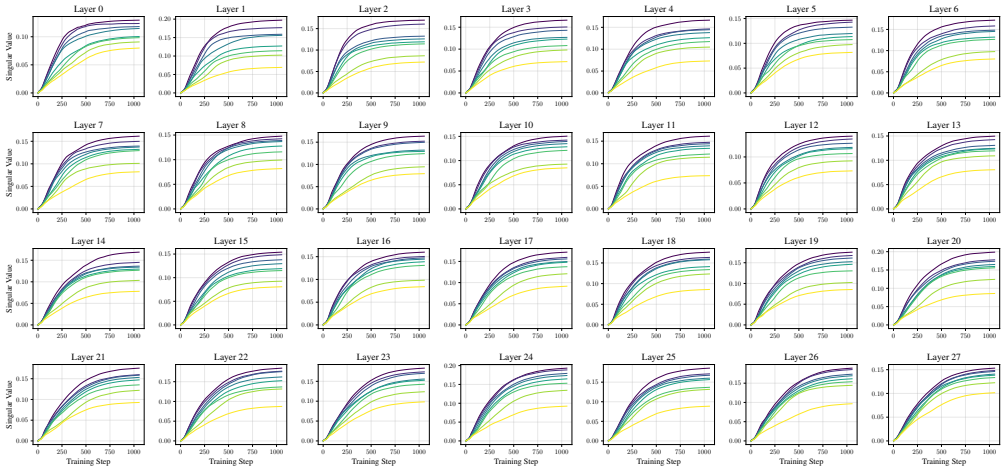



(b) $V$ matrix

Figure 33: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using SOAP on the Alpaca dataset.

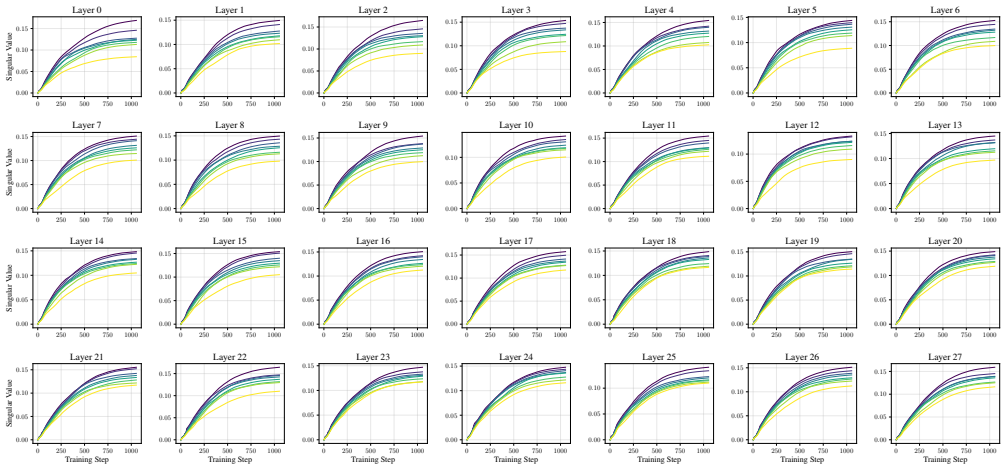



(a) $Q$ matrix

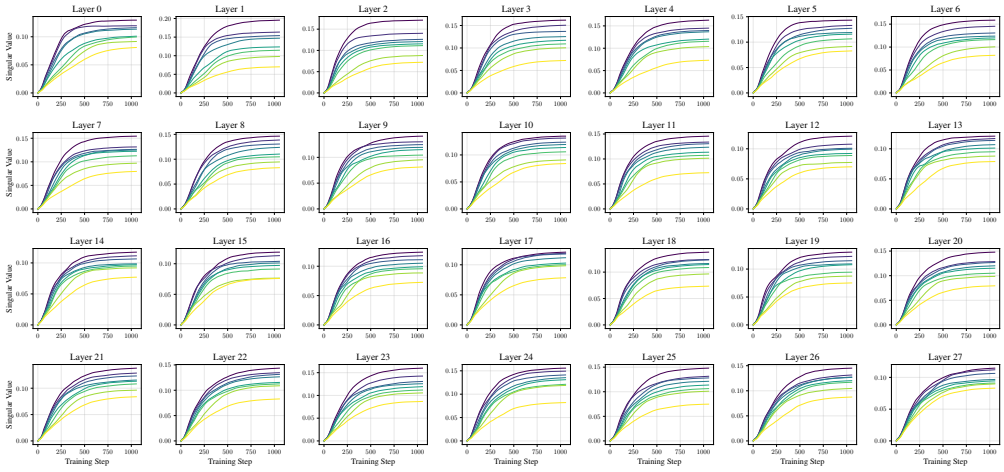



(b) $V$ matrix

Figure 34: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using SOAP on the OpenHermes 2.5 dataset.



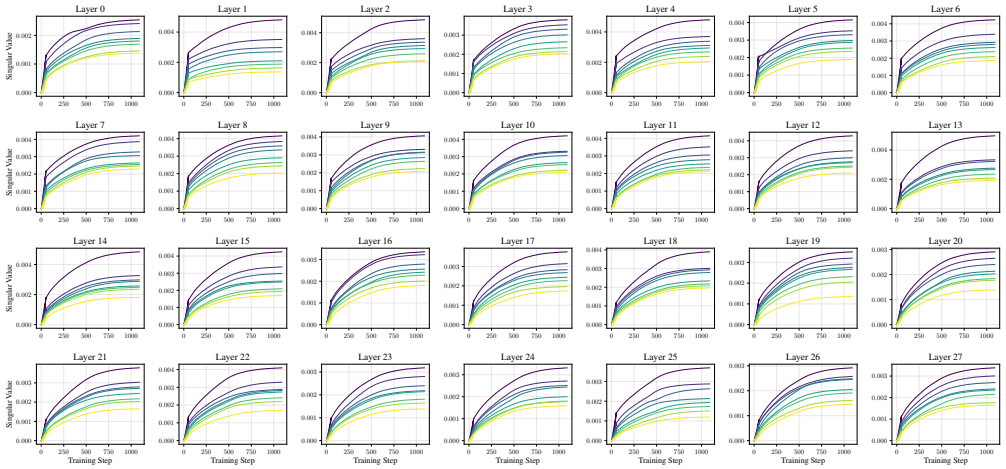

(a) $Q$ matrix



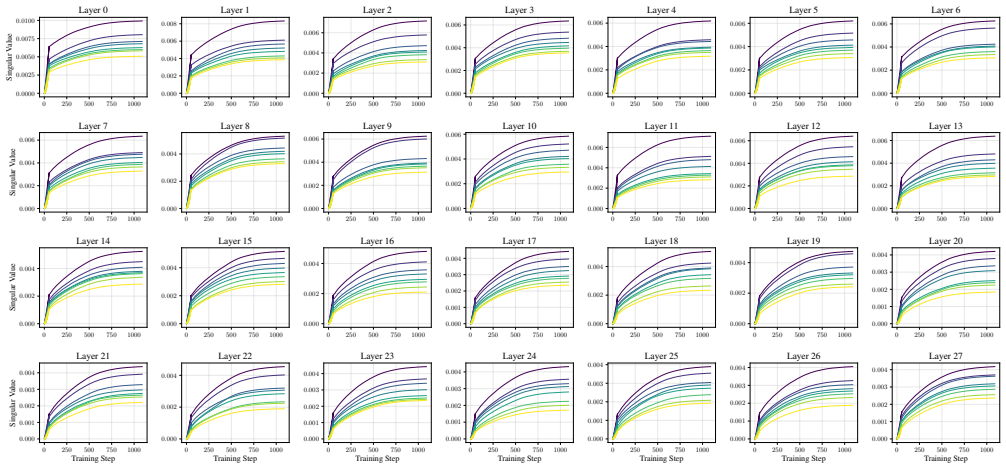

(b) $V$ matrix

Figure 35: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using SOAP on the UltraChat dataset.

**LLaMA-3.2-3B with Shampoo optimizer.**

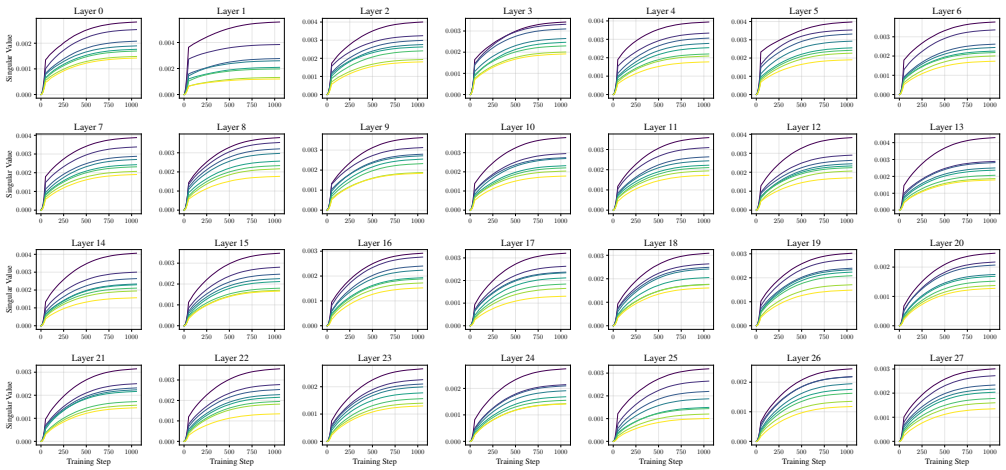



(a) $Q$ matrix

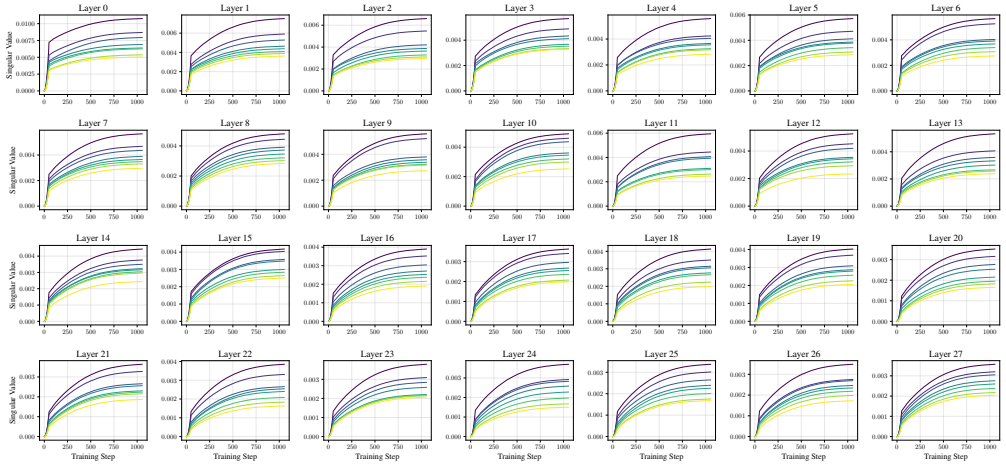



(b) $V$ matrix

Figure 36: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Shampoo on the Alpaca dataset.

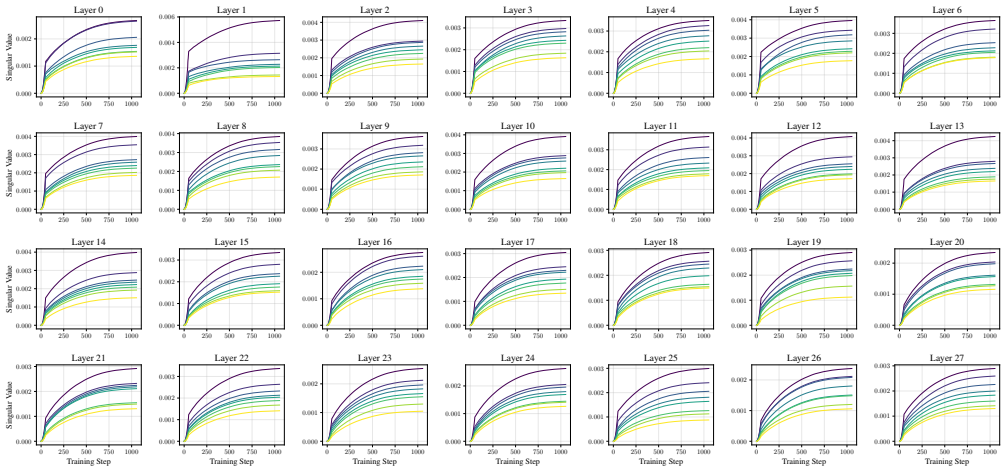



(a) $Q$ matrix

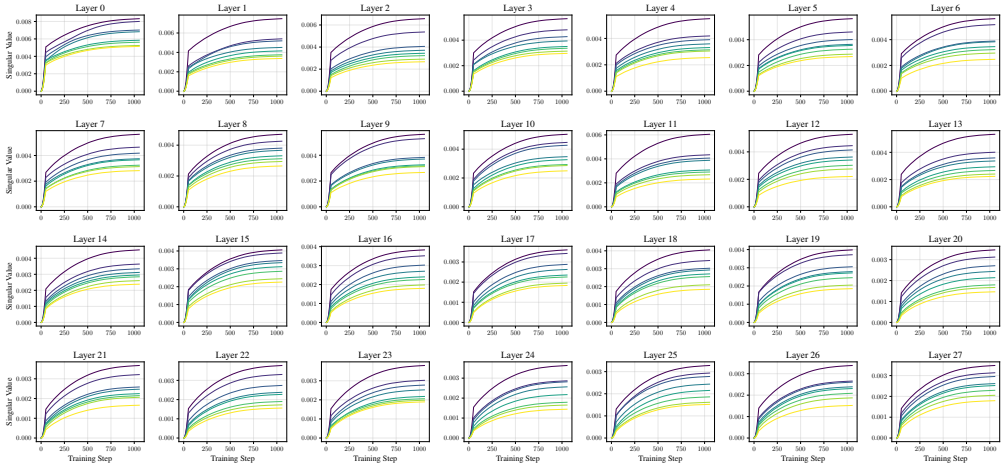



(b) $V$ matrix

Figure 37: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Shampoo on the OpenHermes 2.5 dataset.



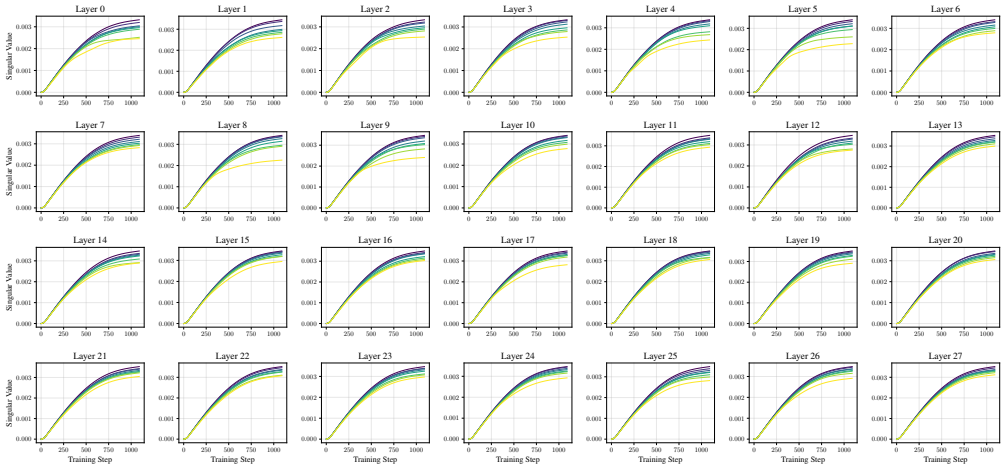

(a) $Q$ matrix



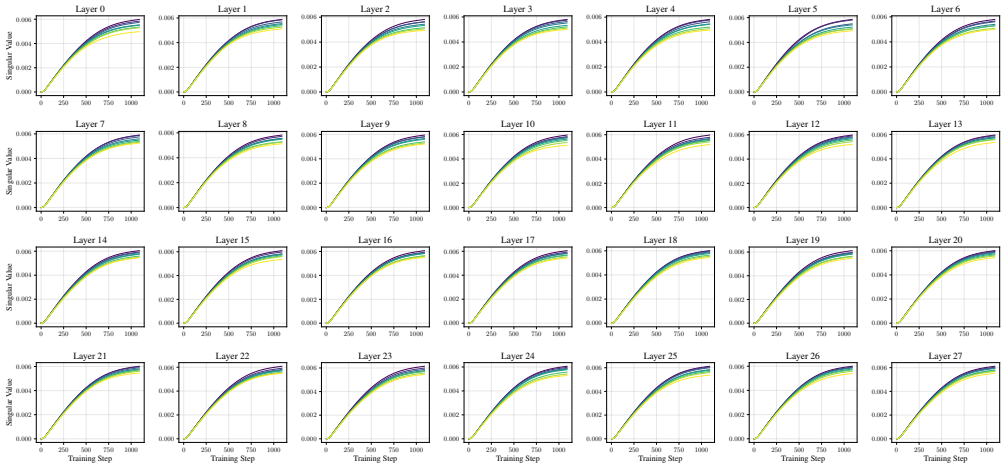

(b) $V$ matrix

Figure 38: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Shampoo on the UltraChat dataset.

**LLaMA-3.2-3B with Dion optimizer.**

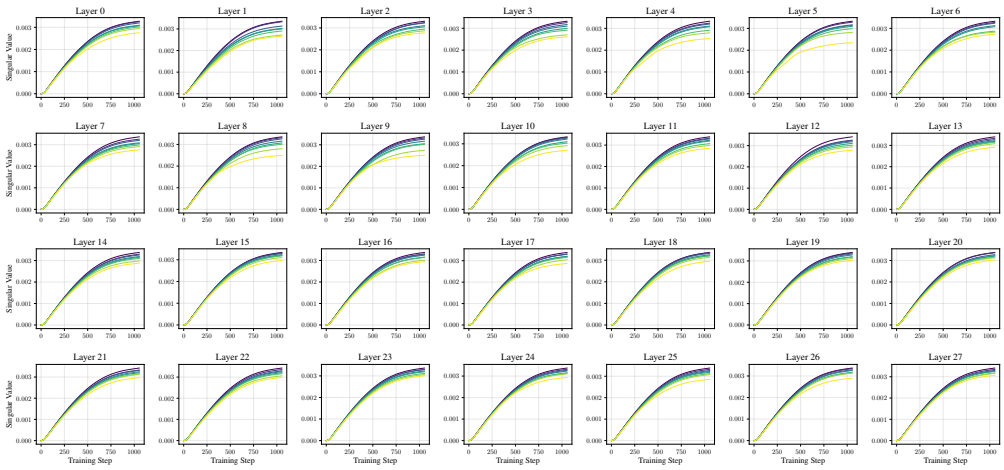



(a) $Q$ matrix

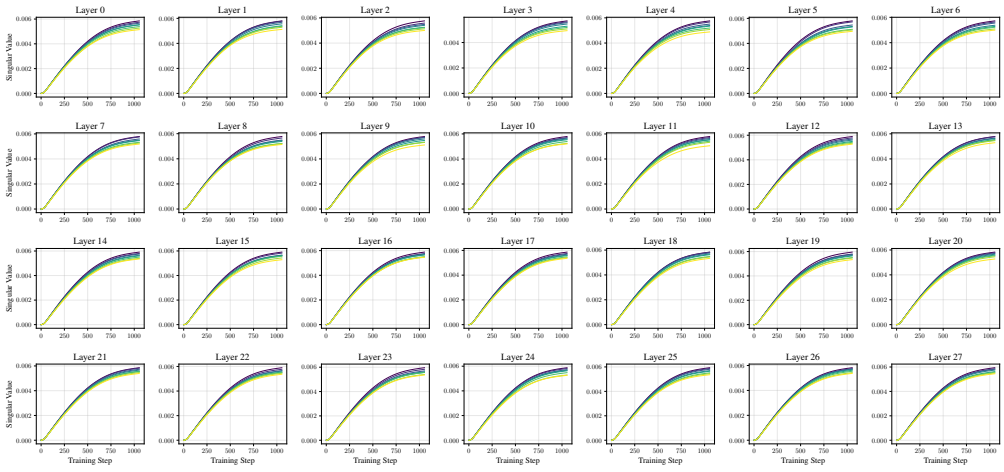



(b) $V$ matrix

Figure 39: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Dion on the Alpaca dataset.

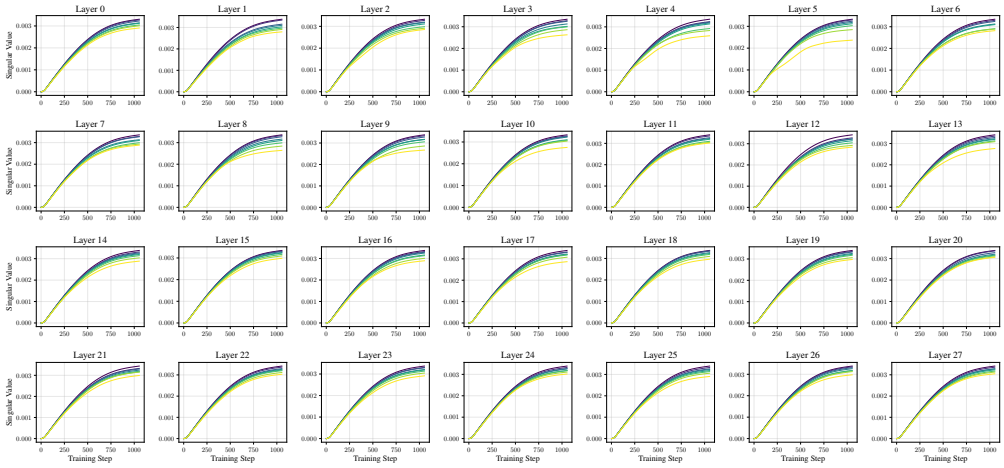



(a) $Q$ matrix

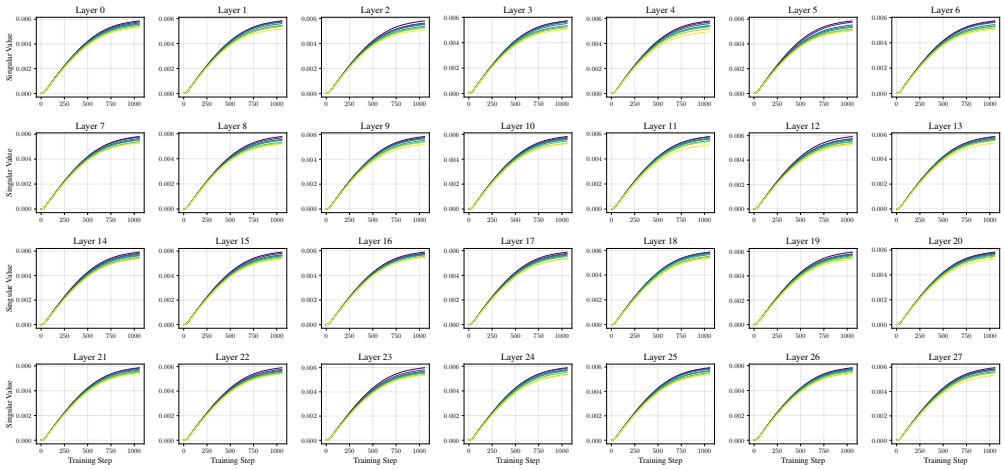



(b) $V$ matrix

Figure 40: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Dion on the OpenHermes 2.5 dataset.




(a) $Q$ matrix




(b) $V$ matrix

Figure 41: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained using Dion on the UltraChat dataset.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The paper formally proves the claims made in the abstract and the introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the authors' work are explicitly discussed in Section 7.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The main text includes formal definitions, assumptions, and proof sketches. Complete proofs for theoretical results are provided in Sections A to D.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The information needed to reproduce the experimental results is disclosed in Section E.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: All experiments use publicly available datasets, models, and codebases. The setup directly follows these public implementations. Section E specifies the model–dataset combinations, optimizer choices, LoRA configurations, and hyperparameters needed to reproduce the reported results.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The experimental details are specified in Section E.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: This work is primarily theoretical; the experiments motivate the theoretical analyses and aid reader comprehension. Therefore, measures of statistical significance for empirical results are not reported. Given the theoretical nature of main contributions, such an omission does not detract from core findings of this work.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the computer resources in Section E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors of this work have reviewed the NeurIPS Code of Ethics. This work conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is primarily theoretical. The main focus is on the mathematical aspect without a direct relation to downstream applications.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve releasing any new data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets, optimizers, and pretrained models, all of which are properly cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: An LLM is used only for editing purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.