Uniform Spectral Growth under Factor-wise Muon Orthogonalization in Matrix Factorization and LoRA

Changmin Kang* CMKANG8128@KAIST.AC.KR
KAIST AI, Seoul, South Korea

Jihun Yun* JIHUNY@KRAFTON.COM
KRAFTON, Seoul, South Korea

Baekrok Shin BR.SHIN@KAIST.AC.KR
KAIST AI, Seoul, South Korea

Yeseul Cho CYSEUL@KAIST.AC.KR
KAIST AI, Seoul, South Korea

Chulhee Yun CHULHEE.YUN@KAIST.AC.KR
KAIST AI, Seoul, South Korea

Abstract

Spectral gradient descent (SpecGD) orthogonalizes matrix parameter updates and has inspired practical optimizers such as Muon. They often perform well in large language model training, but their dynamics remain poorly understood, especially in factorized parameterizations where the product matrix does not receive orthogonalized updates. We study such dynamics through matrix factorization (MF), where the orthogonalization is applied separately to the factor updates. We analyze spectral gradient flow (SpecGF)—a continuous-time analog of SpecGD—in the low-rank MF setting and prove "equal-rate" dynamics: all singular values grow at equal rates up to small deviations. Consequently, smaller singular values attain their target values earlier than larger ones, contrasting with the largest-first stepwise learning observed in standard gradient flow. Moreover, SpecGF converges to global minima from almost all initializations, provided the factor norms remain bounded; with ℓ_2 regularization, we obtain global convergence. Empirically, we observe that LoRA fine-tuning with orthogonalization-based optimizers including Muon exhibit near-uniform growth in the product of LoRA adapters, consistent with the mechanism predicted by our MF analysis.

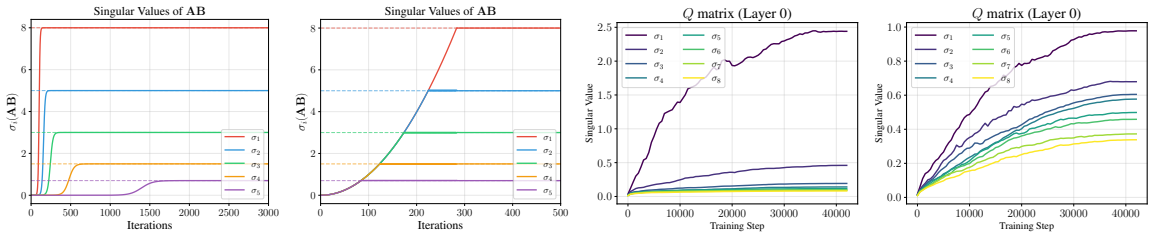
1. Introduction

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 $\mathbf{W} = \mathbf{A}\mathbf{B}$, 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?

*. Equal contribution.

© C. Kang*, J. Yun*, B. Shin, Y. Cho & C. Yun.



(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).

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. We then use Low-Rank Adaptation (LoRA) [24] fine-tuning as a practical testbed for the same factor-wise orthogonalization mechanism. We summarize our contributions as follows:

- We show that the singular values of $\mathbf{AB}$ evolve at a uniform rate under SpecGF with LoRA-style initialization even though SpecGF orthogonalizes the gradients of the factors $\mathbf{A}$ and $\mathbf{B}$ (Theorem 1). The dynamics make smaller singular values converge before larger ones, contrasting with largest-first stepwise learning dynamics from standard gradient flow (GF) [3, 18]. This difference is visualized in Figures 1a and 1b. We then prove that if SpecGF converges, it almost surely converges to global minima provided that the factor norms remain bounded (Theorem 3); with ℓ_2 regularization, SpecGF globally converges (Theorem 4).
- We examine whether the same 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. We further show that the uniform growth phenomenon persists across other orthogonalization-based optimizers (Section 5).

2. Theoretical Setup

We study the product-level dynamics through the MF objective

$$\min_{\mathbf{A}\in\mathbb{R}^{m\times r},\mathbf{B}\in\mathbb{R}^{r\times n}} \left\{\mathcal{L}(\mathbf{A},\mathbf{B}) := \frac{1}{2}\|\mathbf{AB}-\mathbf{Y}\|_{\mathrm{F}}^2\right\}, \tag{1}$$

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 $\mathcal{L}(t) := \mathcal{L}(\mathbf{A}(t),\mathbf{B}(t))$. 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 [22, 24]. We analyze the continuous-time analog of SpecGD, *spectral gradient flow (SpecGF)*:

$$\dot{\mathbf{A}}(t) = -\mathcal{T}\left((\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})\mathbf{B}(t)^\top\right), \quad \dot{\mathbf{B}}(t) = -\mathcal{T}\left(\mathbf{A}(t)^\top(\mathbf{A}(t)\mathbf{B}(t)-\mathbf{Y})\right), \tag{2}$$

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

$$\mathcal{T}(\mathbf{M}) \coloneqq \left((\mathbf{M}\mathbf{M}^\top)^\dagger\right)^{1/2}\mathbf{M} = \mathbf{U}_\mathbf{M}\mathbf{V}_\mathbf{M}^\top, \quad \mathcal{T}(\mathbf{0}) \coloneqq \mathbf{0}. \tag{3}$$

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 $\beta > 0$:

$$\mathcal{T}_\beta(\mathbf{M}) \coloneqq \left(\mathbf{M}\mathbf{M}^\top + \beta\mathbf{I}_m\right)^{-1/2}\mathbf{M}. \tag{4}$$

Adding $\beta\mathbf{I}$ makes the inverse square root well-defined, and makes SpecGF analytic everywhere; see Corollary 19 for details. 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.

3. Uniform Growth of Singular Values

We analyze how the singular values of $\mathbf{W}(t) = \mathbf{A}(t)\mathbf{B}(t)$ evolve under SpecGF with $\mathcal{T}_\beta$, starting from LoRA-style initialization. A key finding is "equal-rate" dynamics, where all active singular values grow at the same rate, implying that the smallest singular value reaches its target first. We begin by establishing near-diagonal alignment:

$$\mathbf{A}(0) = \mathbf{0}, \quad \mathbf{B}(0) = \gamma\mathbf{N} \quad \text{with} \quad \mathbf{N}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{5}$$

where $\mathcal{N}(0,1)$ is a standard Gaussian and γ is sufficiently small. In a coordinate system aligned with $\mathbf{Y}$, the small scale of $\mathbf{B}(0)$ induces a near-diagonal product whose diagonal entries approximate the singular values of $\mathbf{A}(t)\mathbf{B}(t)$, ultimately leading to near-uniform growth.

We focus on the case $r = r^\star$ for simplicity; see Section C for the case $r < r^\star$. Assume $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$ is fixed with $\sigma_1 > \cdots > \sigma_r > 0$. Let $\mathbf{V}_\perp \in \mathbb{R}^{n\times(n-r)}$ denote the orthogonal complement of $\mathbf{V}_r$, so that $[\mathbf{V}_r \mid \mathbf{V}_\perp]$ is orthonormal. We define the *core variables*:

$$\left(\mathbf{X}(t), \mathbf{Z}(t), \mathbf{Z}_\perp(t)\right) \coloneqq \left(\mathbf{U}_r^\top\mathbf{A}(t), \mathbf{B}(t)\mathbf{V}_r, \mathbf{B}(t)\mathbf{V}_\perp\right).$$

The core product is $\mathbf{G}(t) \coloneqq \mathbf{X}(t)\mathbf{Z}(t)$, and under $\mathbf{A}(0) = \mathbf{0}$, we have $\mathbf{A}(t) = \mathbf{U}_r\mathbf{X}(t)$ for all $t \geq 0$. We denote the diagonal entries of $\mathbf{G}(t)$ by $d_i(t) \coloneqq [\mathbf{G}(t)]_{ii}$ for all $i \in [r]$. Further, we let $d_{\min}(t) \coloneqq \min_{i\in[r]} d_i(t)$. For a square matrix $\mathbf{M}$, we write $\mathrm{Off}(\mathbf{M}) \coloneqq \mathbf{M} \odot (\mathbf{1}\mathbf{1}^\top - \mathbf{I})$ where $\odot$ denotes the Hadamard (element-wise) product.

Under the reparametrization, we introduce two tolerances: (i) the *alignment tolerance* $\delta > 0$; and (ii) the *target tolerance* $\varepsilon > 0$.

(i) **Alignment tolerance $\delta > 0$:** We define δ-alignment by

$$\|\mathrm{Off}(\mathbf{G}(t))\|_\mathrm{F} \leq \delta \text{ and } \|\mathbf{X}(t)\mathbf{Z}_\perp(t)\|_\mathrm{F} \leq \delta, \tag{6}$$

i.e., the core product $\mathbf{G}(t)$ is near-diagonal and the residual $\mathbf{X}(t)\mathbf{Z}_\perp(t)$ is near-zero.

(ii) **Target tolerance $\varepsilon > 0$:** We define $d_i(t) \coloneqq [\mathbf{G}(t)]_{ii}$, $e_i(t) \coloneqq d_i(t) - \sigma_i$. The active set is

$$\mathcal{I}_\varepsilon(t) \coloneqq \{i \in [r] : e_i(t) \leq -\varepsilon\} = \{i \in [r] : d_i(t) \leq \sigma_i - \varepsilon\}. \tag{7}$$

We call each index $i \in [r]$ a *mode*, corresponding to the i-th singular value of $\mathbf{AB}$. A mode i is *active* at time t if $i \in \mathcal{I}_\varepsilon(t)$. We show that under δ-alignment, the diagonal entries d_i can approximate the singular values of $\mathbf{AB}$ (Lemma 8). However, under LoRA-style initialization, early gradients are close to zero, thus $\mathcal{T}_\beta$ operates in its near-zero regime, where the update is dominated by β-smoothing. Hence, we introduce a time $\tau > 0$ such that the diagonal entries become positive enough and every mode is active at τ (Lemma 9). We then define termination time $T := \inf\{t \geq \tau : \mathcal{I}_\varepsilon(t) = \emptyset\}$ and show that, with high probability, T is finite and the growth rates of active modes are nearly uniform.

Theorem 1 (Uniform growth) *For a target tolerance such that $\varepsilon \ll \sigma_r$, consider SpecGF with $\mathcal{T}_\beta$ under LoRA-style initialization in Equation (5) with $\beta = O(\varepsilon^3)$ and $\gamma = \Theta(\varepsilon^{1/2})$. Then, with probability at least $1 - e^{-cr^2} - e^{-crn} - e^{-cr}$ with an absolute constant $c > 0$, the following holds:*

(i) ***Finite termination time:*** $T \leq \tau + 2\sqrt{\sigma_1}$.

(ii) ***Approximate unit speed:*** *For all $t \in [\tau, T]$ and all $i \in \mathcal{I}_\varepsilon(t)$,* $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} = 1 + O(\varepsilon^{1/4})$.

(iii) ***Uniform growth:*** *For all $t \in [\tau, T]$ and all $i, j \in \mathcal{I}_\varepsilon(t)$,* $\left|\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} - \frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_j(t)}\right| = O(\varepsilon^{1/4})$.

See Section B for the proof. Since all modes start near zero and σ_r requires the least growth, the smallest singular value reaches its target first. Our analysis naturally extends to the underparameterized setting $r < r^\star$; see Section C. The proofs largely follow those of Theorem 1.

4. Convergence Analysis of SpecGF

Section 3 establishes near-uniform spectral growth, but this alone does not guarantee convergence as it controls only a finite phase of product-level growth, while the individual factor norms may diverge. Hence, we study convergence of SpecGF with $\mathcal{T}_\beta$. 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 $\sqrt{r}$. Thus, the dynamics have bounded speed, and solutions do not blow up in finite time and globally exist. We then show that SpecGF with $\mathcal{T}_\beta$ either converges or diverges to infinity in norm using the Łojasiewicz inequality; see Section D.2 for the proof.

Proposition 2 *Under SpecGF with $\mathcal{T}_\beta$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(t) \leq -\frac{1}{2r}\|\nabla\mathcal{L}(t)\|_{\mathrm{F}}\|(\dot{\mathbf{A}}(t), \dot{\mathbf{B}}(t))\|_{\mathrm{F}}. \tag{8}$$

As a consequence of Equation (8), $(\mathbf{A}(t), \mathbf{B}(t))$ 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 (1) has strict-saddle geometry: every critical point of $\mathcal{L}$ is either a global minimum or a strict saddle point [4, 28]. Moreover, every global minimum corresponds to the best rank-r approximation of $\mathbf{Y}$. As the set of initial points where SpecGF with $\mathcal{T}_\beta$ converges to a strict saddle point has Lebesgue measure zero (Theorem 30), Proposition 2 implies the following theorem.

Theorem 3 *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 standard 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 time-invariant in general; see Proposition 27—thus whether SpecGF globally converges to a critical point is *inconclusive*. We conjecture that there are no polynomial invariants in terms of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ for SpecGF; if there were, then it should translate to the scalar case, $m = n = 1$. However, any such invariant is identically constant; see Propositions 25 and 26 for evidence in the scalar case.

Global Convergence via Regularization. Theorem 3 can be strengthened to state **global** convergence by adding an ℓ_2 regularizer. Consider the regularized objective with $\lambda > 0$:

$$\mathcal{L}_\lambda(\mathbf{A}, \mathbf{B}) \triangleq \frac{1}{2}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{B}\|_{\mathrm{F}}^2. \tag{9}$$

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 λ [43]. The strict-saddle geometry and the avoidance of strict saddle points carry over. Hence, regularized SpecGF with $\mathcal{T}_\beta$ cannot diverge to infinity and almost surely converges to global minima; see Section D.6 for details.

Theorem 4 *For almost all initial values $\mathbf{A}(0)$ and $\mathbf{B}(0)$, SpecGF with $\mathcal{T}_\beta$ under the regularized loss in Equation (9) converges to a global minimum.*

5. Uniform Growth in Practical LoRA Fine-Tuning

We present empirical evidence that near-uniform growth is consistently observed under LoRA fine-tuning with various orthogonalization-based optimizers, including 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 $r \ll \min\{m, n\}$. Hence, applying these optimizers to LoRA factors yields the same factor-wise orthogonalized update structure studied in MF. We observe that near-uniform growth persists across a range of practical settings. Figure 2 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. We defer the experimental details for Figures 1 and 2 and the ablation results to Section E.

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 [25, 50]. 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 6.

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, ℓ_2 regularization induces shrinkage of the product spectrum, suppressing components below the regularization scale λ [43]. 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 [44], uniform growth can be beneficial relative to largest-first dynamics. Although this discussion is stated for MF, the shared product parameterization suggests that a similar optimization–generalization trade-off may arise in LLM fine-tuning with LoRA.

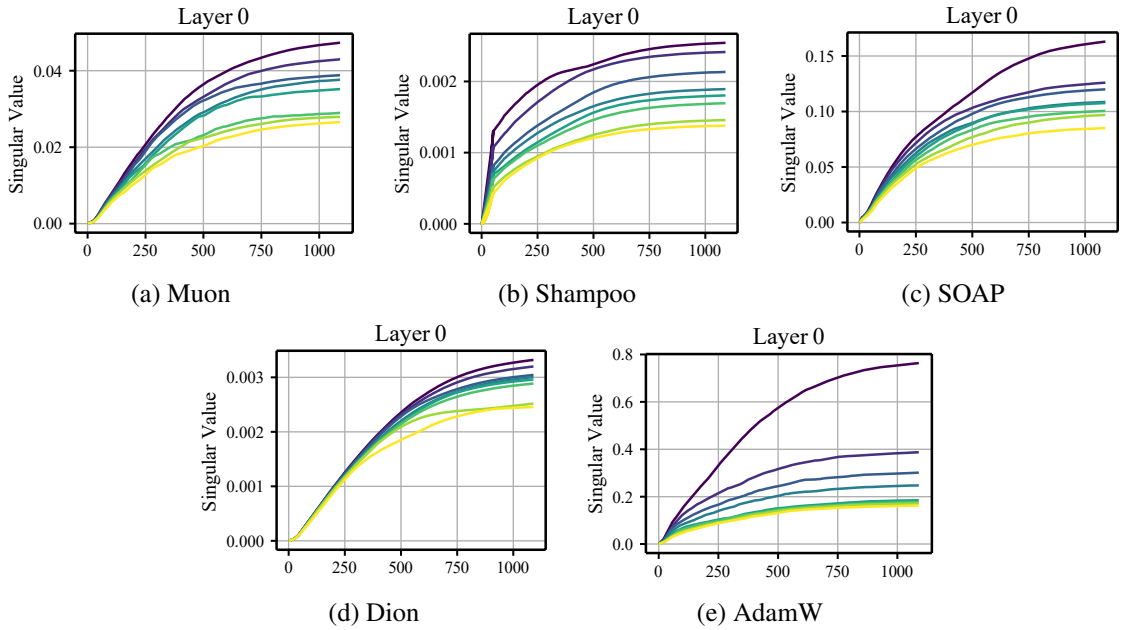


Figure 2: Persistence of uniform growth in LoRA fine-tuning under orthogonalization-based optimizers; AdamW may preferentially grow the leading singular values. The trajectories correspond to singular values of the query projection matrix in the first self-attention layer of LLaMA-3.2-3B.

6. Conclusion

We analyzed spectral gradient flow (SpecGF) for low-rank matrix factorization (MF) and proved that singular values of the product $\mathbf{AB}$ exhibit uniform growth, leading to "equal-rate" dynamics that contrasts with the largest-first behavior of standard GF. We also established that bounded SpecGF trajectories converge almost surely to global minima; with ℓ_2 regularization, global convergence is then guaranteed. We validated our theory via experiments on MF. We also verified that near-uniform growth is widely observed across various orthogonalization-based optimizers and initialization schemes in LoRA fine-tuning. A key limitation is that global convergence without regularization remains open due to the absence of known invariants for SpecGF. Extending our continuous-time analysis to discrete-time SpecGD and more general network architectures is important future work.

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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/.

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

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

[30] 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.

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

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

[33] 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.

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

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

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

Contents

1 Introduction **1**

2 Theoretical Setup **2**

3 Uniform Growth of Singular Values **3**

4 Convergence Analysis of SpecGF **4**

5 Uniform Growth in Practical LoRA Fine-Tuning **5**

6 Conclusion **6**

A Related Work **12**

B Uniform Growth of Singular Values **13**
B.1 Setup and Notation . . . 13
B.2 Alignment from Small Initialization . . . 14
B.3 Initial Growth and Non-degeneracy . . . 15
B.4 Near-Diagonal Factors . . . 16
B.5 Square-Root Dynamics . . . 18
B.6 Non-degeneracy Persistence . . . 20
B.7 Main Theorem: Uniform Growth . . . 20

C Extension to Underparameterized Case ($r < r^\star$) **21**
C.1 Setup and Extended Core Variables . . . 21
C.2 Column Space Invariance . . . 22
C.3 Invariant Manifold . . . 22
C.4 Lipschitz Stability near the Invariant Manifold . . . 22
C.5 Reduction to the Square Case . . . 23
C.6 Extended Uniform Growth Theorem . . . 23

D Convergence to Global Minima of SpecGF **24**
D.1 Analyticity of SpecGF . . . 24
D.2 Convergence Analysis of SpecGF . . . 25
D.3 Saddle Point Avoidance . . . 32
D.4 Convergence Rate of SpecGF . . . 35
D.5 Stability Analysis of SpecGF . . . 39
D.6 SpecGF with ℓ_2 Regularization . . . 42

E Experiments **44**
E.1 Plots for Empirical Validation . . . 44
E.2 Matrix Factorization Experiments Under Various Settings . . . 46
E.3 LLM Experiments . . . 48

Appendix A. 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 [27] and its variants, such as Dion [2] and MuonBP [29] that enhanced the efficiency, as well as preconditioned optimizers like Shampoo [21] and SOAP [45]. Muon normalizes the singular values of the matrix-valued updates, leading to faster convergence and higher performance [33, 47] compared to Adam [31].

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. [14] study implicit bias and max-margin behavior under spectral descent, while Vasudeva et al. [44] analyze generalization benefits in imbalanced classification.

Training Dynamics. Matrix factorization (MF) is a fundamental and widely studied problem in optimization [3, 10, 17, 20, 26, 30, 38, 40] and machine learning [13, 36, 39, 43]. 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, 18], aligning with broader notions of spectral bias [8, 49]. Gradient-based methods can also exhibit saddle-to-saddle trajectories in MF problems [25].

Orthogonalization-based optimizers such as Muon or SpecGD induce different dynamics. Vasudeva et al. [44] 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 a better generalization of SpecGD compared to gradient descent under data imbalance. However, their analysis relies on spectral initialization [51] 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. [37] 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 (LLM). This approach injects trainable low-rank factors into weight updates while keeping the base model fixed [24], motivated by the observation that fine-tuning updates often lie in a low intrinsic-dimensional subspace [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, 48]. However, theoretical guarantees for convergence and training dynamics in this combined setting remain largely unexplored.

Appendix B. Uniform Growth of Singular Values

This section proves the uniform growth results stated in Theorem 1.

B.1. Setup and Notation

β-regularized orthogonalization. Fix $\beta > 0$. For $\mathbf{G} \in \mathbb{R}^{m\times r}$ define

$$\mathcal{T}_{\mathrm{col},\beta}(\mathbf{G}) \coloneqq \mathbf{G}(\mathbf{G}^\top\mathbf{G} + \beta\mathbf{I})^{-1/2},$$

and for $\mathbf{G} \in \mathbb{R}^{r\times n}$ define

$$\mathcal{T}_{\mathrm{row},\beta}(\mathbf{G}) \coloneqq (\mathbf{G}\mathbf{G}^\top + \beta\mathbf{I})^{-1/2}\mathbf{G}.$$

The case $\beta = 0$ (unregularized SpecGF) can be handled by a limiting argument; we focus on $\beta > 0$ throughout.

Lemma 5 (Operator norm bound) *For all $\beta > 0$ and all $\mathbf{G}$,*

$$\|\mathcal{T}_{\mathrm{col},\beta}(\mathbf{G})\|_2 \le 1, \qquad \|\mathcal{T}_{\mathrm{row},\beta}(\mathbf{G})\|_2 \le 1.$$

Proof Let s be any singular value of $\mathbf{G}$. Then s becomes $s/\sqrt{s^2+\beta} \le 1$ under either normalization, hence the operator norm is ≤ 1. ■

Lemma 6 (Invariance) *If $\mathbf{A}(0) = \mathbf{0}$, then $\mathbf{A}(t) \in \mathrm{col}(\mathbf{U}_r)$ for all $t \ge 0$. Consequently, $\mathbf{A}(t) = \mathbf{U}_r\mathbf{X}(t)$ for some $\mathbf{X}(t) \in \mathbb{R}^{r\times r}$.*

Proof The gradient with respect to $\mathbf{A}$ is $\nabla_{\mathbf{A}}\mathcal{L} = (\mathbf{A}\mathbf{B} - \mathbf{Y})\mathbf{B}^\top$. At $t = 0$, we have $\nabla_{\mathbf{A}}\mathcal{L}(0) = -\mathbf{Y}\mathbf{B}(0)^\top = -\mathbf{U}_r\mathbf{\Sigma}\mathbf{V}_r^\top\mathbf{B}(0)^\top$, which lies in $\mathrm{col}(\mathbf{U}_r)$. Since $\mathcal{T}_\beta$ (and $\mathcal{T}$) preserves the column space, $\dot{\mathbf{A}}(0) \in \mathrm{col}(\mathbf{U}_r)$.

By induction, suppose $\mathbf{A}(t) \in \mathrm{col}(\mathbf{U}_r)$, i.e., $\mathbf{A}(t) = \mathbf{U}_r\mathbf{X}(t)$ for some $\mathbf{X}(t) \in \mathbb{R}^{r\times r}$. Then

$$\nabla_{\mathbf{A}}\mathcal{L}(t) = (\mathbf{U}_r\mathbf{X}(t)\mathbf{B}(t) - \mathbf{U}_r\mathbf{\Sigma}\mathbf{V}_r^\top)\mathbf{B}(t)^\top = \mathbf{U}_r(\mathbf{X}(t)\mathbf{B}(t)\mathbf{B}(t)^\top - \mathbf{\Sigma}\mathbf{V}_r^\top\mathbf{B}(t)^\top).$$

This again lies in $\mathrm{col}(\mathbf{U}_r)$, so $\dot{\mathbf{A}}(t) = -\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(t)) \in \mathrm{col}(\mathbf{U}_r)$. By continuity of the flow, $\mathbf{A}(t) \in \mathrm{col}(\mathbf{U}_r)$ for all $t \ge 0$. ■

Core variables. Let $\mathbf{Y} = \mathbf{U}_r\mathbf{\Sigma}\mathbf{V}_r^\top$ be the compact SVD with $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \dots, \sigma_r)$ and $\sigma_1 > \cdots > \sigma_r > 0$, where $\mathbf{U}_r \in \mathbb{R}^{m\times r}$ and $\mathbf{V}_r \in \mathbb{R}^{n\times r}$. Let $\mathbf{V}_\perp \in \mathbb{R}^{n\times(n-r)}$ denote the orthogonal complement of $\mathbf{V}_r$. We define the core variables:

$$\mathbf{X}(t) \coloneqq \mathbf{U}_r^\top\mathbf{A}(t) \in \mathbb{R}^{r\times r}, \quad \mathbf{Z}(t) \coloneqq \mathbf{B}(t)\mathbf{V}_r \in \mathbb{R}^{r\times r}, \quad \mathbf{Z}_\perp(t) \coloneqq \mathbf{B}(t)\mathbf{V}_\perp \in \mathbb{R}^{r\times(n-r)}.$$

The core product is $\mathbf{G}(t) \coloneqq \mathbf{X}(t)\mathbf{Z}(t)$. For a square matrix $\mathbf{M}$, we write $\mathrm{Off}(\mathbf{M}) \coloneqq \mathbf{M} - \mathrm{Diag}(\mathrm{diag}(\mathbf{M}))$ for its off-diagonal part.

We introduce two tolerances:

- **Alignment tolerance $\delta > 0$:** Time t is δ-aligned if $\|\mathrm{Off}(\mathbf{G}(t))\|_{\mathrm{F}} \le \delta$ and $\|\mathbf{X}(t)\mathbf{Z}_\perp(t)\|_{\mathrm{F}} \le \delta$.
- **Target tolerance $\varepsilon > 0$:** We denote $d_i(t) \coloneqq [\mathbf{G}(t)]_{ii}$, $e_i(t) \coloneqq d_i(t) - \sigma_i$, and the active set $\mathcal{I}_\varepsilon(t) \coloneqq \{i \in [r] : e_i(t) \le -\varepsilon\}$.

A mode i is *active* if $i \in \mathcal{I}_\varepsilon(t)$. Let $T \coloneqq \inf\{t \ge \tau : \mathcal{I}_\varepsilon(t) = \emptyset\}$ be the termination time.

B.2. Alignment from Small Initialization

Proposition 7 (High-probability alignment) *Under the Gaussian initialization* $\mathbf{A}(0) = \mathbf{0}$ *and* $\mathbf{B}(0) = \gamma\mathbf{N}$*, there exist deterministic constants* $C_1, C_2 < \infty$ *such that, with probability at least* $1 - \exp(-c_1 r^2) - \exp(-c_1 rn)$ *for some constant* $c_1 > 0$*,*

$$\sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathrm{Off}(\mathbf{G}(t))\|_{\mathrm{F}} \le C_1\gamma, \qquad \sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathbf{X}(t)\mathbf{Z}_\perp(t)\|_{\mathrm{F}} \le C_2\gamma.$$

Consequently, the trajectory is $O(\gamma)$*-aligned for all* $t \le \tau + 2\sqrt{\sigma_1}$*.*

Proof Both bounds follow from Lipschitz stability near the diagonal manifold $\mathcal{M} := \{(\mathbf{X}, \mathbf{Z}, \mathbf{Z}_\perp) : \mathbf{X}, \mathbf{Z} \text{ diagonal}, \mathbf{Z}_\perp = 0\}$. Let $\mathbf{S}(t) := (\mathbf{X}(t), \mathbf{Z}(t), \mathbf{Z}_\perp(t))$ and consider the orthogonal projection onto $\mathcal{M}$:

$$\Pi(\mathbf{X}, \mathbf{Z}, \mathbf{Z}_\perp) := \big(\mathrm{Diag}(\mathrm{diag}(\mathbf{X})), \mathrm{Diag}(\mathrm{diag}(\mathbf{Z})), \mathbf{0}\big).$$

Let $\mathbf{S}^\star(t) := (\mathbf{X}^\star(t), \mathbf{Z}^\star(t), \mathbf{Z}^\star_\perp(t))$ be the trajectory of SpecGF with $\mathcal{T}_\beta$ trajectory, started at $\mathbf{S}^\star(0) := \Pi(\mathbf{S}(0))$. By diagonal manifold invariance, $\mathbf{S}^\star(t) \in \mathcal{M}$ for all t, hence $\mathbf{Z}^\star_\perp(t) \equiv \mathbf{0}$ and, writing $\mathbf{G}^\star(t) := \mathbf{X}^\star(t)\mathbf{Z}^\star(t)$, $\mathrm{Off}(\mathbf{G}^\star(t)) \equiv \mathbf{0}$.

By speed bounds and Gaussian concentration, both trajectories remain in a compact set $\mathcal{K}$ for $t \in [0, \tau + 2\sqrt{\sigma_1}]$. Since SpecGF with $\mathcal{T}_\beta$ defines a $\mathcal{C}^1$ vector field, finite-horizon Lipschitz stability gives

$$\|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_{\mathrm{F}} \le e^{L(\tau + 2\sqrt{\sigma_1})} \|\mathbf{S}(0) - \mathbf{S}^\star(0)\|_{\mathrm{F}}$$

for some finite $L = L(\mathcal{K})$.

Note that $\|\mathbf{S}(0) - \mathbf{S}^\star(0)\|_{\mathrm{F}} \le \|\mathrm{Off}(\mathbf{Z}(0))\|_{\mathrm{F}} + \|\mathbf{Z}_\perp(0)\|_{\mathrm{F}}$ because $\mathbf{X}(0) = \mathbf{0}$. Since $\mathbf{Z}(0) = \gamma\mathbf{N}\mathbf{V}_r$ and $\mathbf{Z}_\perp(0) = \gamma\mathbf{N}\mathbf{V}_\perp$ where $\mathbf{N}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and $\mathbf{V}_r$, $\mathbf{V}_\perp$ have orthonormal columns, the entries of $\mathbf{N}\mathbf{V}_r$ and $\mathbf{N}\mathbf{V}_\perp$ are also i.i.d. $\mathcal{N}(0,1)$. Thus $\|\mathrm{Off}(\mathbf{N}\mathbf{V}_r)\|_{\mathrm{F}}^2 \sim \chi^2_{r(r-1)}$ and $\|\mathbf{N}\mathbf{V}_\perp\|_{\mathrm{F}}^2 \sim \chi^2_{r(n-r)}$. By chi-squared concentration,

$$\mathbb{P}\big(\|\mathrm{Off}(\mathbf{Z}(0))\|_{\mathrm{F}} + \|\mathbf{Z}_\perp(0)\|_{\mathrm{F}} > C\gamma(r + \sqrt{rn})\big) \le \exp(-c_1 r^2) + \exp(-c_1 rn)$$

for some absolute constants $C, c_1 > 0$. Consequently, $\sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_{\mathrm{F}} \le \kappa\gamma$ for some finite $\kappa > 0$ with probability at least $1 - \exp(-c_1 r^2) - \exp(-c_1 rn)$.

Since $\mathbf{Z}^\star_\perp(t) \equiv \mathbf{0}$, it follows that

$$\|\mathbf{Z}_\perp(t)\|_{\mathrm{F}} = \|\mathbf{Z}_\perp(t) - \mathbf{Z}^\star_\perp(t)\|_{\mathrm{F}} \le \|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_{\mathrm{F}} \le \kappa\gamma.$$

By speed bounds $\|\mathbf{X}(t)\|_2 \le t \le \tau + 2\sqrt{\sigma_1}$, hence $\sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathbf{X}(t)\mathbf{Z}_\perp(t)\|_{\mathrm{F}} \le (\tau + 2\sqrt{\sigma_1})\kappa\gamma := C_2\gamma$.

For the off-diagonal bound, using $\mathbf{G} = \mathbf{X}\mathbf{Z}$ and $\mathbf{G}^\star = \mathbf{X}^\star\mathbf{Z}^\star$, we have

$$\mathbf{G} - \mathbf{G}^\star = \mathbf{X}\mathbf{Z} - \mathbf{X}^\star\mathbf{Z}^\star = (\mathbf{X} - \mathbf{X}^\star)\mathbf{Z} + \mathbf{X}^\star(\mathbf{Z} - \mathbf{Z}^\star).$$

Since both trajectories remain in the compact set $\mathcal{K}$ for $t \in [0, \tau + 2\sqrt{\sigma_1}]$, there exists $M = M(\mathcal{K}) < \infty$ such that $\|\mathbf{Z}(t)\|_2, \|\mathbf{X}^\star(t)\|_2 \le M$. Therefore, we have

$$\begin{aligned}
\|\mathbf{G}(t) - \mathbf{G}^\star(t)\|_{\mathrm{F}} &\le \|\mathbf{X}(t) - \mathbf{X}^\star(t)\|_{\mathrm{F}} \|\mathbf{Z}(t)\|_2 + \|\mathbf{X}^\star(t)\|_2 \|\mathbf{Z}(t) - \mathbf{Z}^\star(t)\|_{\mathrm{F}} \\
&\le M\big(\|\mathbf{X}(t) - \mathbf{X}^\star(t)\|_{\mathrm{F}} + \|\mathbf{Z}(t) - \mathbf{Z}^\star(t)\|_{\mathrm{F}}\big) \\
&\le M\sqrt{2}\|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_{\mathrm{F}}.
\end{aligned}$$

Since $\mathrm{Off}(\mathbf{G}^\star(t)) \equiv \mathbf{0}$, we obtain

$$\|\mathrm{Off}(\mathbf{G}(t))\|_{\mathrm{F}} = \|\mathrm{Off}(\mathbf{G}(t) - \mathbf{G}^\star(t))\|_{\mathrm{F}} \le \|\mathbf{G}(t) - \mathbf{G}^\star(t)\|_{\mathrm{F}} \le M\sqrt{2}\|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_{\mathrm{F}}.$$

Combining these bounds yields $\sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathrm{Off}(\mathbf{G}(t))\|_{\mathrm{F}} \le C_1\gamma$. ■

Lemma 8 (Diagonal approximation) *Suppose $\|\mathrm{Off}(\mathbf{G}(t))\|_{\mathrm{F}} \le \delta$ and $\|\mathbf{X}(t)\mathbf{Z}_\perp(t)\|_{\mathrm{F}} \le \delta$. Then there exists a permutation $\pi(t)$ such that*

$$|\sigma_{\pi_i(t)}(\mathbf{A}(t)\mathbf{B}(t)) - d_i(t)| = O(\delta)$$

for all $i \in [r]$, where $\pi_i(t)$ denotes the index satisfying $d_i(t) \approx \sigma_{\pi_i(t)}(\mathbf{A}(t)\mathbf{B}(t))$.

Proof We have

$$\mathbf{AB} = \mathbf{U}_r\mathbf{XZV}_r^\top + \mathbf{U}_r\mathbf{XZ}_\perp\mathbf{V}_\perp^\top.$$

Since $\mathbf{U}_r, \mathbf{V}_r, \mathbf{V}_\perp$ have orthonormal columns, the singular values of $\mathbf{AB}$ equal those of $[\mathbf{G} \mid \mathbf{XZ}_\perp]$. We write $\mathbf{D} = \mathrm{Diag}(d_1, \dots, d_r)$. The singular values of $[\mathbf{D} \mid \mathbf{0}]$ are exactly $\{|d_1|, \dots, |d_r|\}$, and

$$\|[\mathbf{G} \mid \mathbf{XZ}_\perp] - [\mathbf{D} \mid \mathbf{0}]\|_F \le \|\mathrm{Off}(\mathbf{G})\|_F + \|\mathbf{XZ}_\perp\|_F \le 2\delta.$$

By Weyl's inequality (or Mirsky), there exists a permutation π of $[r]$ such that

$$|\sigma_{\pi(i)}([\mathbf{G} \mid \mathbf{XZ}_\perp]) - |d_i|| \le 2\delta \quad \text{for all } i \in [r].$$

Since $d_i > 0$ on our trajectory, we conclude $|\sigma_{\pi(i)}(\mathbf{AB}) - d_i| = O(\delta)$. ■

B.3. Initial Growth and Non-degeneracy

Lemma 9 (Initial growth) *Let $\mathbf{Z}(0) = \gamma\mathbf{M}$ where $\mathbf{M} := \mathbf{NV}_r \in \mathbb{R}^{r\times r}$, with $\mathbf{N}$ having i.i.d. $\mathcal{N}(0,1)$ entries and $\beta = O(\gamma^2)$. Write $\mathbf{M} = [\mathbf{c}_1 \mid \cdots \mid \mathbf{c}_r]$ where $\mathbf{c}_i \in \mathbb{R}^r$ denotes the i-th column. Define the high-probability event*

$$\mathcal{E} := \left\{ \|\mathbf{M}\|_2 \le C_3\sqrt{r}, \quad \min_{i\in[r]} \|\mathbf{c}_i\|^2 \ge c_1 r \right\}.$$

Then $\mathbb{P}(\mathcal{E}) \ge 1 - e^{-c_1 r}$ for some absolute constants $C_3, c_1 > 0$. Let $\tau \asymp \frac{\sigma_r}{\sigma_1\sqrt{r}}$. On the event $\mathcal{E}$,

$$d_{\min}(\tau) \ge c_2\tau\gamma$$

for some constant $c_2 > 0$. Moreover, all modes are active at τ: $\mathcal{I}_\varepsilon(\tau) = [r]$, equivalently $d_i(\tau) < \sigma_i - \varepsilon$ for all i.

Proof For an $r \times r$ Gaussian matrix $\mathbf{M}$, standard concentration gives $\|\mathbf{M}\|_2 \le 3\sqrt{r}$ with probability at least $1 - 2e^{-r/2}$. Each column $\mathbf{c}_i \sim \mathcal{N}(0, \mathbf{I}_r)$ satisfies $\|\mathbf{c}_i\|^2 \sim \chi^2_r$, so $\mathbb{P}(\|\mathbf{c}_i\|^2 < \frac{r}{2}) \le e^{-c_1 r}$ for some absolute constant $c_1 > 0$. By a union bound, $\mathbb{P}(\mathcal{E}) \ge 1 - e^{-c_1 r}$.

We now bound $[\dot{\mathbf{X}}(0)\mathbf{Z}(0)]_{ii}$ on $\mathcal{E}$; note that $\dot{\mathbf{Z}}(0) = \mathbf{0}$. From the dynamics of SpecGF with $\mathcal{T}_\beta$ at $t = 0$ with $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{Z}(0) = \gamma\mathbf{M}$:

$$\dot{\mathbf{X}}(0) = \mathcal{T}_{\mathrm{col},\beta}(\gamma\mathbf{\Sigma}\mathbf{M}^\top) = \gamma\mathbf{\Sigma}\mathbf{M}^\top(\gamma^2\mathbf{M}\mathbf{\Sigma}^2\mathbf{M}^\top + \beta\mathbf{I})^{-1/2}.$$

Define the positive definite matrix $\mathbf{Q} := \gamma^2\mathbf{M}\mathbf{\Sigma}^2\mathbf{M}^\top + \beta\mathbf{I}$. Then

$$\dot{\mathbf{X}}(0)\mathbf{Z}(0) = \gamma^2\mathbf{\Sigma}\mathbf{M}^\top\mathbf{Q}^{-1/2}\mathbf{M},$$

whose i-th diagonal entry equals $\gamma^2\sigma_i \cdot \mathbf{c}_i^\top\mathbf{Q}^{-1/2}\mathbf{c}_i$. Since $\mathbf{Q}^{-1/2}$ is positive definite with minimum eigenvalue $\lambda_{\max}(\mathbf{Q})^{-1/2}$,

$$\mathbf{c}_i^\top\mathbf{Q}^{-1/2}\mathbf{c}_i \ge \frac{\|\mathbf{c}_i\|^2}{\sqrt{\lambda_{\max}(\mathbf{Q})}}.$$

On $\mathcal{E}$, we have $\lambda_{\max}(\mathbf{Q}) \le \gamma^2\sigma_1^2\|\mathbf{M}\|_2^2 + \beta \le C_3^2\gamma^2\sigma_1^2 r + \beta$. Assuming $\beta \le C\gamma^2\sigma_1^2 r$ for some absolute constant $C > 0$, this gives $\lambda_{\max}(\mathbf{Q}) \le C'\gamma^2\sigma_1^2 r$, and hence

$$\sqrt{\lambda_{\max}(\mathbf{Q})} \le \sqrt{C'} \cdot \gamma\sigma_1\sqrt{r}.$$

Combined with $\|\mathbf{c}_i\|^2 \ge c_1 r$, we obtain

$$[\dot{\mathbf{X}}(0)\mathbf{Z}(0)]_{ii} \ge \gamma^2\sigma_r \cdot \frac{c_1 r}{\sqrt{C'} \cdot \gamma\sigma_1\sqrt{r}} = \frac{c_1\gamma\sigma_r\sqrt{r}}{\sqrt{C'}\sigma_1}.$$

Finally, we apply Taylor expansion. For $\beta > 0$, the vector field is C^1, so for sufficiently small τ:

$$\mathbf{X}(\tau) = \tau\dot{\mathbf{X}}(0) + O(\tau^2), \qquad \mathbf{Z}(\tau) = \mathbf{Z}(0) + O(\tau^2),$$

where the second equality uses $\dot{\mathbf{Z}}(0) = \mathbf{0}$ (because $\mathbf{X}(0) = \mathbf{0}$). Expanding $\mathbf{G}(\tau) = \mathbf{X}(\tau)\mathbf{Z}(\tau)$,

$$d_{\min}(\tau) \ge \tau \cdot \min_i [\dot{\mathbf{X}}(0)\mathbf{Z}(0)]_{ii} - O(\tau^2\gamma) \asymp \tau\gamma.$$

For the second claim, note that $d_i(\tau) \asymp \tau\gamma \ll \sigma_r - \varepsilon$ for sufficiently small γ. Hence, it holds that $e_i(\tau) = d_i(\tau) - \sigma_i < -\varepsilon$ for all $i \in [r]$, so all modes are active (i.e., early phase of training). ■

B.4. Near-Diagonal Factors

Lemma 10 (Near-diagonal factors) *Under the hypotheses of Proposition 7, it follows that*

$$\sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathrm{Off}(\mathbf{X}(t))\|_{\mathrm{F}} \le C_4\gamma, \qquad \sup_{t \le \tau + 2\sqrt{\sigma_1}} \|\mathrm{Off}(\mathbf{Z}(t))\|_{\mathrm{F}} \le C_5\gamma.$$

Moreover, on any interval where $d_{\min}(t) \ge d_0 > 0$, the diagonal entries satisfy

$$|[\mathbf{X}(t)]_{ii} - \widetilde{s}_i(t)| \le C_6\Big(\gamma + \frac{\gamma}{\sqrt{d_0}}\Big), \qquad |[\mathbf{Z}(t)]_{ii} - \widetilde{s}_i(t)| \le C_6\Big(\gamma + \frac{\gamma}{\sqrt{d_0}}\Big),$$

where $\widetilde{s}_i(t) := \mathrm{sgn}(\xi_i)\sqrt{d_i(t)}$ is the signed square root with $\xi_i \sim \mathcal{N}(0,1)$, κ is the Lipschitz constant from Proposition 7, and C_6 depends only on κ.

Proof By Lipschitz stability (Proposition 7), $\|\mathbf{X}(t)-\mathbf{X}^\star(t)\|_F \le \kappa\gamma$ where $\mathbf{X}^\star(t) \in \mathcal{M}$ is diagonal. Since $\mathrm{Off}(\mathbf{X}^\star) = \mathbf{0}$,

$$\|\mathrm{Off}(\mathbf{X})\|_F = \|\mathrm{Off}(\mathbf{X}-\mathbf{X}^\star)\|_F \le \|\mathbf{X}-\mathbf{X}^\star\|_F \le \kappa\gamma.$$

The same bound holds for $\mathbf{Z}$ similarly. For the diagonal entries, we let

$$x_i := [\mathbf{X}]_{ii}, \quad z_i := [\mathbf{Z}]_{ii}, \quad x_i^\star := [\mathbf{X}^\star]_{ii}, \quad z_i^\star := [\mathbf{Z}^\star]_{ii}.$$

The reference trajectory starts from $x_i^\star(0) = 0$ and $z_i^\star(0) = \gamma\xi_i$. Since $\xi_i \sim \mathcal{N}(0,1)$, Gaussian concentration gives $|\xi_i| = O(1)$ with high probability, so it holds that $|x_i^\star(0) - z_i^\star(0)| = O(\gamma)$.

On diagonal manifold $\mathcal{M}$, the SpecGF dynamics for mode i are

$$\dot{x}_i^\star = -\frac{e_i^\star z_i^\star}{\sqrt{(e_i^\star z_i^\star)^2+\beta}}, \qquad \dot{z}_i^\star = -\frac{e_i^\star x_i^\star}{\sqrt{(e_i^\star x_i^\star)^2+\beta}},$$

where $e_i^\star(t) = x_i^\star(t)z_i^\star(t) - \sigma_i$. Define the strictly increasing function $g(t) := \frac{t}{\sqrt{t^2+\beta}}$. The imbalance $\delta_i^\star := x_i^\star - z_i^\star$ satisfies

$$\dot{\delta}_i^\star = g(e_i^\star x_i^\star) - g(e_i^\star z_i^\star).$$

Since $e_i^\star x_i^\star - e_i^\star z_i^\star = e_i^\star \delta_i^\star$, for $e_i^\star < 0$, the sign of $(e_i^\star x_i^\star) - (e_i^\star z_i^\star)$ is opposite to that of $\delta_i^\star$. By the monotonicity of g, this implies $\delta_i^\star \cdot \dot{\delta}_i^\star \le 0$, hence $|\delta_i^\star(t)|$ is non-increasing.

We now show that $e_i^\star(t) \le 0$ for all $t \ge 0$. At $t = 0$, we have $e_i^\star(0) = x_i^\star(0)z_i^\star(0) - \sigma_i = -\sigma_i < 0$. If $e_i^\star(t_0) = 0$ for some $t_0 > 0$, then $\dot{x}_i^\star(t_0) = \dot{z}_i^\star(t_0) = 0$. By ODE uniqueness (due to locally Lipschitz), $(x_i^\star, z_i^\star)$ remains constant for $t \ge t_0$, so $e_i^\star(t) = 0$ for all $t \ge t_0$. Thus, $e_i^\star(t)$ cannot become positive, and $e_i^\star(t) \le 0$ for all t.

Combining the above, $\delta_i^\star \cdot \dot{\delta}_i^\star \le 0$ holds for all t, and thus

$$|x_i^\star(t) - z_i^\star(t)| \le |x_i^\star(0) - z_i^\star(0)| = O(\gamma) \quad \text{for all } t.$$

Now, we define

$$\widetilde{s}_i^\star := \mathrm{sgn}(\xi_i)\sqrt{d_i^\star}, \qquad \widetilde{s}_i := \mathrm{sgn}(\xi_i)\sqrt{d_i}.$$

We claim that $x_i^\star, z_i^\star, \widetilde{s}_i^\star$ all share the same sign as ξ_i. To see this, note that $z_i^\star(0) = \gamma\xi_i$ has the sign of ξ_i, while $x_i^\star(0) = 0$. For $e_i^\star \le 0$, the dynamics give

$$\dot{x}_i^\star = -\frac{e_i^\star z_i^\star}{\sqrt{(e_i^\star z_i^\star)^2+\beta}},$$

which has the same sign as $z_i^\star$. Thus, $x_i^\star$ immediately enters the same sign as $z_i^\star$ and cannot cross zero thereafter (since $\dot{x}_i^\star$ would push it back). Similarly, $\dot{z}_i^\star$ has the same sign as $x_i^\star$, so $z_i^\star$ also maintains its sign.

Since $\|\mathbf{G} - \mathbf{G}^\star\|_F \le \kappa\gamma$, we have $|d_i - d_i^\star| \le \kappa\gamma$. On intervals where $d_{\min}(t) \ge d_0$,

$$|\widetilde{s}_i - \widetilde{s}_i^\star| = |\sqrt{d_i} - \sqrt{d_i^\star}| = \frac{|d_i - d_i^\star|}{\sqrt{d_i} + \sqrt{d_i^\star}} \le \frac{\kappa\gamma}{\sqrt{d_i}} \le \frac{\kappa\gamma}{\sqrt{d_0}}.$$

Since $x_i^\star$ and $\widetilde{s}_i^\star$ have the same sign, we obtain

$$|x_i^\star - \widetilde{s}_i^\star| = \Big||x_i^\star| - \sqrt{|x_i^\star||z_i^\star|}\Big| \le |x_i^\star - z_i^\star| = O(\gamma).$$

Combining with $|x_i - x_i^\star| \le \|\mathbf{X} - \mathbf{X}^\star\|_F \le \kappa\gamma$, the triangle inequality gives

$$|x_i - \widetilde{s}_i| \le |x_i - x_i^\star| + |x_i^\star - \widetilde{s}_i^\star| + |\widetilde{s}_i^\star - \widetilde{s}_i| \le C_6\Big(\gamma + \frac{\gamma}{\sqrt{d_0}}\Big),$$

where C_6 depends only on κ. The same argument gives $|z_i - \widetilde{s}_i| \le C_6(\gamma + \gamma/\sqrt{d_0})$. ■

B.5. Square-Root Dynamics

Lemma 11 (Square-root dynamics) *Suppose δ-alignment holds with $\delta = O(\gamma)$, $\gamma = \Theta(\varepsilon^{1/2})$, and $\beta = O(\varepsilon^3)$. Let $[\tau', T']$ be any subinterval of $[\tau, \tau + 2\sqrt{\sigma_1}]$ on which*

$$d_{\min}(t) := \min_{i\in[r]} d_i(t) \ge d_0 = \Theta(\gamma).$$

Then for each active mode i (i.e., $e_i(t) \le -\varepsilon$), the square-root coordinate $s_i(t) := \sqrt{d_i(t)}$ satisfies, for a.e. $t \in [\tau', T']$,

$$\dot{s}_i(t) = -\mathrm{sgn}(e_i(t))\,\alpha_i(t) + R_i(t), \qquad \alpha_i(t) := \frac{|e_i(t)|\,s_i(t)}{\sqrt{e_i(t)^2 s_i(t)^2 + \beta}} \in (0, 1]. \tag{10}$$

Moreover, the remainder obeys

$$|R_i(t)| \le C_7\varepsilon^{1/4}, \tag{11}$$

where $C_7 = C_7(\mathbf{Y}, r)$ depends only on constants from Lemma 10 (and earlier constants).

In particular, for every active mode i (i.e., $e_i(t) \le -\varepsilon$),

$$\dot{s}_i(t) = 1 + O(\varepsilon^{1/4}).$$

Proof Since $d_i(t) = [\mathbf{X}(t)\mathbf{Z}(t)]_{ii}$ and $s_i(t) = \sqrt{d_i(t)}$, the chain rule gives, for a.e. t with $d_i(t) > 0$,

$$\dot{d}_i(t) = [\dot{\mathbf{X}}(t)\mathbf{Z}(t)]_{ii} + [\mathbf{X}(t)\dot{\mathbf{Z}}(t)]_{ii}, \qquad \dot{s}_i(t) = \frac{\dot{d}_i(t)}{2s_i(t)}. \tag{12}$$

Let $\mathbf{E}(t) := \mathbf{G}(t) - \mathbf{\Sigma}$ and $\mathbf{G_X} := \mathbf{E}\mathbf{Z}^\top + \mathbf{X}\mathbf{Z}_\perp\mathbf{Z}_\perp^\top = \mathbf{U}_r^\top\nabla_{\mathbf{A}}\mathcal{L}$. By Lemma 10, $\mathbf{X}$ and $\mathbf{Z}$ are nearly diagonal:

$$\mathbf{X} = \mathrm{Diag}(x_i) + \mathbf{\Delta}_X, \qquad \mathbf{Z} = \mathrm{Diag}(z_i) + \mathbf{\Delta}_Z,$$

where $x_i, z_i = \mathrm{sgn}(\xi_i)s_i + O(\gamma + \gamma/\sqrt{d_0})$ and $\|\mathbf{\Delta}_X\|_F, \|\mathbf{\Delta}_Z\|_F = O(\gamma)$. The SpecGF dynamics give

$$\dot{\mathbf{X}} = -\mathcal{T}_{\mathrm{col},\beta}(\mathbf{G_X}), \qquad \dot{\mathbf{Z}} = -\mathcal{T}_{\mathrm{row},\beta}(\nabla_{\mathbf{B}}\mathcal{L})\,\mathbf{V}_r.$$

Since $\|\mathbf{X}\mathbf{Z}_\perp\|_F = O(\delta)$ (by δ-alignment), we have $\mathbf{G}_\mathbf{X} = \mathbf{E}\mathbf{Z}^\top + O(\delta)$. Combined with $\mathbf{E} = \mathrm{Diag}(e_i) + O(\delta)$ and $z_i = \mathrm{sgn}(\xi_i)s_i + O(\gamma)$, this yields

$$\mathbf{G}_\mathbf{X} = \mathrm{Diag}(e_i z_i) + O(\delta+\gamma), \qquad \mathbf{G}_\mathbf{X}^\top \mathbf{G}_\mathbf{X} = \mathrm{Diag}(e_i^2 s_i^2) + O(\delta+\gamma).$$

Since $\mathcal{T}_{\mathrm{col},\beta}(\mathbf{G}) = \mathbf{G}\,(\mathbf{G}^\top\mathbf{G} + \beta\mathbf{I})^{-1/2}$, we obtain

$$[\dot{\mathbf{X}}]_{ii} = -\frac{e_i z_i}{\sqrt{e_i^2 s_i^2 + \beta}} + O(\delta+\gamma) = -\frac{e_i\,\mathrm{sgn}(\xi_i)s_i}{\sqrt{e_i^2 s_i^2 + \beta}} + O(\delta+\gamma). \tag{13}$$

For $\dot{\mathbf{Z}}$, write $\nabla_\mathbf{B}\mathcal{L} = \mathbf{X}^\top\mathbf{E}\mathbf{V}_r^\top + \mathbf{X}^\top\mathbf{X}\mathbf{Z}_\perp\mathbf{V}_\perp^\top$. Using $\mathbf{V}_r^\top\mathbf{V}_r = \mathbf{I}$ and $\mathbf{V}_\perp^\top\mathbf{V}_r = \mathbf{0}$, we have $\nabla_\mathbf{B}\mathcal{L}\,\mathbf{V}_r = \mathbf{X}^\top\mathbf{E}$, so

$$\dot{\mathbf{Z}} = \dot{\mathbf{B}}\mathbf{V}_r = -\mathcal{T}_{\mathrm{row},\beta}(\nabla_\mathbf{B}\mathcal{L})\,\mathbf{V}_r = -\big((\mathbf{X}^\top\mathbf{E})(\mathbf{X}^\top\mathbf{E})^\top + \beta\mathbf{I}\big)^{-1/2}\,\mathbf{X}^\top\mathbf{E}.$$

Moreover, $(\mathbf{X}^\top\mathbf{E})(\mathbf{X}^\top\mathbf{E})^\top = \mathbf{X}^\top\mathbf{E}\mathbf{E}^\top\mathbf{X} = \mathrm{Diag}(s_i^2 e_i^2) + O(\delta+\gamma)$, so

$$[\dot{\mathbf{Z}}]_{ii} = -\frac{x_i e_i}{\sqrt{x_i^2 e_i^2 + \beta}} + O(\delta+\gamma) = -\frac{\mathrm{sgn}(\xi_i)s_i\,e_i}{\sqrt{e_i^2 s_i^2 + \beta}} + O(\delta+\gamma). \tag{14}$$

Using Lemma 10 and (13)–(14), we compute the diagonal entries of $\dot{\mathbf{X}}\mathbf{Z}$ and $\mathbf{X}\dot{\mathbf{Z}}$:

$$[\dot{\mathbf{X}}\mathbf{Z}]_{ii} = [\dot{\mathbf{X}}]_{ii}\,[\mathbf{Z}]_{ii} + O(\|\dot{\mathbf{X}}\|_F\|\mathrm{Off}(\mathbf{Z})\|_F) = -\frac{e_i s_i^2}{\sqrt{e_i^2 s_i^2 + \beta}} + O\Big(\delta+\gamma+\frac{\gamma}{\sqrt{d_0}}\Big), \tag{15}$$

$$[\mathbf{X}\dot{\mathbf{Z}}]_{ii} = [\mathbf{X}]_{ii}\,[\dot{\mathbf{Z}}]_{ii} + O(\|\mathbf{X}\|_F\|\mathrm{Off}(\dot{\mathbf{Z}})\|_F) = -\frac{e_i s_i^2}{\sqrt{e_i^2 s_i^2 + \beta}} + O\Big(\delta+\gamma+\frac{\gamma}{\sqrt{d_0}}\Big). \tag{16}$$

Summing (15)–(16) and using (12),

$$\dot{d}_i = -\frac{2e_i s_i^2}{\sqrt{e_i^2 s_i^2 + \beta}} + O\Big(\delta+\gamma+\frac{\gamma}{\sqrt{d_0}}\Big), \qquad \dot{s}_i = -\frac{e_i s_i}{\sqrt{e_i^2 s_i^2 + \beta}} + O\Big(\delta+\frac{\gamma}{\sqrt{d_0}}\Big),$$

where we used $s_i \geq \sqrt{d_0}$ on $[\tau', T']$ to absorb the $O(\gamma)$ contribution into $O(\gamma/\sqrt{d_0})$ after dividing by $2s_i$. Under the scaling $\gamma = \Theta(\varepsilon^{1/2})$, $d_0 = \Theta(\gamma)$, and $\delta = O(\gamma)$, the dominant error term is $\gamma/\sqrt{d_0} = O(\varepsilon^{1/4})$.

Defining $\alpha_i(t) := \frac{|e_i(t)|\,s_i(t)}{\sqrt{e_i(t)^2 s_i(t)^2+\beta}} \in (0,1]$ and $R_i(t) := \dot{s}_i(t) + \mathrm{sgn}(e_i(t))\,\alpha_i(t)$, we see that (10) holds and (11) follows.

Finally, for active modes $e_i(t) \leq -\varepsilon$, we have $\mathrm{sgn}(e_i) = -1$ and

$$\alpha_i = \frac{|e_i|\,s_i}{\sqrt{e_i^2 s_i^2 + \beta}} = \Big(1 + \frac{\beta}{e_i^2 s_i^2}\Big)^{-1/2} \geq 1 - \frac{\beta}{2\varepsilon^2 d_0} = 1 - O(\varepsilon^{1/2}),$$

using $\beta = O(\varepsilon^3)$ and $d_0 = \Theta(\varepsilon^{1/2})$. Since $\alpha_i \leq 1$ trivially and $|R_i| = O(\varepsilon^{1/4})$, we conclude $\dot{s}_i = 1 + O(\varepsilon^{1/4})$ for active modes. ■

B.6. Non-degeneracy Persistence

Lemma 12 (Non-degeneracy persistence) *Suppose $d_{\min}(\tau) \geq d_0 = \Theta(\gamma)$ with $\gamma = \Theta(\varepsilon^{1/2})$ and $\beta = O(\varepsilon^3)$. Let $\underline{d} := \min\{d_0, \sigma_r - \varepsilon\}$. On the high-probability alignment event in Proposition 7, for all $t \in [\tau, \min\{T, \tau + 2\sqrt{\sigma_1}\}]$,*

$$d_{\min}(t) \geq \underline{d} > 0.$$

Proof Note that we can possibly decrease $d_0 = \Theta(\tau\gamma)$ so that $d_0 \leq \sigma_r - \epsilon$; hence, without loss of generality, we may assume $\underline{d} = d_0$. We define the exit time $T_{\text{exit}} := \inf\{t > \tau : d_{\min}(t) < \underline{d}\}$. Let $H := \tau + 2\sqrt{\sigma_1}$ for simplicity.

We show that $T_{\text{exit}} > \min\{T, H\}$, which implies $d_{\min}(t) \geq \underline{d}$ for all $t \in [\tau, \min\{T, H\}]$. On $[\tau, T_{\text{exit}})$, we have $d_{\min}(t) \geq \underline{d} > 0$ by definition, so Lemma 11 applies. For any active mode i (i.e., $e_i(t) \leq -\varepsilon$), we have $\dot{s}_i(t) = 1 + O(\varepsilon^{1/4}) > 0$ for small ε. Hence, $s_i(t)$ is strictly increasing, and so is $d_i(t) = s_i(t)^2$.

For contradiction, suppose that some mode i reaches $d_i(t) = \underline{d}$ from above. Then $e_i(t) = d_i(t) - \sigma_i \leq \underline{d} - \sigma_r < -\varepsilon$ (since $\underline{d} \leq \sigma_r - \varepsilon$), so i is active and $\dot{d}_i(t) > 0$. This contradicts the decrease after reaching $\underline{d}$. By continuity, $d_{\min}(t)$ cannot reach $\underline{d}$ on $[\tau, \min\{T, H\}]$. It follows that $T_{\text{exit}} > \min\{T, H\}$, and hence $d_{\min}(t) \geq \underline{d}$ for all $t \in [\tau, \min\{T, H\}]$. ■

B.7. Main Theorem: Uniform Growth

Theorem 13 (Uniform growth) *For a target tolerance such that $\varepsilon \ll \sigma_r$, consider SpecGF with $\mathcal{T}_\beta$ under LoRA-style initialization (Equation (5)) where $\beta = O(\varepsilon^3)$ and $\gamma = \Theta(\varepsilon^{1/2})$. Then, the following holds with probability at least $1 - e^{-cr^2} - e^{-crn} - e^{-cr}$ with an absolute constant $c > 0$:*

(i) ***Finite termination time:*** $T \leq \tau + 2\sqrt{\sigma_1}$.

(ii) ***Approximate unit speed:*** *For all $t \in [\tau, T]$ and all active modes $i \in \mathcal{I}_\varepsilon(t)$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} = 1 + O(\varepsilon^{1/4}).$$

(iii) ***Uniform growth:*** *For all $t \in [\tau, T]$ and any two active modes $i, j \in \mathcal{I}_\varepsilon(t)$,*

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} - \frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_j(t)}\right| = O(\varepsilon^{1/4}).$$

Proof Let us work on the intersection of the alignment event in Proposition 7 on the deterministic interval $[0, \tau + 2\sqrt{\sigma_1}]$ and the initial-growth event in Lemma 9. By a union bound, this event has probability at least $1 - \exp(-cr^2) - \exp(-crn) - \exp(-cr)$.

On this event, Lemma 12 gives $d_{\min}(t) \geq \underline{d} > 0$ for all $t \in [\tau, \min\{T, \tau + 2\sqrt{\sigma_1}\}]$. Hence, Lemma 11 applies on this interval. In particular, for every active mode $i \in \mathcal{I}_\varepsilon(t)$, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}s_i(t) = 1 + O(\varepsilon^{1/4}), \qquad s_i(t) := \sqrt{d_i(t)}.$$

Taking ε sufficiently small, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}s_i(t) \geq \frac{1}{2}$$

whenever $i \in \mathcal{I}_\varepsilon(t)$ and $t \in [\tau, \min\{T, \tau + 2\sqrt{\sigma_1}\}]$, especially at the boundary $d_i = \sigma_i - \varepsilon$. Thus, a mode cannot cross from $d_i \geq \sigma_i - \varepsilon$ back into the active region. Since all modes are active at time τ by Lemma 9, and since a mode cannot re-enter the active region after reaching the threshold, any mode active at $\tau + 2\sqrt{\sigma_1}$ must have remained active throughout $[\tau, \tau + 2\sqrt{\sigma_1}]$.

We now prove that $T \leq \tau + 2\sqrt{\sigma_1}$. Suppose, for contradiction, that $T > \tau + 2\sqrt{\sigma_1}$. Then at time $\tau + 2\sqrt{\sigma_1}$ there exists an active mode i, so $d_i(\tau + 2\sqrt{\sigma_1}) < \sigma_i - \varepsilon$. Since mode i remains below the threshold whenever it is active, the above lower bound on $\dot{s}_i$ gives

$$s_i(\tau + 2\sqrt{\sigma_1}) \geq s_i(\tau) + \frac{1}{2}(2\sqrt{\sigma_1}) \geq \sqrt{\sigma_1}.$$

Therefore

$$d_i(\tau + 2\sqrt{\sigma_1}) \geq \sigma_1 \geq \sigma_i,$$

which contradicts $d_i(\tau + 2\sqrt{\sigma_1}) < \sigma_i - \varepsilon$. Thus

$$T \leq \tau + 2\sqrt{\sigma_1}.$$

Since $T \leq \tau + 2\sqrt{\sigma_1}$, the estimates above apply on the whole interval $[\tau, T]$. Hence every active mode satisfies

$$\dot{s}_i(t) = 1 + O(\varepsilon^{1/4})$$

for all $t \in [\tau, T]$, proving (ii). For (iii), for any active i, j,

$$|\dot{s}_i(t) - \dot{s}_j(t)| \leq |\dot{s}_i(t) - 1| + |\dot{s}_j(t) - 1| = O(\varepsilon^{1/4}),$$

which concludes the proof. ■

Appendix C. Extension to Underparameterized Case ($r < r^\star$)

We extend the uniform growth analysis to the underparameterized regime where the LoRA rank r is smaller than the target rank $r^\star := \mathrm{rank}(\mathbf{Y})$. The key insight is that the column space invariance and diagonal manifold structure still hold within the relevant subspace.

C.1. Setup and Extended Core Variables

Let $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the compact SVD with $\mathbf{U} \in \mathbb{R}^{m \times r^\star}$, $\mathbf{V} \in \mathbb{R}^{n \times r^\star}$, and $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_{r^\star})$ with $\sigma_1 > \cdots > \sigma_{r^\star} > 0$. The LoRA factorization has rank $r \leq r^\star$: $\mathbf{A} \in \mathbb{R}^{m \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times n}$.

We partition the target singular vectors as $\mathbf{U} = [\mathbf{U}_r \mid \mathbf{U}_{>r}]$ and $\mathbf{V} = [\mathbf{V}_r \mid \mathbf{V}_{>r}]$, where $\mathbf{U}_r, \mathbf{V}_r \in \mathbb{R}^{\cdot \times r}$ correspond to the top r singular values and $\mathbf{U}_{>r}, \mathbf{V}_{>r}$ correspond to the remaining $r^\star - r$ singular values. Let $\mathbf{V}_\perp \in \mathbb{R}^{n \times (n - r^\star)}$ denote the orthogonal complement of $\mathbf{V}$ in $\mathbb{R}^n$.

The extended core variables are:

$$\mathbf{X}_r := \mathbf{U}_r^\top \mathbf{A} \in \mathbb{R}^{r \times r}, \qquad \mathbf{X}_{>r} := \mathbf{U}_{>r}^\top \mathbf{A} \in \mathbb{R}^{(r^\star - r) \times r},$$
$$\mathbf{Z}_r := \mathbf{B}\mathbf{V}_r \in \mathbb{R}^{r \times r}, \qquad \mathbf{Z}_{>r} := \mathbf{B}\mathbf{V}_{>r} \in \mathbb{R}^{r \times (r^\star - r)},$$
$$\mathbf{Z}_\perp := \mathbf{B}\mathbf{V}_\perp \in \mathbb{R}^{r \times (n - r^\star)}.$$

In this setup, the core product is defined by $\mathbf{G}_r \coloneqq \mathbf{X}_r\mathbf{Z}_r \in \mathbb{R}^{r\times r}$.

Similar to δ-alignment, we define *extended δ-alignment* by

$$\|\mathrm{Off}(\mathbf{G}_r(t))\|_{\mathrm{F}} \le \delta, \quad \|\mathbf{X}_{>r}(t)\|_{\mathrm{F}} \le \delta, \quad \|\mathbf{Z}_{>r}(t)\|_{\mathrm{F}} \le \delta, \quad \|\mathbf{X}_r(t)\mathbf{Z}_\perp(t)\|_{\mathrm{F}} \le \delta.$$

This generalizes the original δ-alignment condition to the underparameterized setting by additionally controlling the components outside the top-r subspace. Note that, given a convergence to the best rank-r approximation (cf. Proposition 28 and Theorem 30), for any $\delta > 0$ we can always find some finite $t_0 > 0$ such that extended δ-alignment holds for all $t \ge t_0$.

C.2. Column Space Invariance

Lemma 14 (Column space invariance) *If $\mathbf{A}(0) = \mathbf{0}$, then $\mathbf{A}(t) \in \mathrm{col}(\mathbf{U})$ for all $t \ge 0$. Consequently, $\mathbf{A}(t) = \mathbf{U}_r\mathbf{X}_r(t) + \mathbf{U}_{>r}\mathbf{X}_{>r}(t)$.*

Proof The gradient $\nabla_{\mathbf{A}}\mathcal{L} = (\mathbf{AB} - \mathbf{Y})\mathbf{B}^\top$ lies in $\mathrm{col}(\mathbf{A}) + \mathrm{col}(\mathbf{Y}) \subseteq \mathrm{col}(\mathbf{U})$ whenever $\mathbf{A} \in \mathrm{col}(\mathbf{U})$. At $t = 0$, $\nabla_{\mathbf{A}}\mathcal{L}(0) = -\mathbf{YB}(0)^\top \in \mathrm{col}(\mathbf{U})$. Since $\mathcal{T}_\beta$ preserves column spaces, $\dot{\mathbf{A}}(0) \in \mathrm{col}(\mathbf{U})$. By induction and continuity of the flow, $\mathbf{A}(t) \in \mathrm{col}(\mathbf{U})$ for all $t \ge 0$. ■

C.3. Invariant Manifold

Lemma 15 (Invariant manifold) *Under the initialization $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{B}(0) = \gamma\mathbf{N}$ with $\mathbf{N}$ having i.i.d. $\mathcal{N}(0,1)$ entries, the extended manifold*

$$\mathcal{M}_{\mathrm{ext}} \coloneqq \left\{(\mathbf{X}_r, \mathbf{X}_{>r}, \mathbf{Z}_r, \mathbf{Z}_{>r}, \mathbf{Z}_\perp) : \mathbf{X}_{>r} = \mathbf{0},\ \mathbf{Z}_{>r} = \mathbf{0},\ \mathbf{Z}_\perp = \mathbf{0}\right\}$$

is invariant under the SpecGF dynamics.

Proof Let $\mathbf{\Sigma}_r = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$ and $\mathbf{\Sigma}_{>r} = \mathrm{Diag}(\sigma_{r+1}, \ldots, \sigma_{r^\star})$. On $\mathcal{M}_{\mathrm{ext}}$, we have $\mathbf{A} = \mathbf{U}_r\mathbf{X}_r$ and $\mathbf{B} = \mathbf{Z}_r\mathbf{V}_r^\top$, so

$$\begin{aligned}\mathbf{AB} - \mathbf{Y} &= \mathbf{U}_r\mathbf{X}_r\mathbf{Z}_r\mathbf{V}_r^\top - \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top - \mathbf{U}_{>r}\mathbf{\Sigma}_{>r}\mathbf{V}_{>r}^\top \\ &= \mathbf{U}_r(\mathbf{X}_r\mathbf{Z}_r - \mathbf{\Sigma}_r)\mathbf{V}_r^\top - \mathbf{U}_{>r}\mathbf{\Sigma}_{>r}\mathbf{V}_{>r}^\top.\end{aligned}$$

The gradient $\nabla_{\mathbf{A}}\mathcal{L} = (\mathbf{AB} - \mathbf{Y})\mathbf{B}^\top = \mathbf{U}_r(\mathbf{X}_r\mathbf{Z}_r - \mathbf{\Sigma}_r)\mathbf{Z}_r^\top$ lies entirely in $\mathrm{col}(\mathbf{U}_r)$, implying $\dot{\mathbf{X}}_{>r} = \mathbf{U}_{>r}^\top\dot{\mathbf{A}} = \mathbf{0}$.

Similarly, $\nabla_{\mathbf{B}}\mathcal{L} = \mathbf{X}_r^\top(\mathbf{X}_r\mathbf{Z}_r - \mathbf{\Sigma}_r)\mathbf{V}_r^\top - \mathbf{X}_{>r}^\top\mathbf{\Sigma}_{>r}\mathbf{V}_{>r}^\top$. When $\mathbf{X}_{>r} = \mathbf{0}$, this lies entirely in the row space of $\mathbf{V}_r^\top$, so $\dot{\mathbf{Z}}_{>r} = \dot{\mathbf{B}}\mathbf{V}_{>r} = \mathbf{0}$ and $\dot{\mathbf{Z}}_\perp = \dot{\mathbf{B}}\mathbf{V}_\perp = \mathbf{0}$. ■

C.4. Lipschitz Stability near the Invariant Manifold

Proposition 16 (Alignment in the underparameterized case) *Under the initialization $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{B}(0) = \gamma\mathbf{N}$, there exist constants $C_{>r}, C_\perp < \infty$ such that, with probability at least $1 - e^{-cr(r^\star - r)} - e^{-cr(n - r^\star)}$,*

$$\sup_{t\le\tau+2\sqrt{\sigma_1}} \|\mathbf{X}_{>r}(t)\|_F \le C_{>r}\gamma, \qquad \sup_{t\le\tau+2\sqrt{\sigma_1}} \|\mathbf{Z}_{>r}(t)\|_F \le C_{>r}\gamma, \qquad \sup_{t\le\tau+2\sqrt{\sigma_1}} \|\mathbf{Z}_\perp(t)\|_F \le C_\perp\gamma.$$

Proof Let $\mathbf{S}(t) := (\mathbf{X}_r(t), \mathbf{X}_{>r}(t), \mathbf{Z}_r(t), \mathbf{Z}_{>r}(t), \mathbf{Z}_\perp(t))$ denote the true trajectory. At initialization with $\mathbf{A}(0) = \mathbf{0}$ and $\mathbf{B}(0) = \gamma\mathbf{N}$, the core variables satisfy

$$\mathbf{X}_r(0) = \mathbf{0}, \quad \mathbf{X}_{>r}(0) = \mathbf{0}, \quad \mathbf{Z}_r(0) = \gamma\mathbf{N}\mathbf{V}_r, \quad \mathbf{Z}_{>r}(0) = \gamma\mathbf{N}\mathbf{V}_{>r}, \quad \mathbf{Z}_\perp(0) = \gamma\mathbf{N}\mathbf{V}_\perp.$$

We define the projection onto $\mathcal{M}_{\text{ext}}$ by $\Pi(\mathbf{X}_r, \mathbf{X}_{>r}, \mathbf{Z}_r, \mathbf{Z}_{>r}, \mathbf{Z}_\perp) := (\mathbf{X}_r, \mathbf{0}, \mathbf{Z}_r, \mathbf{0}, \mathbf{0})$, and consider the reference trajectory $\mathbf{S}^\star(t)$ initialized at $\mathbf{S}^\star(0) = \Pi(\mathbf{S}(0)) = (\mathbf{0}, \mathbf{0}, \gamma\mathbf{N}\mathbf{V}_r, \mathbf{0}, \mathbf{0})$. By Lemma 15, $\mathbf{S}^\star(t) \in \mathcal{M}_{\text{ext}}$ for all $t \geq 0$.

The initial discrepancy is

$$\|\mathbf{S}(0) - \mathbf{S}^\star(0)\|_F = \|\mathbf{Z}_{>r}(0)\|_F + \|\mathbf{Z}_\perp(0)\|_F = O(\gamma\sqrt{r(n-r)})$$

with high probability by Gaussian concentration.

The SpecGF vector field with $\mathcal{T}_\beta$ is $\mathcal{C}^1$, so finite-horizon Lipschitz stability gives

$$\|\mathbf{S}(t) - \mathbf{S}^\star(t)\|_F \leq e^{L(\tau+2\sqrt{\sigma_1})}\|\mathbf{S}(0) - \mathbf{S}^\star(0)\|_F \leq \kappa\gamma$$

for some finite $\kappa = \kappa(\mathbf{Y}, r)$. Since $\mathbf{X}^\star_{>r}(t) \equiv \mathbf{0}$, $\mathbf{Z}^\star_{>r}(t) \equiv \mathbf{0}$, and $\mathbf{Z}^\star_\perp(t) \equiv \mathbf{0}$, the bounds naturally hold. ■

C.5. Reduction to the Square Case

On the invariant manifold $\mathcal{M}_{\text{ext}}$, the dynamics reduce to the $r \times r$ system:

$$\dot{\mathbf{X}}_r = -\mathcal{T}_{\text{col},\beta}\big((\mathbf{X}_r\mathbf{Z}_r - \mathbf{\Sigma}_r)\mathbf{Z}_r^\top\big),$$
$$\dot{\mathbf{Z}}_r = -\mathcal{T}_{\text{row},\beta}\big(\mathbf{X}_r^\top(\mathbf{X}_r\mathbf{Z}_r - \mathbf{\Sigma}_r)\big).$$

This is equivalent to the original rank-$r^\star$ dynamics with target $\mathbf{\Sigma}_r = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$.

By Proposition 16, perturbations from $\mathcal{M}_{\text{ext}}$ are controlled: $\|\mathbf{X}_{>r}\|_F, \|\mathbf{Z}_{>r}\|_F, \|\mathbf{Z}_\perp\|_F = O(\gamma)$. These contribute only at most $O(\gamma)$ errors to the effective dynamics of $(\mathbf{X}_r, \mathbf{Z}_r)$.

C.6. Extended Uniform Growth Theorem

Theorem 17 (Uniform growth for $r < r^\star$) *Consider SpecGF with $\mathcal{T}_\beta$ on the rank-r factorization $\mathbf{AB}$ approximating $\mathbf{Y}$ of rank $r^\star \geq r$. Let $\mathbf{G}_r(t) := \mathbf{X}_r(t)\mathbf{Z}_r(t)$ and $d_i(t) := [\mathbf{G}_r(t)]_{ii}$ for $i \in [r]$. Under the same scaling $\gamma = \Theta(\varepsilon^{1/2})$ and $\beta = O(\varepsilon^3)$, the following holds for all $t \in [\tau, T]$ and all active modes $i \in \mathcal{I}_\varepsilon(t) \cap [r]$, with probability at least $1 - e^{-cr}$:*

(i) ***Approximate unit speed:*** $\dfrac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} = 1 + O(\varepsilon^{1/4})$.

(ii) ***Uniform growth:*** *For any two active modes $i, j \in \mathcal{I}_\varepsilon(t) \cap [r]$,*

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_i(t)} - \frac{\mathrm{d}}{\mathrm{d}t}\sqrt{d_j(t)}\right| = O(\varepsilon^{1/4}).$$

Proof The trajectory remains $O(\gamma)$-close to $\mathcal{M}_{\text{ext}}$ by Proposition 16. On this manifold, the $(\mathbf{X}_r, \mathbf{Z}_r)$ dynamics are identical to the full-rank case with target $\mathbf{\Sigma}_r$. Therefore, all lemmas from the full-rank analysis (Lemmas 9 to 12) naturally hold with $r^\star$ replaced by r and target singular values $(\sigma_1, \ldots, \sigma_r)$. The perturbations from $\mathbf{X}_{>r}$, $\mathbf{Z}_{>r}$, $\mathbf{Z}_\perp$ contribute only at most $O(\gamma)$ errors, which are absorbed into the existing $O(\gamma)$ terms. The proof then follows Theorem 13. ■

Appendix 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 18 *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 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 ω. $\sigma(\mathbf{S})$ 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 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 $\mathcal{H}^\infty(S_\omega)$ for any $\omega \in [0, \pi)$, because for $f \in \mathcal{H}^\infty(S_\omega) \setminus \{0\}$,

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

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 real-analytic restricted to the set of real matrices. As the dimension is finite, the fractional domain spaces $\dot{X}_p$ coincides 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 19 *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 18, the fact that the polynomials are analytic, and the composition of analytic functions is again 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 $\mathbf{X}$, $1 \le i \le r \le \min\{m, n\}$. 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 20 (Analyticity of $\mathcal{T}$ and $\mathcal{T}_\beta$) *SpecGF with $\mathcal{T}_\beta$ is analytic everywhere. SpecGF with $\mathcal{T}$ is analytic on the set of full-rank matrices. Moreover, the solution uniquely exists for all $t \ge 0$ where they are analytic.*

Proof Analyticity is straightforward from Corollary 19, 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

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

If the maximal solution does not exist for all $t \ge 0$, then the iterate should leave every compact set of the domain. Therefore, the solution must exist for all $t \ge 0$. Uniqueness follows from the local Lipschitzness in $(\mathbf{A}, \mathbf{G})$ of SpecGF (by Picard-Lindelöf theorem). ■

D.2. Convergence Analysis of SpecGF

We restate the setup of SpecGF.

Lemma 21 *On SpecGF with $\mathcal{T}$, the loss is monotonically non-increasing:*

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

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 $\mathbf{U}'\mathbf{S}'\mathbf{V}'^\top$.

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

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 22 *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 $\sigma_1, \ldots, \sigma_r \geq 0$.

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

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 $\mathcal{L}(\infty) \leq \mathcal{L}(0)$ should exist. Moreover,

$$\|\mathbf{A}(t)\mathbf{B}(t)\|_{\mathrm{F}} \leq \|\mathbf{A}(t)\mathbf{B}(t) - \mathbf{Y}\|_{\mathrm{F}} + \|\mathbf{Y}\|_{\mathrm{F}} = 2\sqrt{\mathcal{L}(t)} + \|\mathbf{Y}\|_{\mathrm{F}} \leq 2\sqrt{\mathcal{L}(0)} + \|\mathbf{Y}\|_{\mathrm{F}}. \quad (17)$$

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

Lemma 23 (Łojasiewicz gradient inequality, [15, 35]) *Let $F : U \to \mathbb{R}$ be a real analytic function on an open set $U \subset \mathbb{R}^d$. Then, for every point $\mathbf{p} \in U$ of F, there exists a (possibly small) neighborhood W of $\mathbf{p}$, constants $C > 0$ and $b \in [0, 1)$ such that for all $\mathbf{x} \in W$,*

$$|F(\mathbf{x}) - F(\mathbf{p})|^b \leq C\|\nabla F(\mathbf{x})\|. \tag{18}$$

In addition, if $\mathbf{p}$ is a critical point of F, then Equation (18) holds for $b \in [\frac{1}{2}, 1)$.

Proposition 24 *Let $\mathbf{x}(t)$ be a differentiable and absolutely continuous curve on $\mathbb{R}^d$ and F be a real analytic function on an open set of $\mathbb{R}^d$. Assume that $F(\mathbf{x}(t))$ is well-defined for all $t \geq 0$ and the condition*

$$\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t} = 0 \quad \Rightarrow \quad \dot{\mathbf{x}}(t) = 0. \tag{19}$$

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

$$\frac{\mathrm{d}F(\mathbf{x}(t))}{\mathrm{d}t} \leq -c\|\nabla F(\mathbf{x}(t))\|\|\dot{\mathbf{x}}(t)\| \tag{20}$$

for all $t \geq 0$, 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 (20), $F(\mathbf{x}(t)) \downarrow F(\bar{\mathbf{x}})$. If there exists a finite time $t' \geq 0$ such that $F(\mathbf{x}(t')) = F(\bar{\mathbf{x}})$, then $F(\mathbf{x}(t)) = F(\bar{\mathbf{x}})$ for $t \geq t'$, hence $\mathbf{x}(t) = \bar{\mathbf{x}}$ by Equation (19). From here we assume that $F(\mathbf{x}(t)) > F(\bar{\mathbf{x}})$ for all t.

From Lemma 23 and Equation (20), we see that

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

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

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

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

$$-\|\dot{\mathbf{x}}(t)\| \geq \frac{C}{c(1-b)}\frac{\mathrm{d}}{\mathrm{d}t}\left(G(t)^{1-b}\right). \tag{21}$$

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 $t_1 \geq 0$ 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

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

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 (21). The above yields a contradiction:

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

Therefore, for all $t \geq t_1$, $\mathbf{x}(t) \in B(r)$. 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 derivative $-\mathcal{T}_\beta(\nabla_{\mathbf{A}} \mathcal{L}(t))$ 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 (19) is satisfied. Moreover, let the rank of $\nabla_{\mathbf{A}} \mathcal{L}(t)$ and $\nabla_{\mathbf{B}} \mathcal{L}(t)$ be r_A, 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:

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

The first inequality follows from Chebyshev's sum inequality. Hence, Equation (20) 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),

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

The last inequality follows from the fact that $\|\mathcal{T}(\mathbf{M})\|_{\mathrm{F}}^2 \leq \mathrm{rank}(\mathbf{M})$ 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, the terminology "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. If there exists an invariant polynomial in $\mathbf{A}$ and $\mathbf{B}$, it should work for the $m = n = 1$ case. However, it can be shown that any polynomial should be identically constant on whole $\mathbb{R} \times \mathbb{R}$. 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 25 (Absence of polynomial invariant 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

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

Plugging $a = 0$ gives $(\partial_a P(0, b))^2 \times b^2\beta = 0$, 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 some polynomials $P'_{(a)}$ and $P'_{(b)}$ 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 $\mathbb{R}[a]$. For a quadratic with a non-perfect square discriminant, it has no roots in $\mathbb{R}[a]$. $\mathbb{R}[a]$ being 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 some polynomials $P''_{(a)}$ and $P''_{(b)}$ 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, with similar argument, there exist some polynomials $P'''_{(a)}$ and $P'''_{(b)}$ 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. ■

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

Proposition 26 (Absence of polynomial invariant 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 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 $\mathrm{sgn}(a) = \mathrm{sgn}(b)$, 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 $\mathrm{sgn}(a) = -\mathrm{sgn}(b)$, 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 $\partial_b P(a,b)$ is 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 27 *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 derivative of $\mathbf{A}$ and $\mathbf{B}$ translate to that 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 derivative of $\mathbf{A}$ and $\mathbf{B}$ translate to that 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

$$L^N(\mathbf{W}_1,\ldots,\mathbf{W}_N) = \frac{1}{2}\|\mathbf{Y}-\mathbf{W}_N\cdots\mathbf{W}_1\mathbf{X}\|_{\mathrm{F}}^2, \tag{22}$$

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 $N = 2$ case.

Proposition 28 ([4], Proposition 6.6, 6.9, 6.11) *Let $r \triangleq \min\{d_0, d_1, d_2\}$. Assume that $\mathbf{X}\mathbf{X}^\top$ has full rank d_0, $\mathbf{Y}$ is of rank q, and $\mathbf{W}_k$ is the global minimum of L^1 restricted to a set of matrices whose rank is at most k. Let K be the rank of $\mathbf{W}_2\mathbf{W}_1$.*

(i) *$\mathbf{W}_k$ is the best rank-k approximation of $\mathbf{Y}$.*

(ii) *If $K = \min\{r, q\}$, every critical point $(\mathbf{W}_1, \mathbf{W}_2)$ that satisfies $\mathbf{W}_2\mathbf{W}_1 = \mathbf{W}_K$ is the global minimum of L^2. If $\mathbf{W}_2\mathbf{W}_1 \neq \mathbf{W}_K$, then $(\mathbf{W}_1, \mathbf{W}_2)$ is a strict saddle point.*

(iii) *Any other critical point $(\mathbf{W}_1, \mathbf{W}_2)$, including the origin, 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 28 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 (22) 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. [9] that claims a similar result.

Proposition 29 ([9], 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 30 *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 29 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 Jacobian matrix of f_β, Df_β:

$$\frac{\mathrm{d}}{\mathrm{d}t} D\varphi_t(\mathbf{x}_0) = Df_\beta\left(\varphi_t(\mathbf{x}_0)\right) D\varphi_t(\mathbf{x}_0), \quad D\varphi_0(\mathbf{x}_0) = \mathbf{I}.$$

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_β makes $|D\varphi_t(\mathbf{x}_0)|$ nonzero for all $\mathbf{x}_0$ and all $t \geq 0$, hence the conditions meet 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

$$D\mathcal{T}_\beta(0)[\mathbf{\Delta}] = \beta^{-1/2}\mathbf{\Delta}, \tag{23}$$

for all 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 $c > 0$,

$$|s(x+c) - s(x)| = |s'(x+rc)|c \leq \frac{c}{2\beta^{\frac{3}{2}}} \tag{24}$$

for some $r \in (0,1)$. Applying Equation (24) to the eigenvalues of $\mathbf{\Delta}^\top\mathbf{\Delta}$ yields

$$\begin{aligned}\left\|\frac{\mathcal{T}_\beta(h\mathbf{\Delta})}{h} - \beta^{-\frac{1}{2}}\mathbf{\Delta}\right\|_{\mathrm{F}} &= \left\|\mathbf{\Delta}\left(h^2\mathbf{\Delta}^\top\mathbf{\Delta} + \beta\mathbf{I}\right)^{-\frac{1}{2}} - \beta^{-\frac{1}{2}}\mathbf{\Delta}\right\|_{\mathrm{F}} \\ &\leq \|\mathbf{\Delta}\|_{\mathrm{F}}\left\|\left(h^2\mathbf{\Delta}^\top\mathbf{\Delta} + \beta\mathbf{I}\right)^{-\frac{1}{2}} - \beta^{-\frac{1}{2}}\mathbf{I}\right\|_2 \\ &\leq \|\mathbf{\Delta}\|_{\mathrm{F}}\frac{h^2\|\mathbf{\Delta}\|_2^2}{2\beta^{\frac{3}{2}}}.\end{aligned}$$

Equation (23) directly yields that

$$Df_\beta(\bar{\mathbf{x}}) = -\beta^{-1/2}\nabla^2\mathcal{L}(\bar{\mathbf{x}}) \tag{25}$$

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 (25), $Df_\beta\left(\bar{\mathbf{x}}\right)$ 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 24; SpecGF iterate will get close to some critical point of $\mathcal{L}$ after a finite time, where the Łojasiewicz gradient inequality holds. On 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 hold: 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 31 *For all 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 $t \geq T$,

$$\begin{aligned}
\frac{d}{\mathrm{d}t}\mathcal{L}(t) &= \langle \nabla_{\mathbf{A}}\mathcal{L}(t), -\mathcal{T}_\beta\left(\nabla_{\mathbf{A}}\mathcal{L}(t)\right)\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}(t), -\mathcal{T}_\beta\left(\nabla_{\mathbf{B}}\mathcal{L}(t)\right)\rangle \\
&\leq -\frac{1}{\sqrt{2\beta}}\left(\|\nabla_{\mathbf{A}}\mathcal{L}(t)\|_{\mathrm{F}}^2 + \|\nabla_{\mathbf{B}}\mathcal{L}(t)\|_{\mathrm{F}}^2\right) = -\frac{1}{\sqrt{2\beta}}\|\nabla\mathcal{L}(t)\|_{\mathrm{F}}^2 \\
&\leq -\frac{1}{C^2\sqrt{2\beta}}(\mathcal{L}(t) - \mathcal{L}(\bar{\mathbf{A}}, \bar{\mathbf{B}}))^{2b}.
\end{aligned}$$

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 $\mathcal{E}$ differs by b. In our case, we established the almost sure convergence towards global minima (under assuming convergence), hence assume $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is the global minimum. Then, $\mathcal{E}$ is the Morse-Bott function, and thus the (optimal) exponent b is exactly $\frac{1}{2}$ [15]. 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 the Morse-Bott function.

Assumption 32 *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 a $m \times m$ orthogonal matrix; similarly, $[\mathbf{V}_r|\mathbf{V}_\perp]$ forms a $n \times n$ orthogonal matrix. Note that zero singular values may appear in $\mathbf{\Sigma}_\perp$, but never appear in $\mathbf{\Sigma}_r$.

The solution set

$$S \triangleq \left\{\mathbf{M} \in \mathbb{R}^{m\times r}, \mathbf{N} \in \mathbb{R}^{r\times n} : \mathbf{M}\mathbf{N} = \bar{\mathbf{Y}}_r\right\}$$

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 $\mathbf{Q}\mathbf{N} = \mathbf{L}_{\bar{\mathbf{A}}}\bar{\mathbf{Y}}_r = \bar{\mathbf{B}}$; 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 Φ 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 $T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S = \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ [15, Definition 1.5]; Other conditions for being a Morse-Bott function are automatically satisfied [15, 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, thus a Fredholm operator index 0, hence the range condition is automatically satisfied.

Let $\mathbf{R}(\mathbf{A}, \mathbf{B}) \triangleq \mathbf{A}\mathbf{B} - \bar{\mathbf{Y}}_r$ 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 a 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 $T_{(\bar{\mathbf{A}},\bar{\mathbf{B}})}S \subset \ker\nabla^2\mathcal{E}(\bar{\mathbf{A}},\bar{\mathbf{B}})$ [15, 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,

$$\Delta\mathbf{A}_1 = \mathbf{0} \Rightarrow \bar{\mathbf{A}}^\top(\Delta\mathbf{A}_0\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_0) + \Delta\mathbf{A}_0^\top(\bar{\mathbf{A}}\bar{\mathbf{B}} - \mathbf{Y}) = \mathbf{0}, \tag{26}$$

$$\Delta\mathbf{B}_1 = \mathbf{0} \Rightarrow (\Delta\mathbf{A}_0\bar{\mathbf{B}} + \bar{\mathbf{A}}\Delta\mathbf{B}_0)\bar{\mathbf{B}}^\top + (\bar{\mathbf{A}}\bar{\mathbf{B}} - \mathbf{Y})\Delta\mathbf{B}_0^\top = \mathbf{0}. \tag{27}$$

There exists $\mathbf{C}_r, \mathbf{D}_r \in \mathbb{R}^{r\times r}$, $\mathbf{C}_\perp \in \mathbb{R}^{(m-r)\times r}$, 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 (26) and (27), 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

$$\mathbf{C}_r\mathbf{\Sigma}_r^{1/2} + \mathbf{\Sigma}_r^{1/2}\mathbf{D}_r = \mathbf{0}, \tag{28}$$

$$\mathbf{\Sigma}_r\mathbf{D}_\perp - \mathbf{C}_\perp^\top\mathbf{\Sigma}_\perp = \mathbf{0}, \tag{29}$$

$$\mathbf{C}_\perp\mathbf{\Sigma}_r - \mathbf{\Sigma}_\perp\mathbf{D}_\perp^\top = \mathbf{0}. \tag{30}$$

Since no singular values of $\mathbf{\Sigma}_r$ equals with that of $\mathbf{\Sigma}_\perp$, Equations (29) and (30) implies that $\mathbf{C}_\perp\mathbf{\Sigma}_r^2 = \mathbf{\Sigma}_\perp^2\mathbf{C}_\perp$ 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—the gradients of $\mathcal{L}$ with respect to both $\mathbf{A}$ and $\mathbf{B}$ are never rank-deficient—and converges, then the convergence should happen at a finite time.

Proposition 33 *Assume that SpecGF with $\mathcal{T}$ is analytic for all time $t \geq 0$. If SpecGF with $\mathcal{T}$ converges, then it converges finitely to a critical point.*

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 23: there are a neighborhood W of $(\mathbf{A}(\infty), \mathbf{B}(\infty))$, 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,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta(t)^{1-b} = (1-b)\Delta(t)^{-b}\Delta'(t) \leq -\frac{1-b}{C}$$
$$\Rightarrow \Delta(t)^{1-b} - \Delta(t_W)^{1-b} \leq -\frac{1-b}{C}(t - t_W).$$

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 34 (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,

$$\langle \mathbf{X}, \mathcal{T}_\beta(\mathbf{X})\rangle \geq \sqrt{\beta}\|\mathcal{T}_\beta(\mathbf{X})\|_{\mathrm{F}}^2 \quad \Rightarrow \quad \|\mathcal{T}_\beta(\mathbf{X})\|_{\mathrm{F}} \leq \beta^{-\frac{1}{2}}\|\mathbf{X}\|_{\mathrm{F}}. \tag{31}$$

Proposition 35 ([6], Theorem 5.2) *Consider the system of differential equations*

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)) \tag{32}$$

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

$$c\dot{V}(\mathbf{z}) + \|f(\mathbf{z})\| \leq 0 \tag{33}$$

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 36 (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 23: at each global minima $\mathbf{x}_*$ of $\mathcal{L}$, there exist a neighborhood W of $\mathbf{x}_*$, 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 minima $\mathbf{x}_*$. Without loss of generality, take W as an open ball centered at $\mathbf{x}_*$ and of (finite) positive radius. On W, define

$$V(\mathbf{x}) \triangleq (\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)^{1-b}. \tag{34}$$

Then V is continuous, nonnegative, minimized at $\mathbf{x}_*$ and

$$\dot{V}(\mathbf{x}) = (1-b)\left(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*\right)^{-b} \dot{\mathcal{L}}(\mathbf{x}).$$

On W, $\nabla_{\mathbf{A}}\mathcal{L}$ and $\nabla_{\mathbf{B}}\mathcal{L}$ are both bounded; take the bound as C_W so that

$$\|\nabla_{\mathbf{A}}\mathcal{L}\|_{\mathrm{F}} \le C_W, \quad \|\nabla_{\mathbf{B}}\mathcal{L}\|_{\mathrm{F}} \le C_W.$$

Observe

$$\begin{aligned}
-\dot{\mathcal{L}} &= \langle \nabla_{\mathbf{A}}\mathcal{L}, \mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L})\rangle + \langle \nabla_{\mathbf{B}}\mathcal{L}, \mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L})\rangle \\
&= \sum_i \frac{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L})}{\sqrt{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L}) + \beta}} + \sum_j \frac{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L})}{\sqrt{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L}) + \beta}} \\
&\ge \sum_i \frac{\sigma_i^2(\nabla_{\mathbf{A}}\mathcal{L})}{\sqrt{C_W^2 + \beta}} + \sum_j \frac{\sigma_j^2(\nabla_{\mathbf{B}}\mathcal{L})}{\sqrt{C_W^2 + \beta}} \\
&= \frac{1}{\sqrt{C_W^2 + \beta}} \|\nabla\mathcal{L}\|_{\mathrm{F}}^2.
\end{aligned}$$

In addition, for the same f_β in the proof of Theorem 30,

$$\begin{aligned}
\|f_\beta(\mathbf{x})\|_{\mathrm{F}} &= \sqrt{\|\mathcal{T}_\beta(\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{x}))\|_{\mathrm{F}}^2 + \|\mathcal{T}_\beta(\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{x}))\|_{\mathrm{F}}^2} \\
&\le \sqrt{\beta^{-1}\left(\|\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2 + \|\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2\right)} = \beta^{-\frac{1}{2}}\|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}.
\end{aligned}$$

Hence, on W,

$$\begin{aligned}
-\dot{V}(\mathbf{x}) &= -(1-b)\left(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*\right)^{-b} \dot{\mathcal{L}}(\mathbf{x}) \\
&\ge (1-b)\left(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*\right)^{-b} \frac{1}{\sqrt{C_W^2 + \beta}} \|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}}^2 \\
&\ge \frac{1-b}{C\sqrt{C_W^2 + \beta}} \|\nabla\mathcal{L}(\mathbf{x})\|_{\mathrm{F}} \\
&\ge \frac{1-b}{C\sqrt{C_W^2 + \beta}} \beta^{\frac{1}{2}} \|f_\beta(\mathbf{x})\|_{\mathrm{F}}.
\end{aligned}$$

$c = \frac{C\sqrt{C_W^2+\beta}}{(1-b)\sqrt{\beta}}$ 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. For now, let ρ_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 $\|\mathbf{z} - \mathbf{x}_*\|_{\mathrm{F}} \le \delta$. We have

$$\|\Delta\mathbf{A}\|_{\mathrm{F}} \triangleq \|\mathbf{A} - \mathbf{A}_*\|_{\mathrm{F}} \le \delta, \quad \|\Delta\mathbf{B}\|_{\mathrm{F}} \triangleq \|\mathbf{B} - \mathbf{B}_*\|_{\mathrm{F}} \le \delta.$$

Then, for some constants $c' > 0$ (may depend on $\mathbf{x}_*$)

$$\begin{aligned}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}} &= \|\Delta\mathbf{AB}_* + \mathbf{A}_*\Delta\mathbf{B} + \Delta\mathbf{A}\Delta\mathbf{B}\|_{\mathrm{F}} \\ &\le \|\Delta\mathbf{AB}_*\|_{\mathrm{F}} + \|\mathbf{A}_*\Delta\mathbf{B}\|_{\mathrm{F}} + \|\Delta\mathbf{A}\Delta\mathbf{B}\|_{\mathrm{F}} \\ &\le \delta\left(\|\mathbf{A}_*\|_2 + \|\mathbf{B}_*\|_2\right) + \delta^2 \\ &\le c' \max\{\delta, \delta^2\}.\end{aligned}$$

(e.g., take $c' = \|\mathbf{A}_*\|_2 + \|\mathbf{B}_*\|_2 + 1$.) Hence,

$$\mathcal{L}(\mathbf{z}) = \frac{1}{2}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}^2 \le \frac{1}{2}c'^2 \max\{\delta^2, \delta^4\} \quad \Rightarrow \quad V(\mathbf{z}) \le \left(\frac{1}{2}c'^2 \max\{\delta^2, \delta^4\} - \mathcal{L}_*\right)^{1-b}.$$

In the proof of Bhat and Bernstein [6, 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

$$\delta \le \frac{\rho}{4}, \quad \left(\frac{1}{2}c'^2 \max\{\delta^2, \delta^4\} - \mathcal{L}_*\right)^{1-b} \le \frac{\rho}{4c},$$

which translates to

$$\delta(\epsilon, \mathbf{x}_*) = \min\left\{\frac{\rho}{4}, \sqrt{\frac{2}{c'^2}\left(\left(\frac{\rho}{4c}\right)^{\frac{1}{1-b}} + \mathcal{L}_*\right)}, \sqrt[4]{\frac{2}{c'^2}\left(\left(\frac{\rho}{4c}\right)^{\frac{1}{1-b}} + \mathcal{L}_*\right)}\right\}, \quad \rho = \min\{\rho_W, \epsilon\}. \tag{35}$$

We conjecture that ρ_W is sufficiently large for each and every minimizer $\mathbf{x}_*$. If then, δ increases as ϵ increases, and there is a possibility that initial points near (but not exactly) the origin fall in some basin of attraction. We empirically checked that the basin of attraction is sufficiently large.

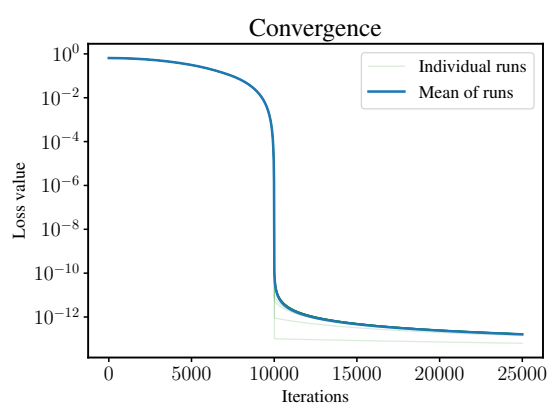


Figure 3: Loss trajectories started at $(\mathbf{A}_i(0), \mathbf{B}_i(0))$.

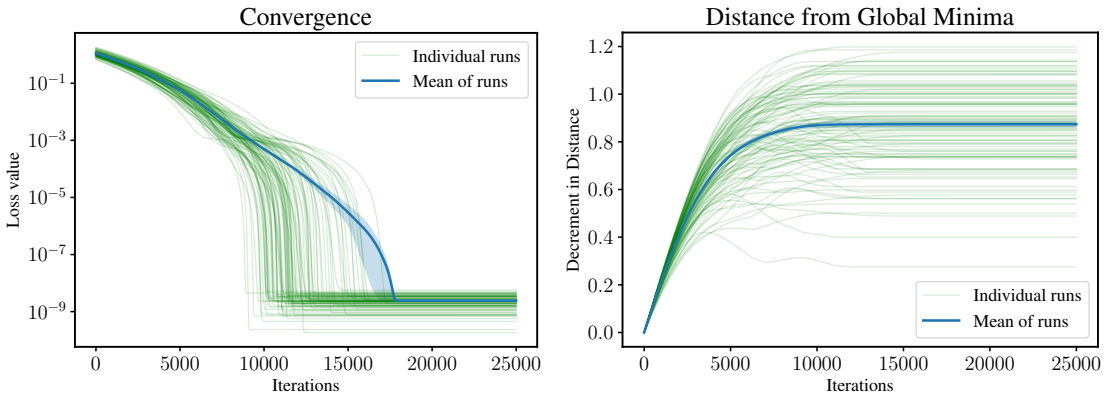


Figure 4: Lyapunov stability of global minima. **Left**: Loss trajectory. **Right**: Decrement of the distance from global minima. The attraction of the basin of some global minima is sufficiently large that, empirically, the origin lies in there.

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 3. 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) loss trajectory $\mathcal{L}(t)$ and (ii) 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 4.

We see that all SpecGF simulations initialized at $(\mathbf{A}_{ij}(0), \mathbf{B}_{ij}(0))$ converged to the global minima, and the decrement is positive, supporting the conjecture that the near-origin points lies in a basin of attraction of some global minima.

D.6. SpecGF with ℓ_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 the same 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 $\mathcal{L}_\lambda : \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n} \to \mathbb{R}$ for a fixed $\lambda > 0$ is defined as

$$\mathcal{L}_\lambda(\mathbf{A}, \mathbf{B}) \triangleq \frac{1}{2}\|\mathbf{AB} - \mathbf{Y}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{B}\|_{\mathrm{F}}^2. \tag{36}$$

  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 derivative of $\mathbf{A}$ and $\mathbf{B}$ is governed by either $\mathcal{T}$ or $\mathcal{T}_\beta$ applied to the gradient:

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

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 21 and 22 readily follows for the regularized SpecGF with $\mathcal{T}_\beta$. Moreover,

$$\|\mathbf{A}(t)\|_{\mathrm{F}}^2 + \|\mathbf{B}(t)\|_{\mathrm{F}}^2 \leq \frac{2}{\lambda}\mathcal{L}_\lambda(t) \leq \frac{2}{\lambda}\mathcal{L}_\lambda(0) = \frac{1}{\lambda}\|\mathbf{Y}\|_{\mathrm{F}}^2 + \|\mathbf{A}(0)\|_{\mathrm{F}}^2 + \|\mathbf{B}(0)\|_{\mathrm{F}}^2.$$

   Thus, $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are both bounded for all $t \geq 0$.

2. Equation (20) in Proposition 24 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 29 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 36 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 LoRA rank r is smaller than the target rank $r^\star$. Although it is specified in Udell et al. [43], we reproduce their result. There, $\mathbf{M}_+ \triangleq \max\{\mathbf{M}, 0\}$ where $\max$ is applied coordinate-wise.

Proposition 37 (ℓ_2-Regularized MF, [43]) *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 (36) is of the form

$$\mathbf{A} = \mathbf{U}_\Omega \left(\mathbf{\Sigma}_\Omega - \lambda \mathbf{I}\right)_+^{\frac{1}{2}}, \quad \mathbf{B} = \left(\mathbf{\Sigma}_\Omega - \lambda \mathbf{I}\right)_+^{\frac{1}{2}} \mathbf{V}_\Omega^\top.$$

up to orthogonal transformation by $\mathbf{R} \in \mathsf{O}(r)$. We denote the submatrix of $\mathbf{U}$ with columns indexed by Ω as $\mathbf{U}_\Omega$, and similarly for $\mathbf{\Sigma}$ and $\mathbf{V}$.

(ii) Every critical point except global minima has a descent direction.

Item 37.(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 37 is found in Udell et al. [43, Section 2.3 and A.1].

Appendix E. Experiments

E.1. Plots for Empirical Validation

In this section, we validate our theoretical predictions, confirming uniform growth under SpecGF.

Experimental Setup. We train $(\mathbf{A}, \mathbf{B})$ with SpecGD on the objective in Equation (1). 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 $\beta = 0, 10^{-8}$. We track core variables $\mathbf{X}$ and $\mathbf{Z}$ with the diagonal entries $d_i = [\mathbf{G}]_{ii}$ and $\sigma_i(\mathbf{AB})$.

Results. We highlight three observations. First, SpecGD with $\mathcal{T}_\beta$ successfully minimizes Equation (1) (see Figure 5), consistent with the convergence behavior of Section 4. Second, Figure 6a shows that $d_i(t) = [\mathbf{X}(t)\mathbf{Z}(t)]_{ii}$ closely tracks $\sigma_i(\mathbf{A}(t)\mathbf{B}(t))$. Third, the curves $\sqrt{d_i(t)}$ exhibit nearly identical slopes, consistent with our Theorem 1.

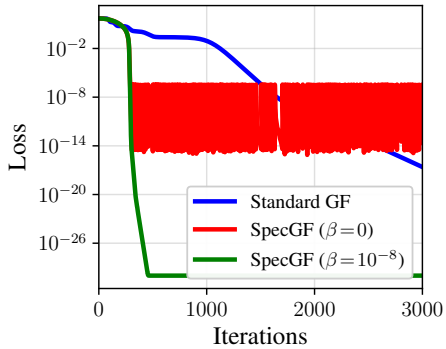
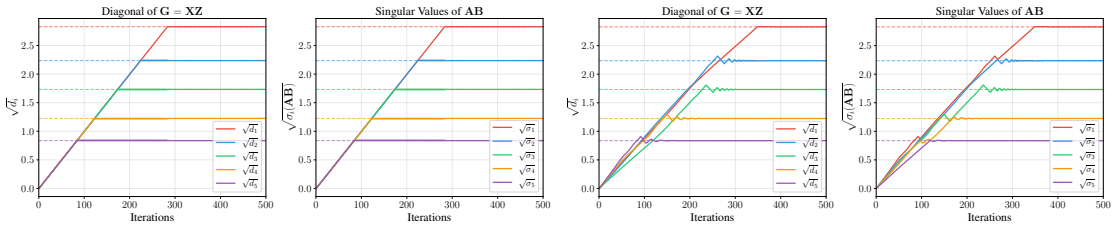


Figure 5: Loss comparison.



(a) SpecGD ($\beta = 10^{-8}$) (b) Muon with NS approximation ($\mu = 0.9$)

Figure 6: Comparison of SpecGD and Muon with NS approximation on matrix factorization with a rank-5 target. The dashed lines indicate the square root of singular values of rank-5 target.

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.

We list additional plots relevant to our theory. Figure 7 shows the singular value evolution under SpedGD with $\mathcal{T}$; it is highly similar to Figure 6a, extending our validation of uniform growth. Figure 8 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 $\sigma_i(\mathbf{A}(t)\mathbf{B}(t))$.

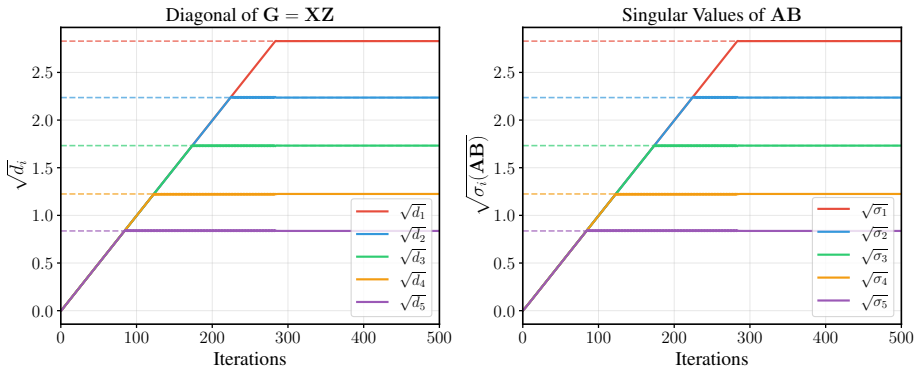
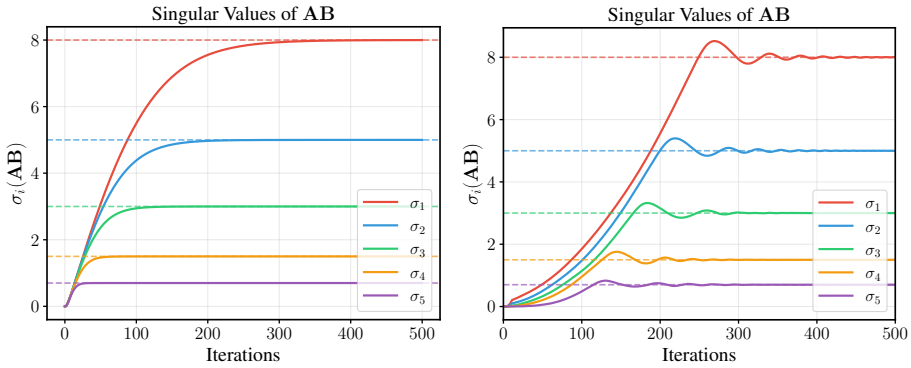


Figure 7: SpecGF ($\beta = 0$) on matrix factorization with a rank-5 target.



(a) Shampoo (b) SOAP

Figure 8: 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 6. We first note that, at each iteration t, Muon updates a weight matrix $\mathbf{W}$ for a loss function $f(\mathbf{W})$ as

$$\begin{aligned} \mathbf{M}_t &= \nabla f(\mathbf{W}_t) + \mu \mathbf{M}_{t-1}, \\ \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_t \mathcal{T}(\mathbf{M}_t), \end{aligned} \tag{37}$$

where $\mu \in [0, 1]$ tunes the momentum, η_t is the step size, and $\mathcal{T}$ orthogonalizes the momentum.

In Figure 9, we vary the momentum parameter μ in Equation (37), 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 μ, indicating that this behavior is robust to the strength of momentum under exact orthogonalization.

In Figure 10, 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 11, 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 γ. In Figure 12, we observe that the uniform growth of singular values persists as depth increases, and Figure 13 further reveals that the growth rate scales as $\Theta(t^L)$ for depth L.

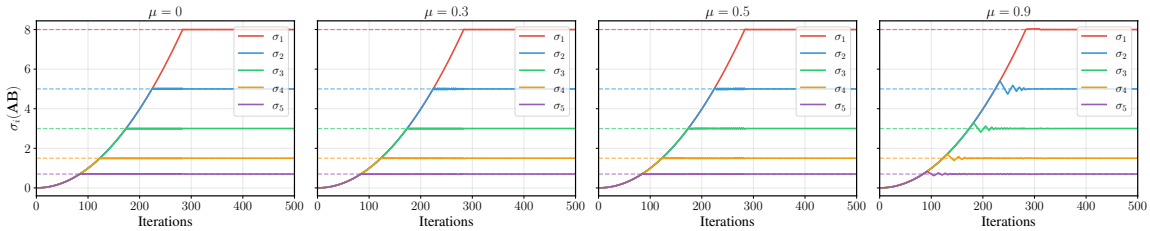


Figure 9: Comparison of singular value evolutions across $\mu \in \{0, 0.3, 0.5, 0.9\}$ from exact orthogonalization $\mathcal{T}$. Notice the consistent uniform growth.

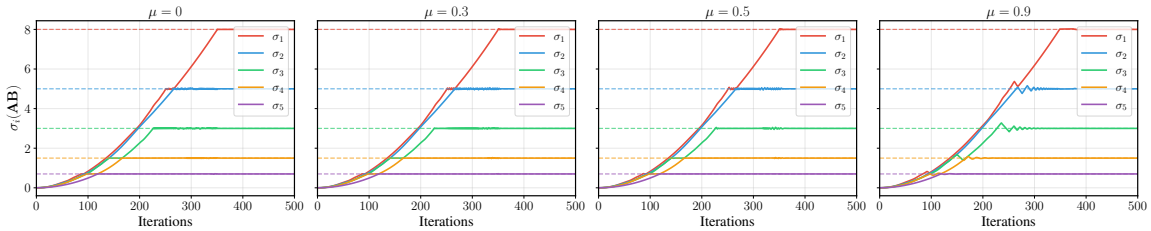


Figure 10: Comparison of singular value evolutions across $\mu \in \{0, 0.3, 0.5, 0.9\}$ from Newton-Schulz iterations. Notice the consistent uniform growth.

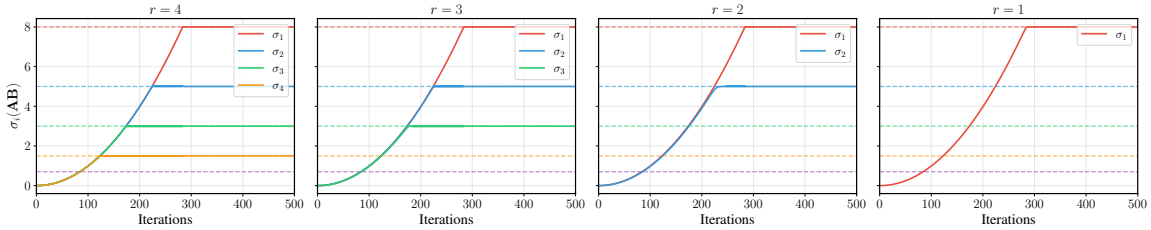


Figure 11: Comparison of singular value evolutions across LoRA rank $r \in \{1, 2, 3, 4\}$. Notice that SpecGF converges toward best rank-r approximation.

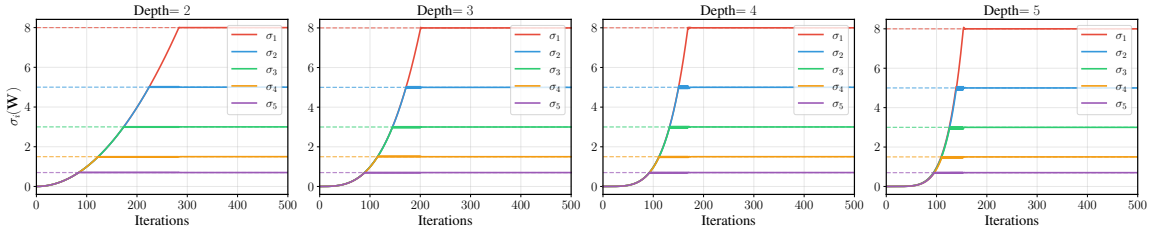


Figure 12: Comparison of singular value evolutions across the depth $L \in \{2, 3, 4, 5\}$. Notice the consistent uniform growth.

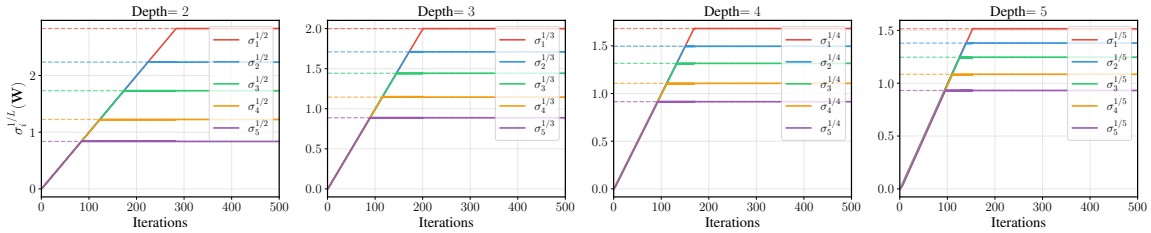


Figure 13: Comparison of the L-th root of the singular value evolutions across the depth $L \in \{2, 3, 4, 5\}$. Notice the consistent uniform *linear* growth.

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) [34] on the SST-2, RTE, and CoLA datasets [46] 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 [12] on CIFAR100 [32], Caltech101 [16], and EuroSAT [23] 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 trainings.

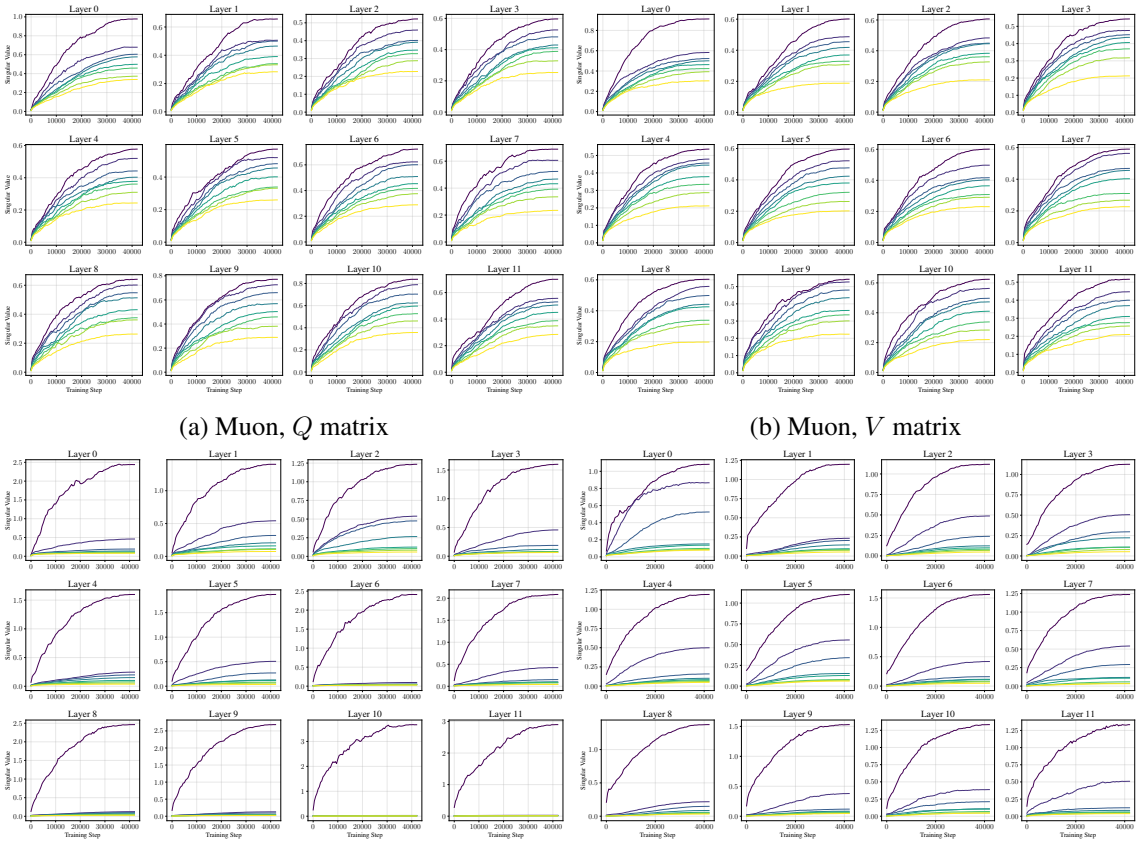
For larger-scale experiments, we train LLaMA-3.2-1B [19] on Alpaca dataset [41] for supervised fine-tuning. We also train a larger model, LLaMA-3.2-3B, on Alpaca, OpenHermes 2.5 [42], and UltraChat [11] datasets. We randomly sampled one million instances from OpenHermes 2.5. As in RoBERTa-base, we also use the same LoRA configuration, i.e., rank $r = 8$ and the scaling factor $\alpha = 16$. We train 3 epochs with a batch size 128 using AdamW and Muon optimizers and with weight decay of 0.01. We use cosine learning rate scheduling with 4% warmup steps of total training iterations. We used NVIDIA A100 GPUs for these trainings.

E.3.2. SINGULAR VALUE DYNAMICS

We provide the full training dynamics of the singular values of projection matrices where LoRA is applied. 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.

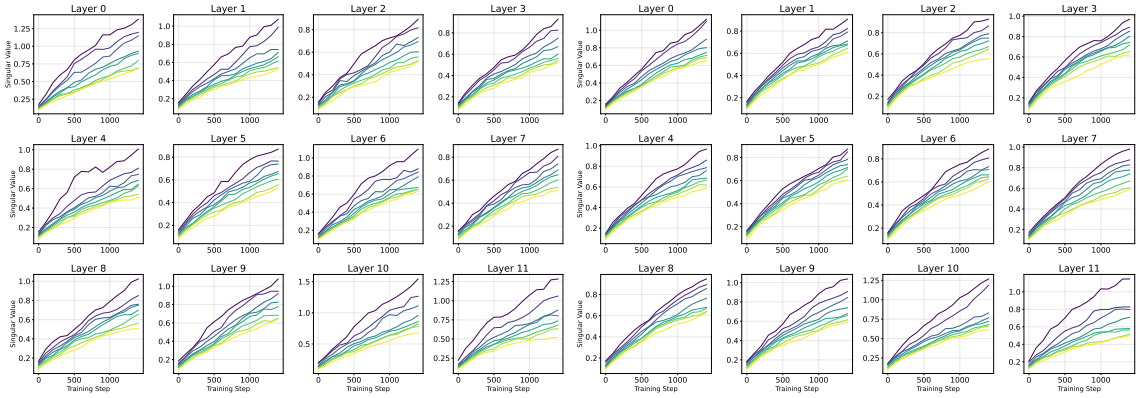
RoBERTa-Base with SST-2/CoLA/RTE dataset.



(a) Muon, Q matrix (b) Muon, V matrix

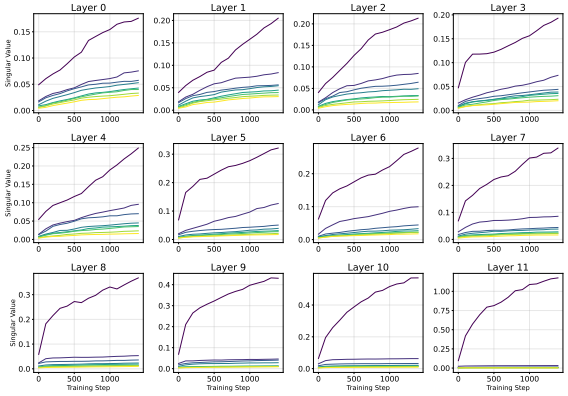
(c) AdamW, Q matrix (d) AdamW, V matrix

Figure 14: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with SST-2 dataset.

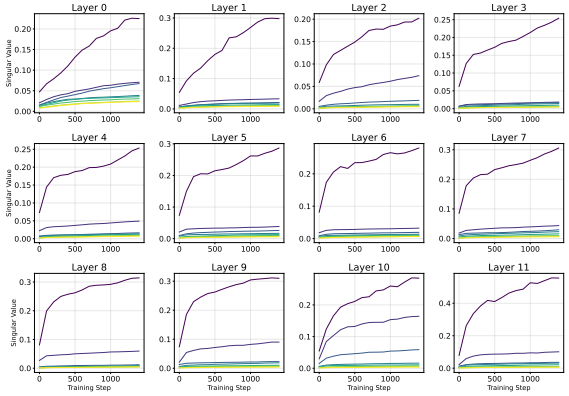


(a) Muon, Q matrix

(b) Muon, V matrix

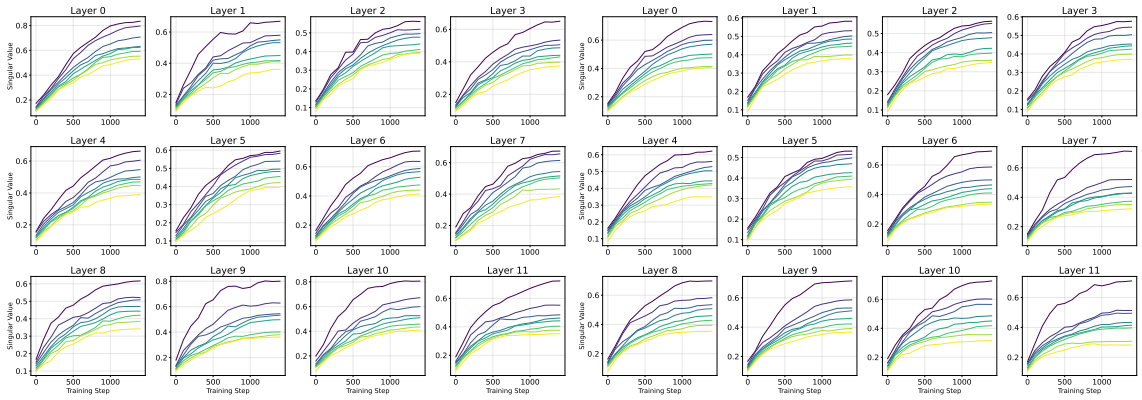


(c) AdamW, Q matrix



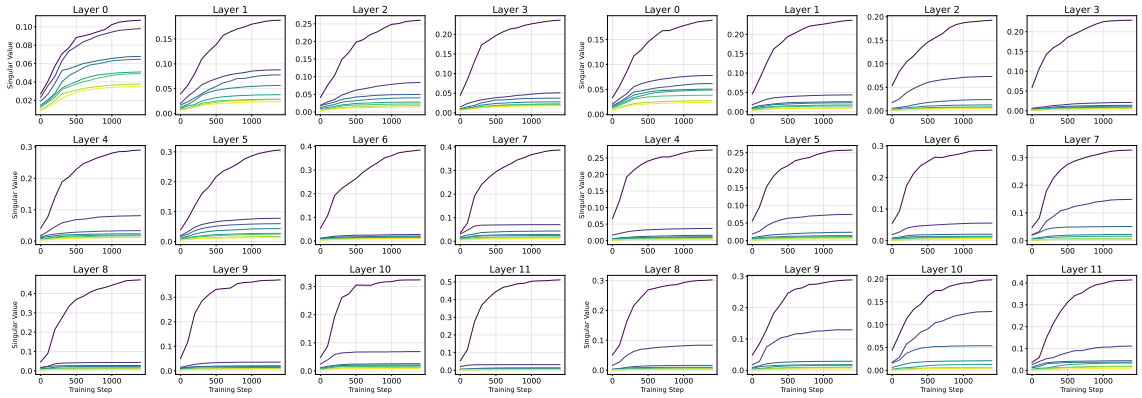
(d) AdamW, V matrix

Figure 15: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with CoLA dataset.



(a) Muon, Q matrix

(b) Muon, V matrix

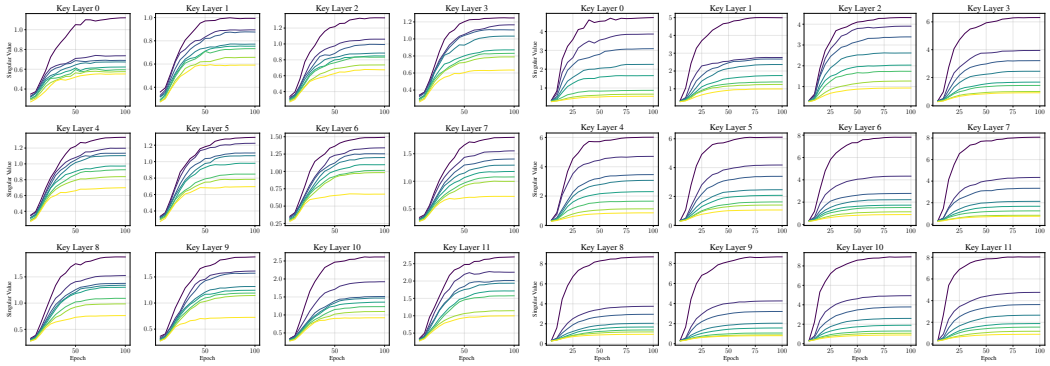


(c) AdamW, Q matrix

(d) AdamW, V matrix

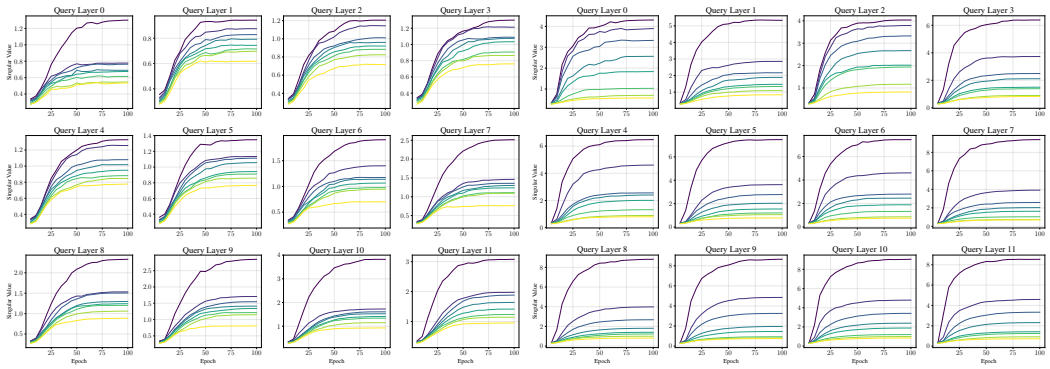
Figure 16: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with RTE dataset.

Vit-B with CIFAR100/Caltech101/EuroSAT dataset.



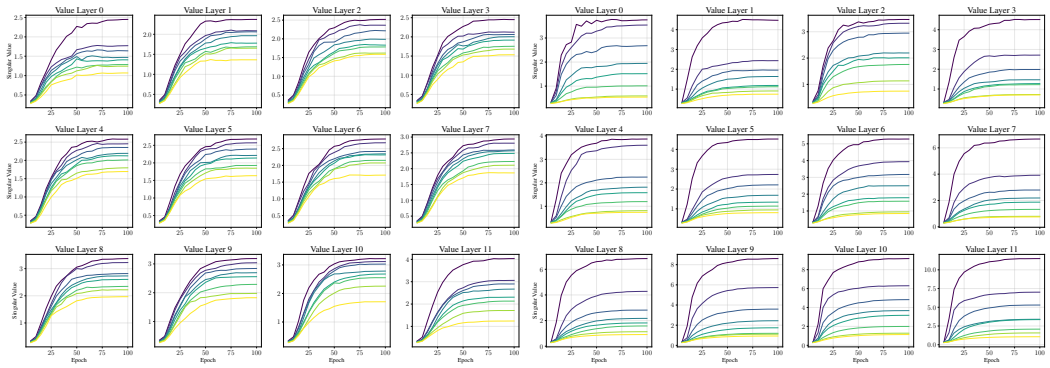
(a) Muon, K matrix

(b) AdamW, K matrix



(c) Muon, Q matrix

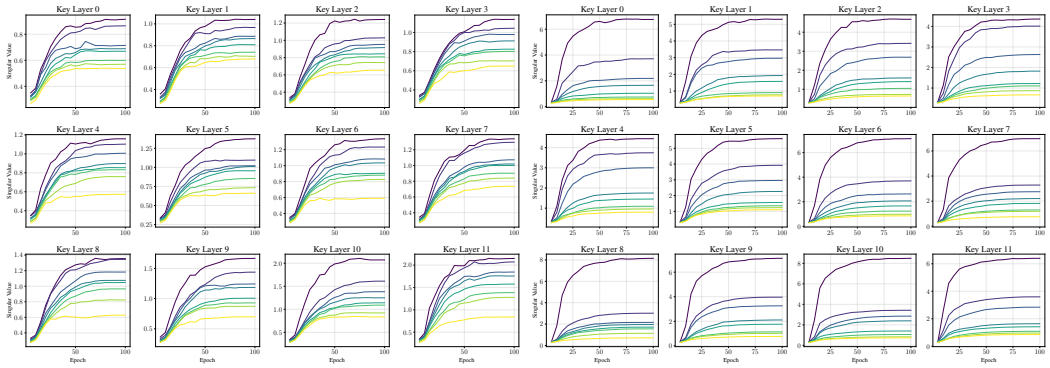
(d) AdamW, Q matrix



(e) Muon, V matrix

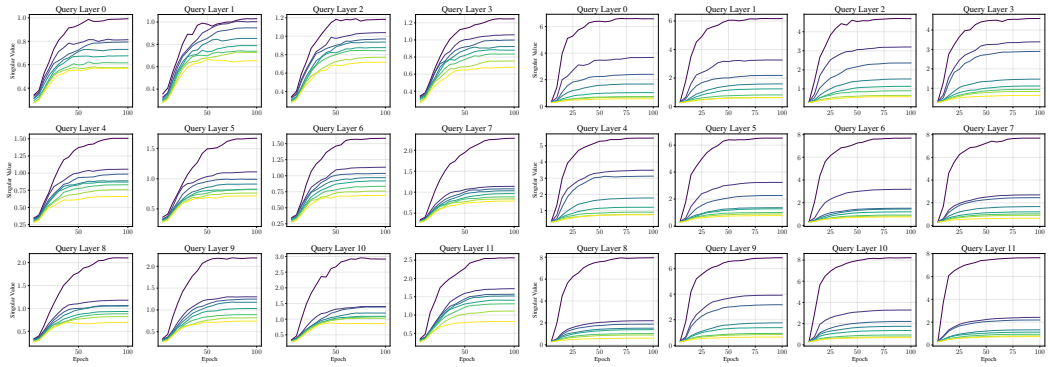
(f) AdamW, V matrix

Figure 17: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained with AdamW, CIFAR100 dataset.



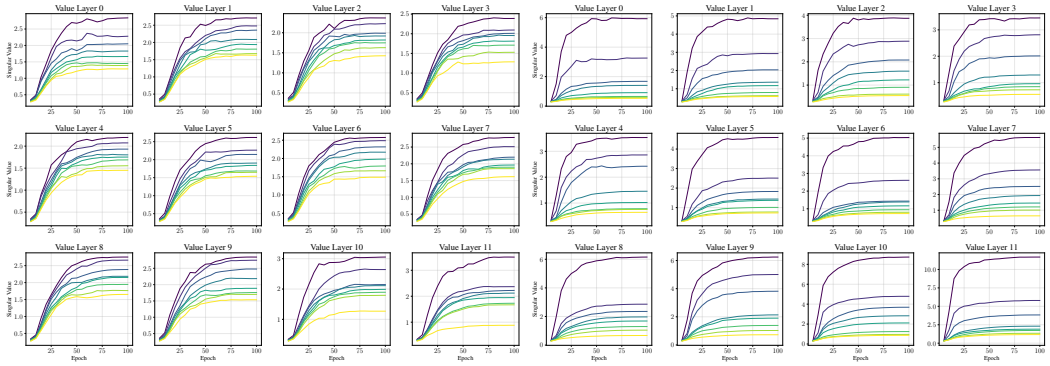
(a) Muon, K matrix

(b) AdamW, K matrix



(c) Muon, Q matrix

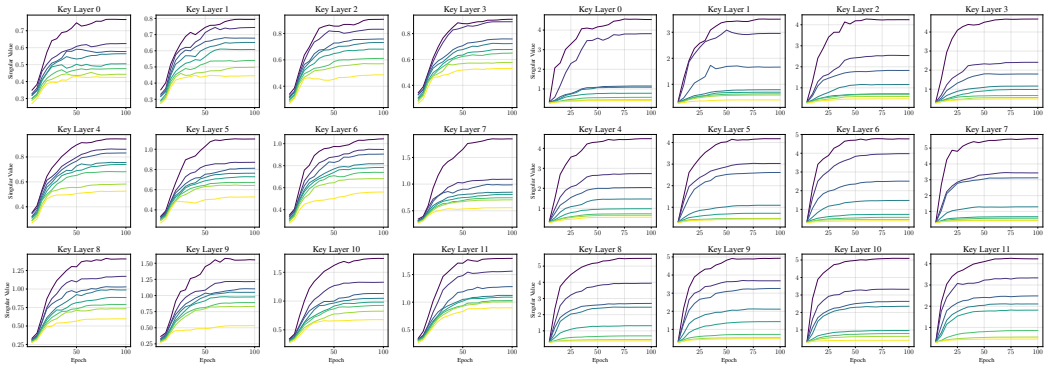
(d) AdamW, Q matrix



(e) Muon, V matrix

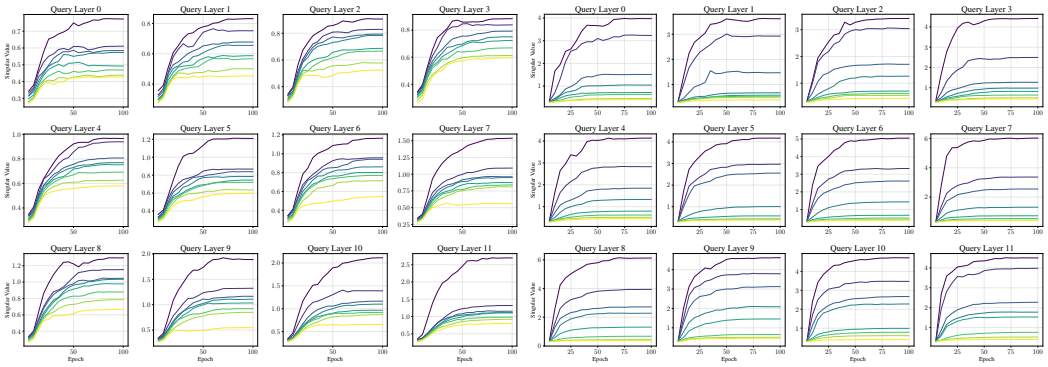
(f) AdamW, V matrix

Figure 18: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained with Caltech101 dataset.



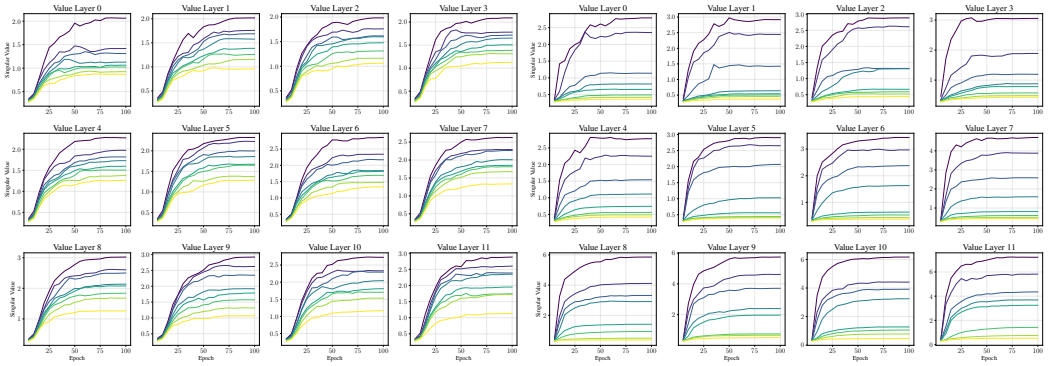
(a) Muon, K matrix

(b) AdamW, K matrix



(c) Muon, Q matrix

(d) AdamW, Q matrix

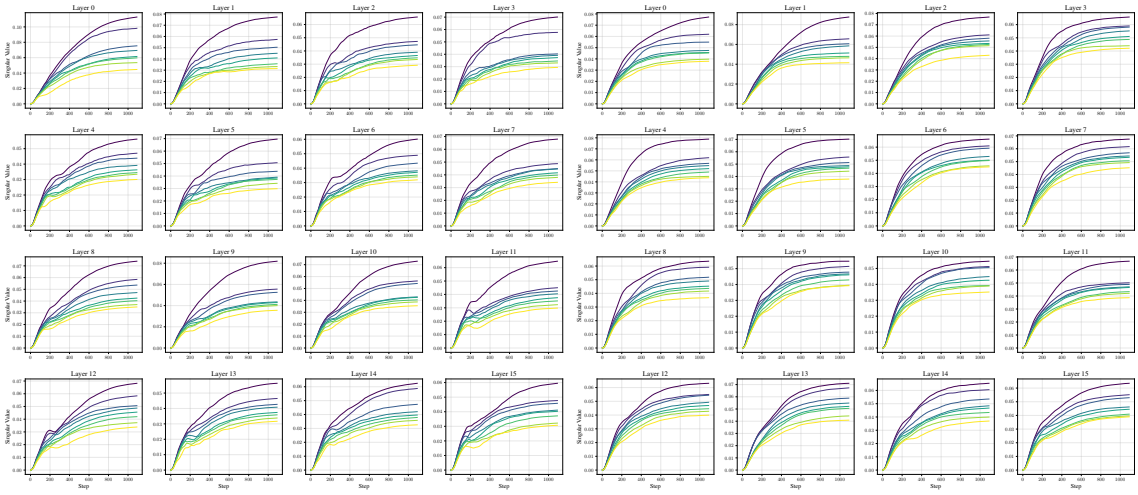


(e) Muon, V matrix

(f) AdamW, V matrix

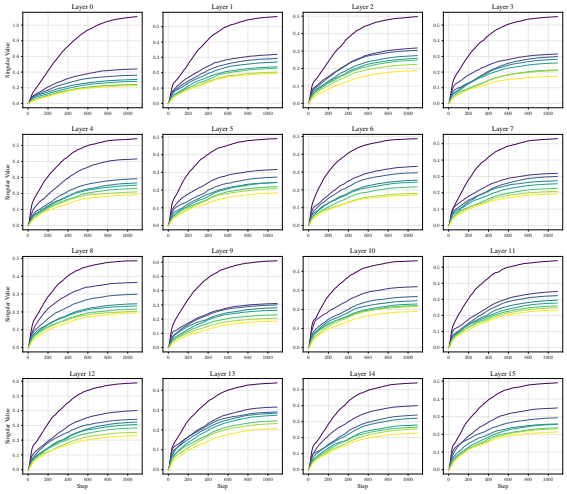
Figure 19: Singular values for all LoRA $\mathbf{AB}$ layers in ViT-B trained with EuroSAT dataset.

LLaMA-3.2-1B with Alpaca dataset.

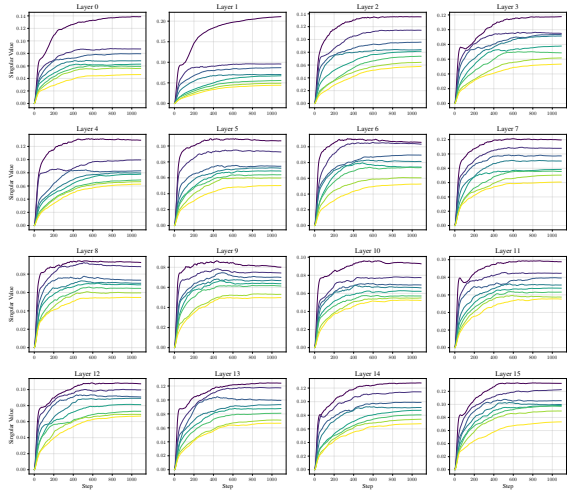


(a) Muon, Q matrix

(b) Muon, V matrix



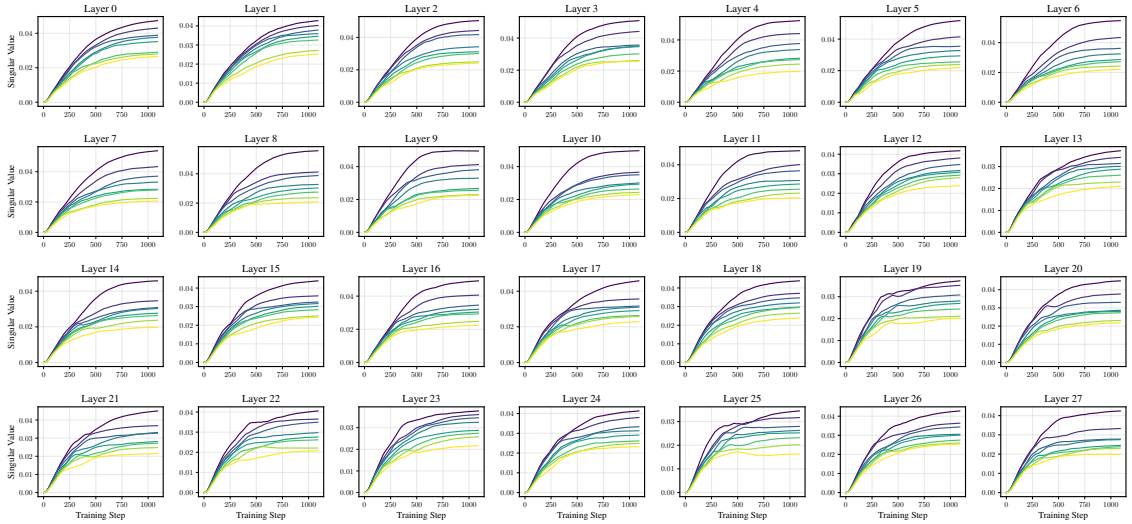
(c) AdamW, Q matrix



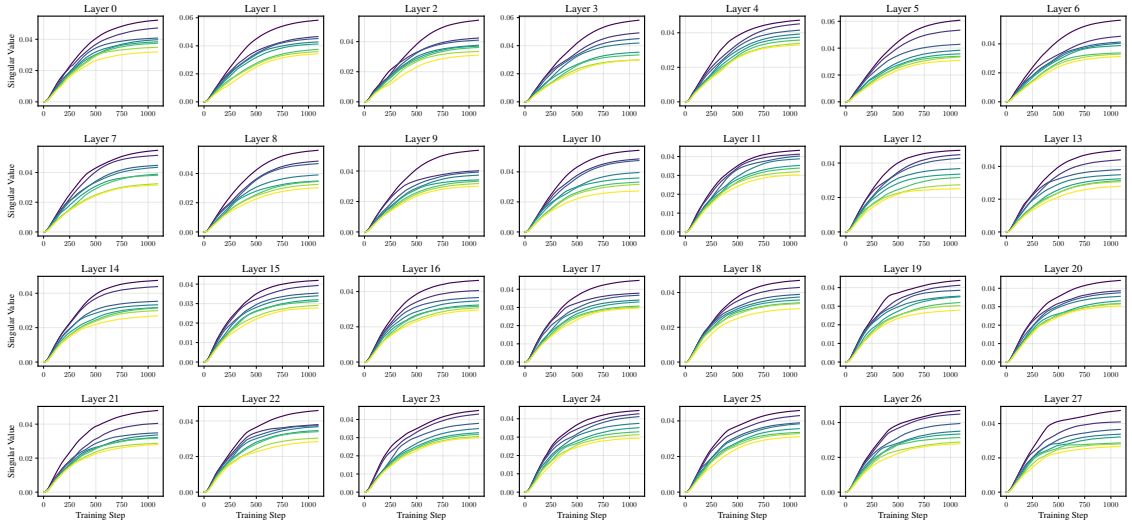
(d) AdamW, V matrix

Figure 20: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-1B trained with Alpaca dataset.

LLaMA-3.2-3B with Alpaca/OpenHermes 2.5/UltraChat dataset.

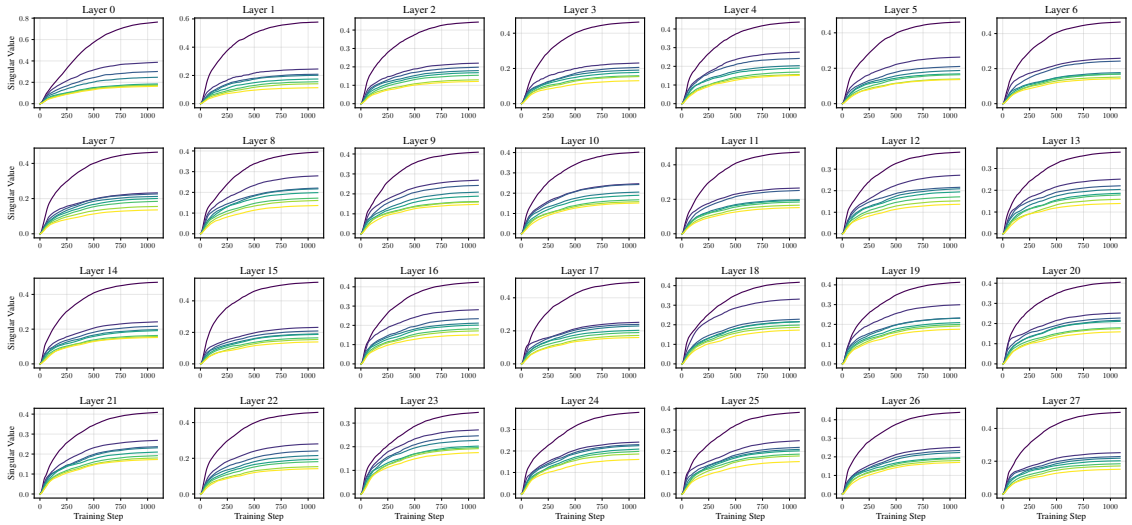


(a) Q matrix

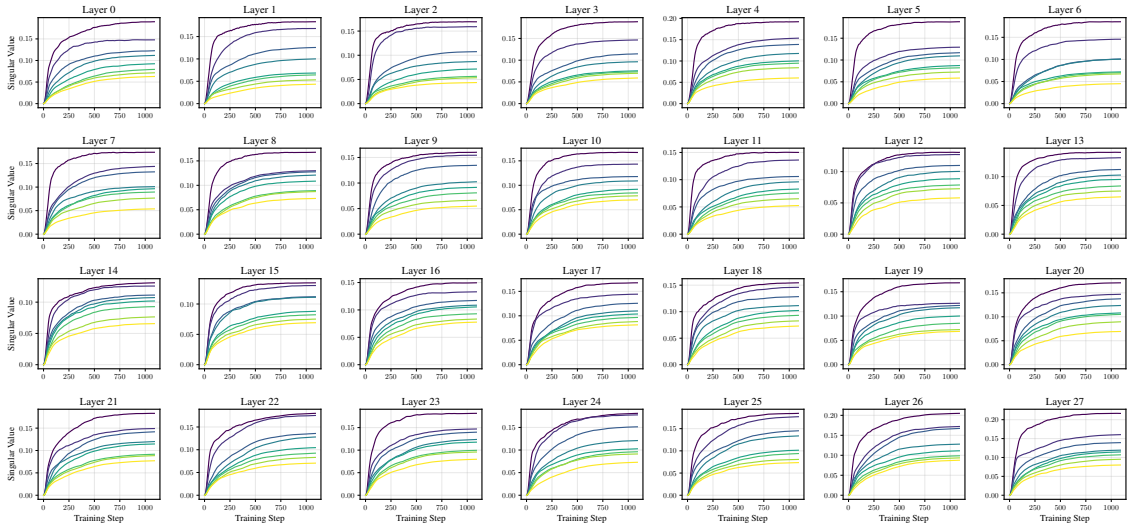


(b) V matrix

Figure 21: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Muon and Alpaca dataset.

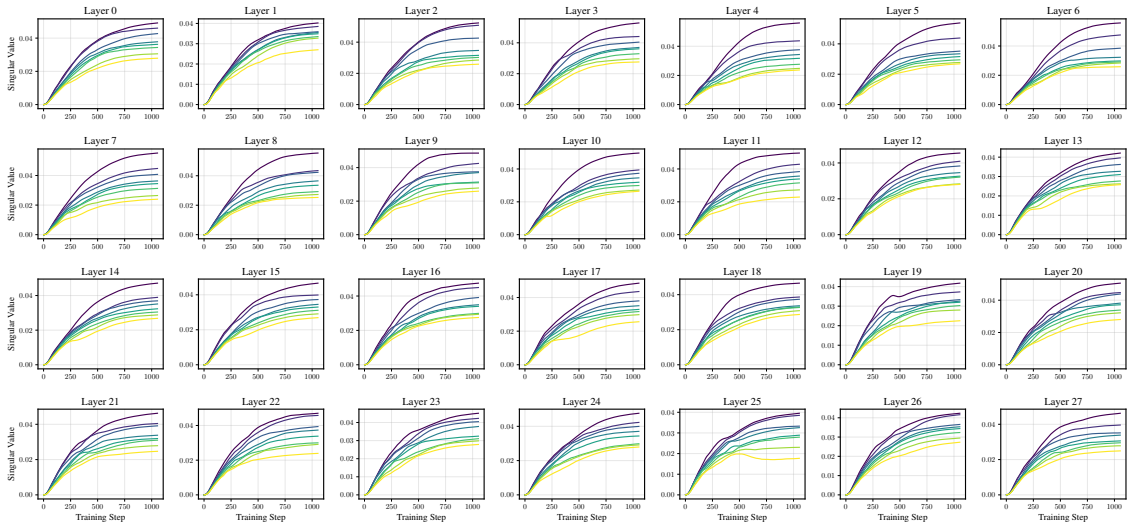


(a) Q matrix

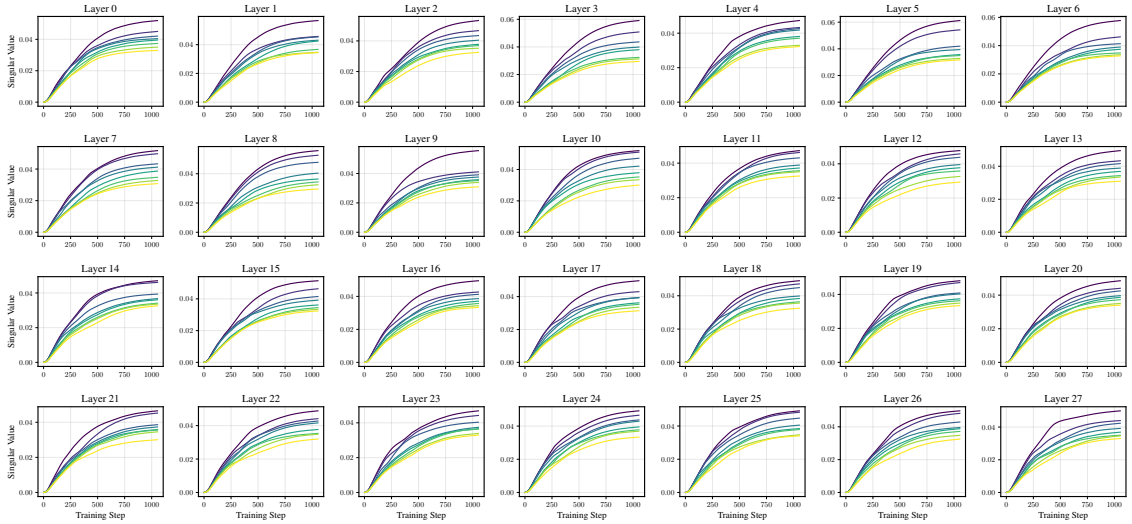


(b) V matrix

Figure 22: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with AdamW and Alpaca dataset.

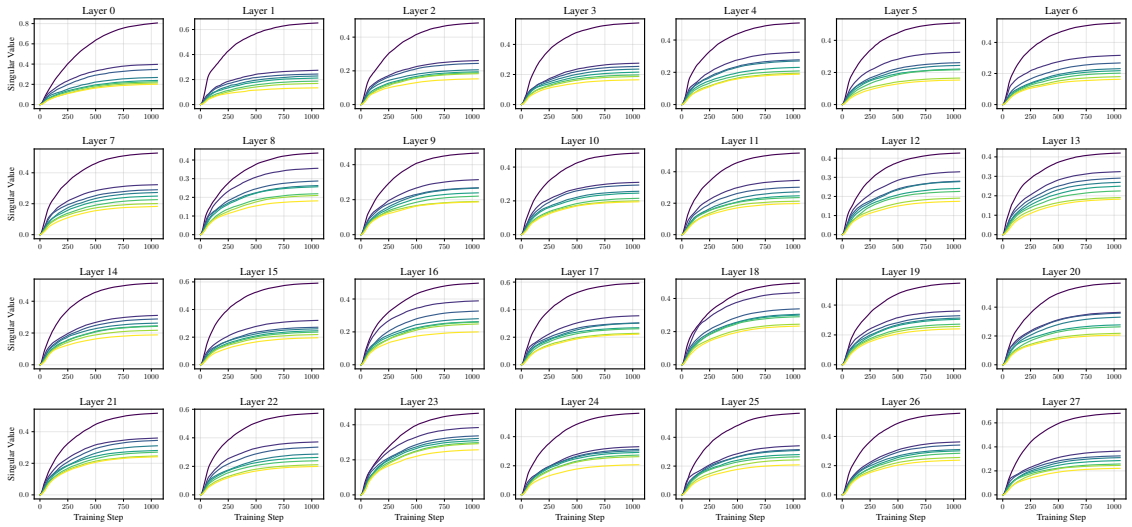


(a) Q matrix

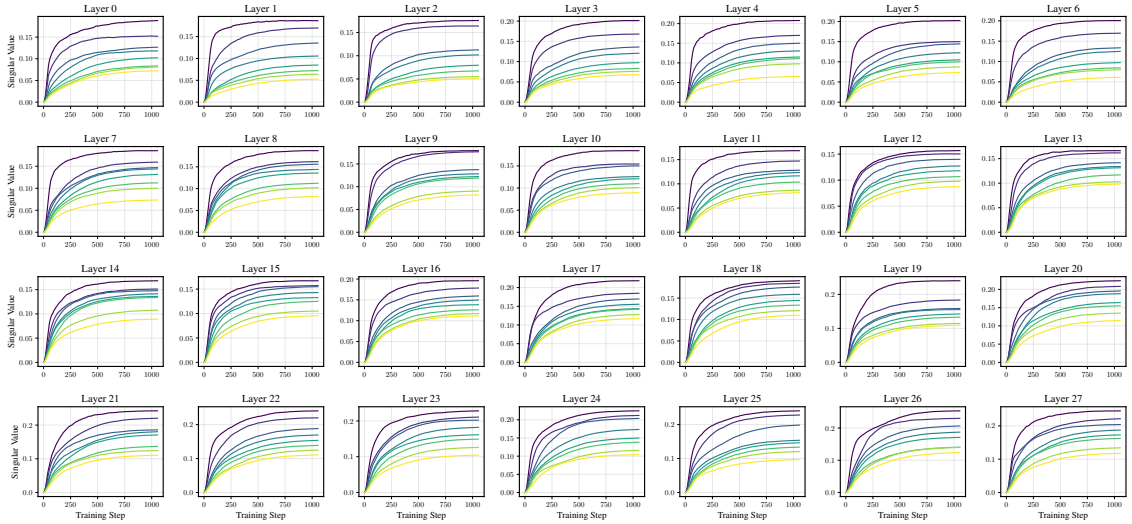


(b) V matrix

Figure 23: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Muon and OpenHermes dataset.

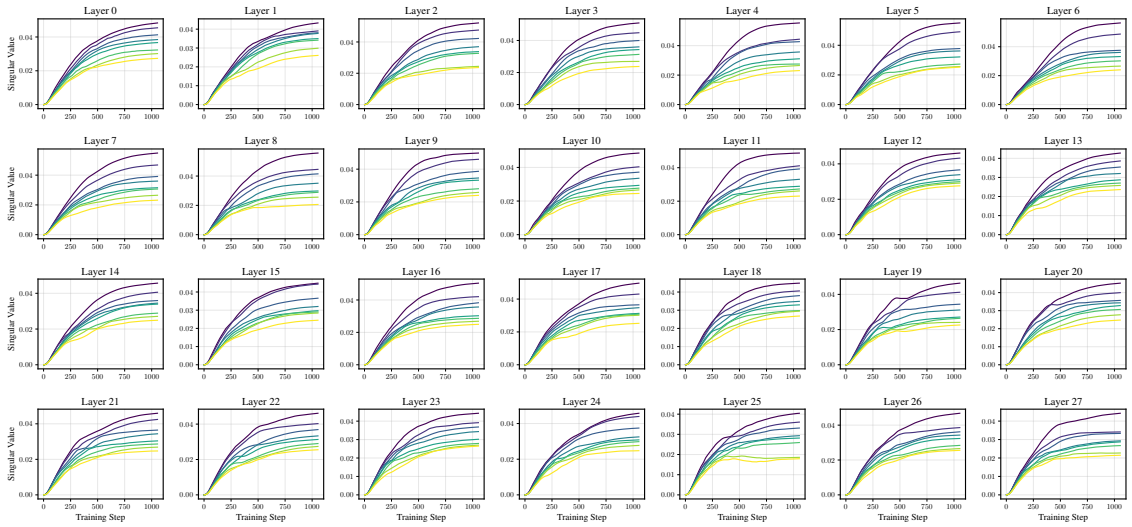


(a) Q matrix

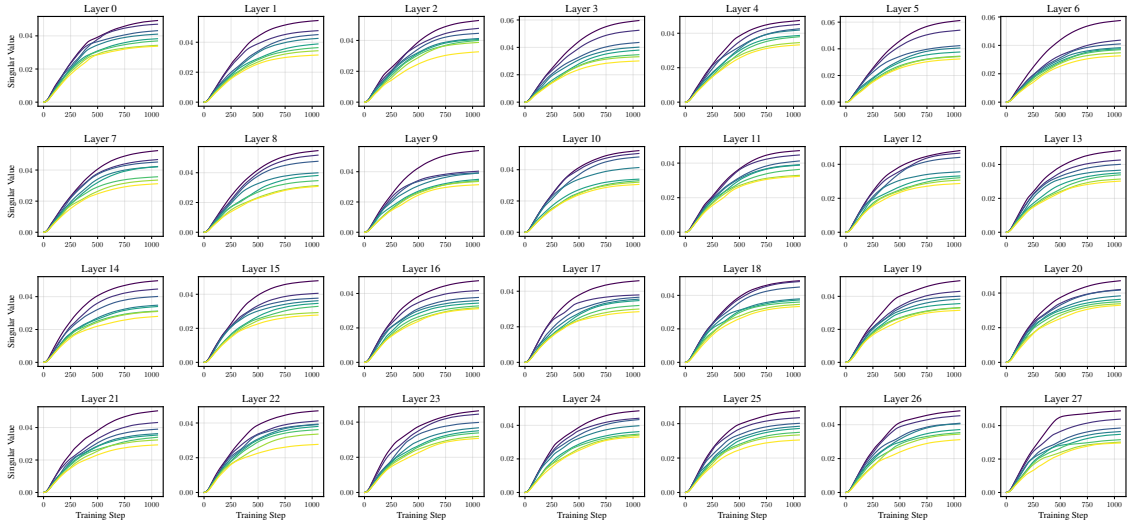


(b) V matrix

Figure 24: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with AdamW and OpenHermes dataset.

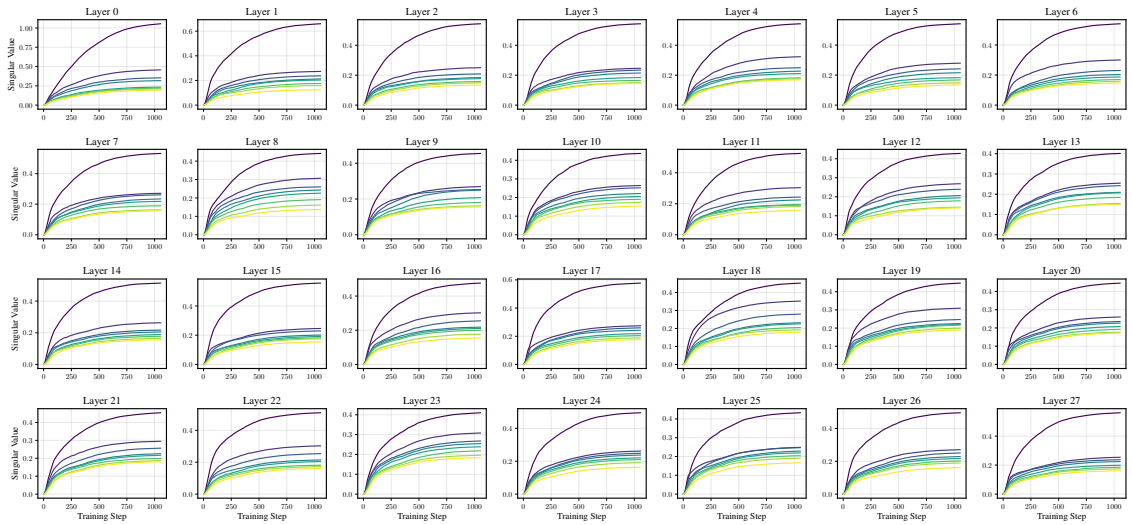


(a) Q matrix

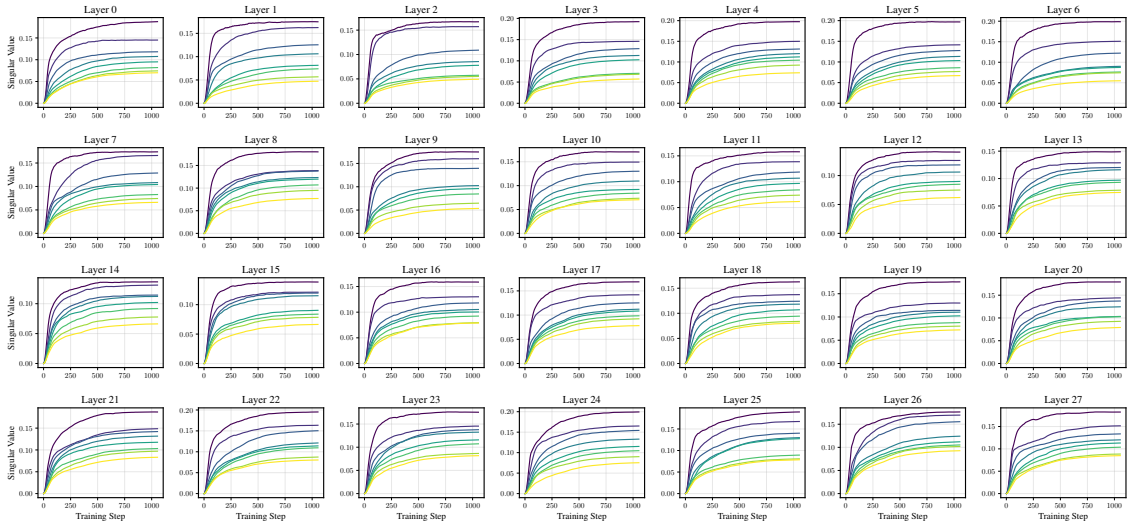


(b) V matrix

Figure 25: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Muon and UltraChat dataset.



(a) Q matrix



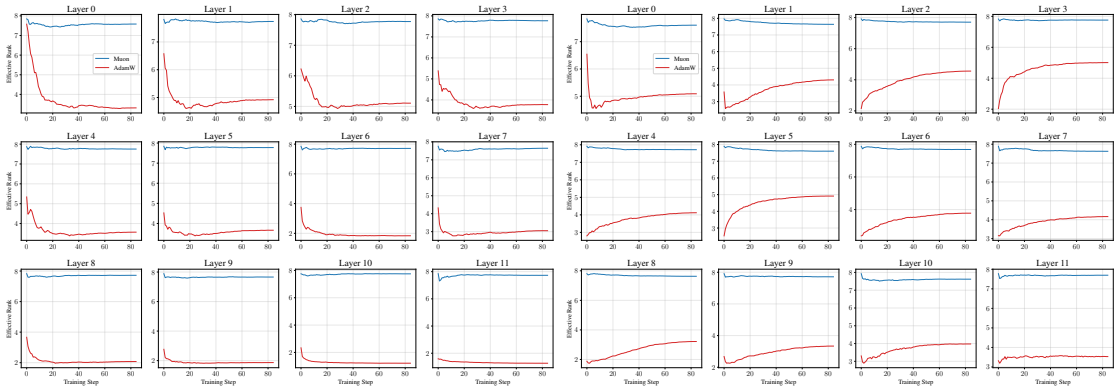
(b) V matrix

Figure 26: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with AdamW and UltraChat dataset.

E.3.3. EFFECTIVE RANK

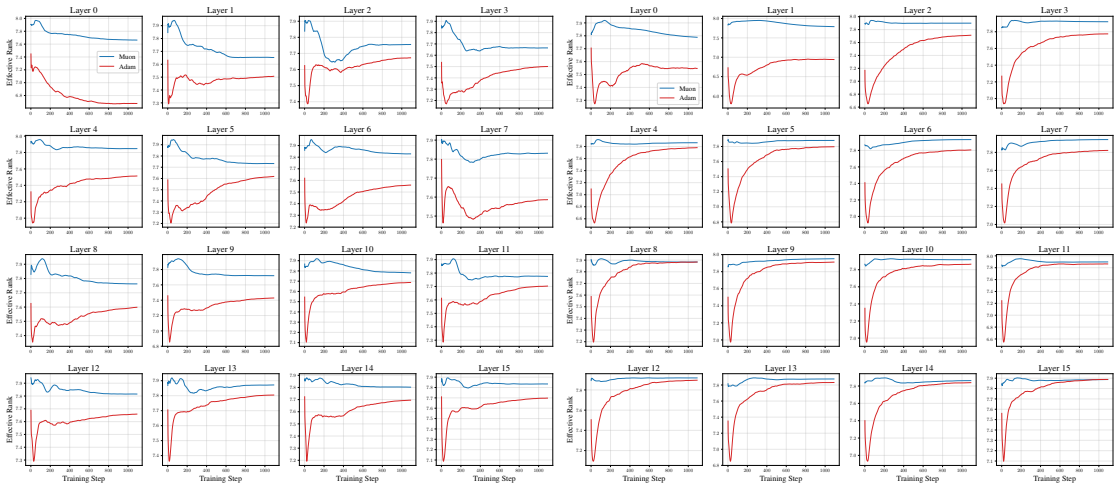
We provide the evolution of the effective rank of query Q and value V projection matrices where LoRA is applied under the following settings: RoBERTa-Base with SST-2 dataset and LLaMA-3.2-1B with Alpaca dataset. Blue and red lines are the effective ranks of the matrix trained under Muon and AdamW, respectively.

Figure 27 shows the effective rank evolutions for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-1B; Figure 28 shows those in RoBERTa-Base. 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 27: Effective rank of the LoRA $\mathbf{AB}$ layers for the Q and V matrices in RoBERTa-Base.



(a) Q matrix (b) V matrix

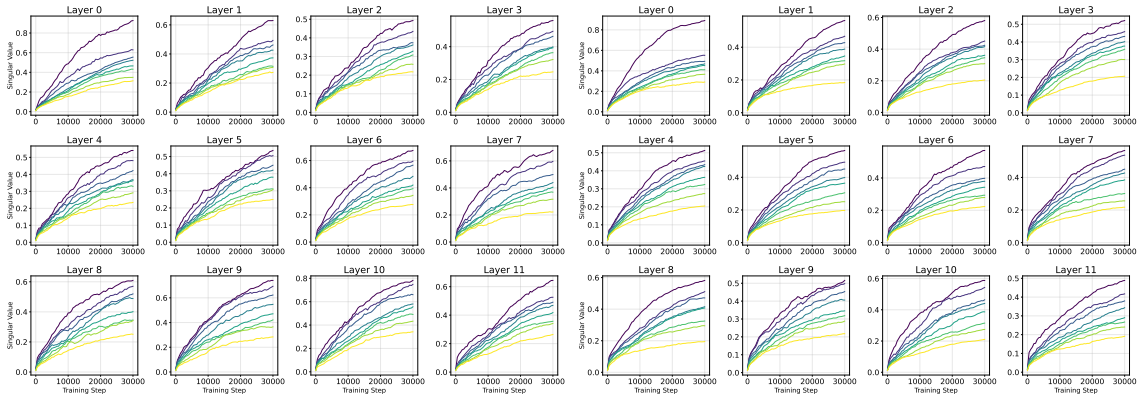
Figure 28: Effective rank of the LoRA $\mathbf{AB}$ matrices for the Q and V in LLaMA-3.2-1B.

E.3.4. UNIFORM GROWTH UNDER LORA INITIALIZATION SCHEME

We fine-tune a pre-trained RoBERTa-base model on the SST-2 dataset using LoRA adapters applied to the query and key 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 Optimized Initialization (LOI) where $\mathbf{AB}$ follows Locally Optimized Initialization (LOI) [7]. Note that this initialization injects spectral information regarding the target matrix to the initialization. We used NVIDIA RTX A6000 GPUs for these trainings.

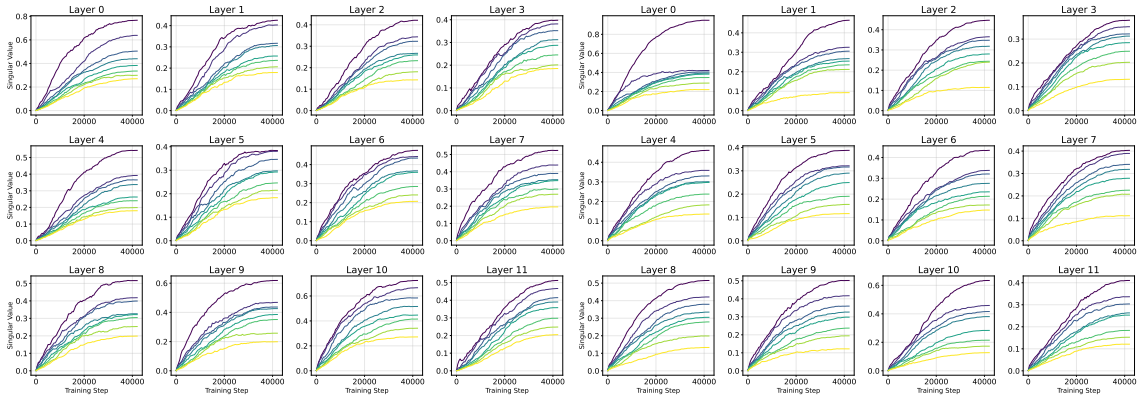
We provide the full training dynamics of the singular values of query Q and value V projection matrices where LoRA is applied. Each line in each subplot stands for a single singular value.

Figure 29 shows the singular value evolutions for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under nonzero random initialization; Figure 30 shows those under LOI. Note that the uniform growth persists for these nonzero LoRA initialization schemes.



(a) Q matrix (b) V matrix

Figure 29: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under NRI.



(a) Q matrix (b) V matrix

Figure 30: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with Muon under LOI.

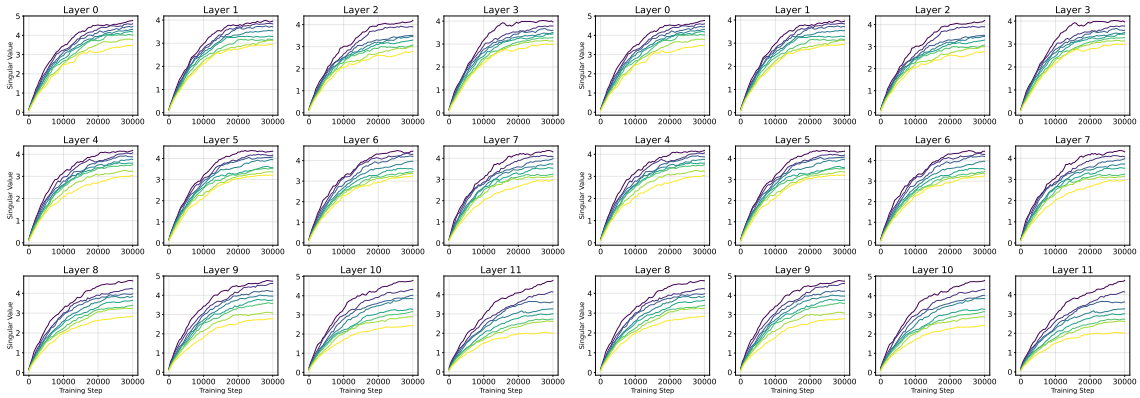
E.3.5. UNIFORM GROWTH UNDER OTHER ORTHOGONALIZATION OPTIMIZERS

We fine-tune LLaMA-3.2-3B with Alpaca, OpenHermes, and UltraChat datasets under other optimizers that utilize the orthogonalized updates. We chose the Shampoo [21], SOAP [45], and Dion [2] optimizers. We also fine-tune RoBERTa-Base with the SST-2 dataset under SOAP. We use the same LoRA configuration as Section E.3.1. We used NVIDIA A100 GPUs for these trainings.

We provide the full training dynamics of the singular values of query Q and value V matrices where LoRA is applied. 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.

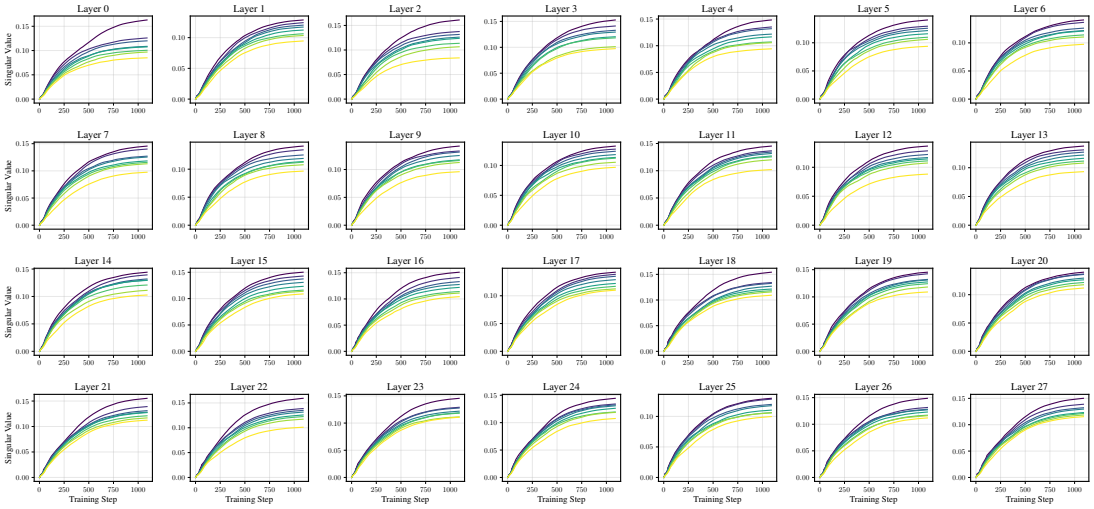
RoBERTa-Base with SOAP optimizer.



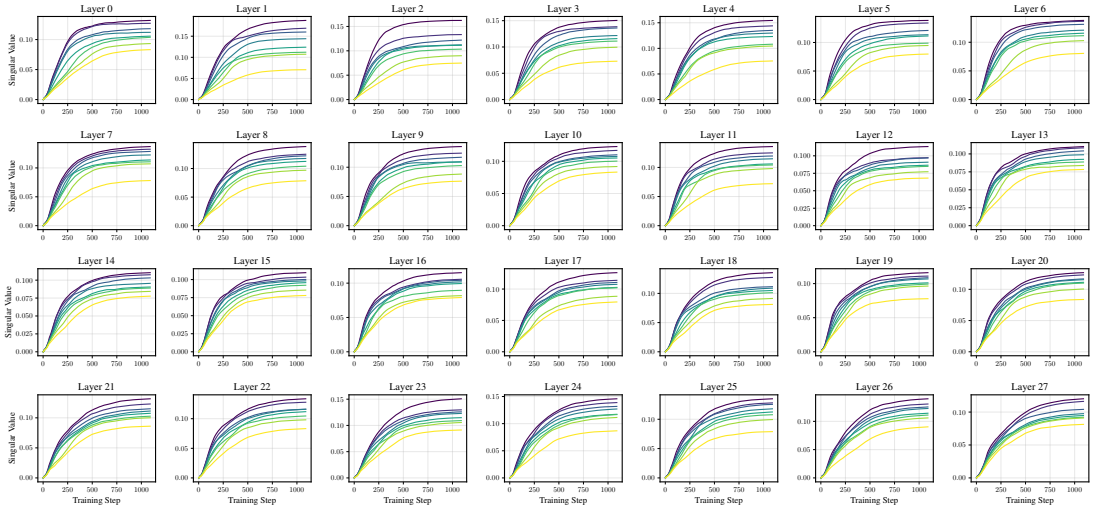
(a) Q matrix (b) V matrix

Figure 31: Singular values for all LoRA $\mathbf{AB}$ layers in RoBERTa-Base trained with SOAP and SST-2 dataset.

LLaMA-3.2-3B with SOAP optimizer.

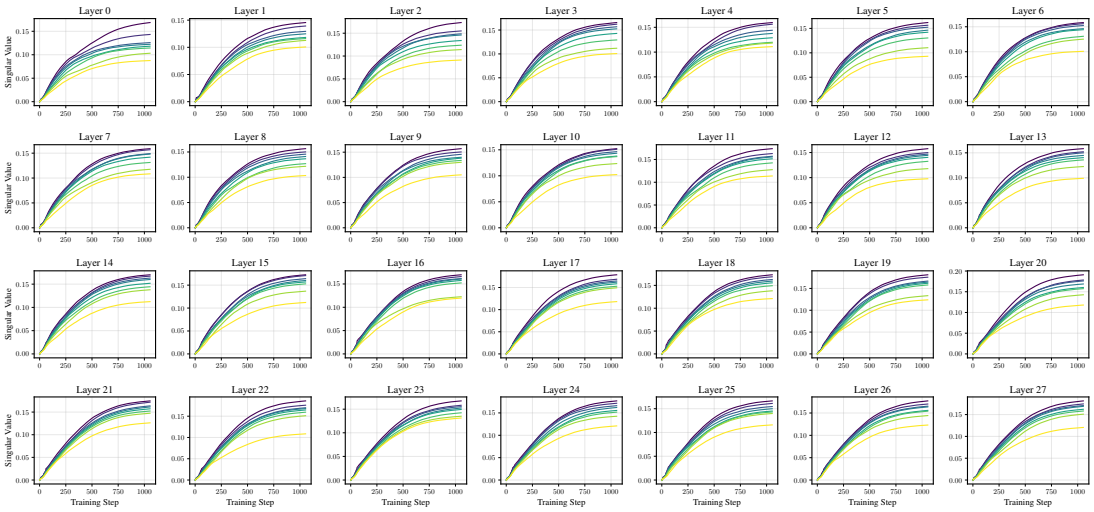


(a) Q matrix

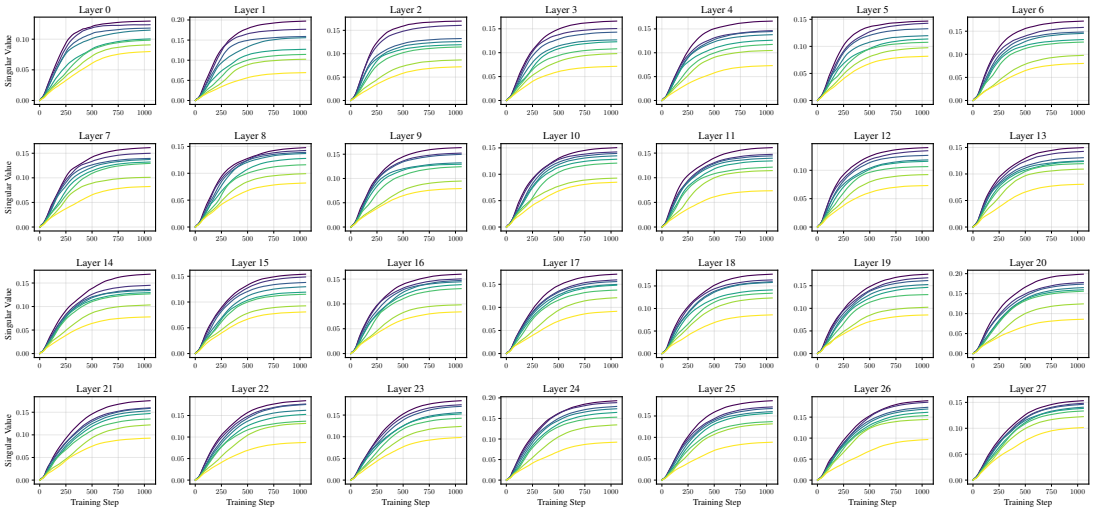


(b) V matrix

Figure 32: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with SOAP and Alpaca dataset.

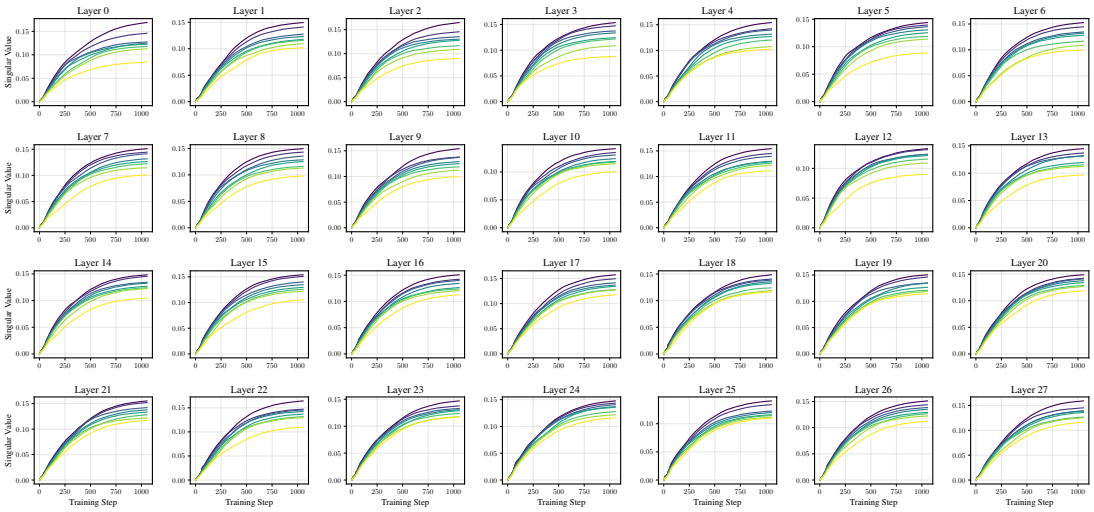


(a) Q matrix

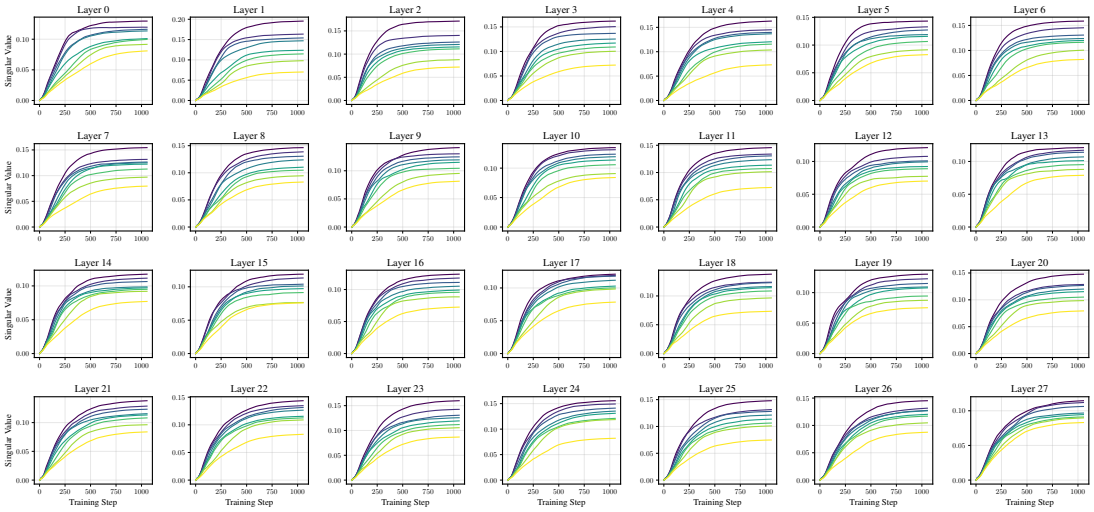


(b) V matrix

Figure 33: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with SOAP and OpenHermes 2.5 dataset.



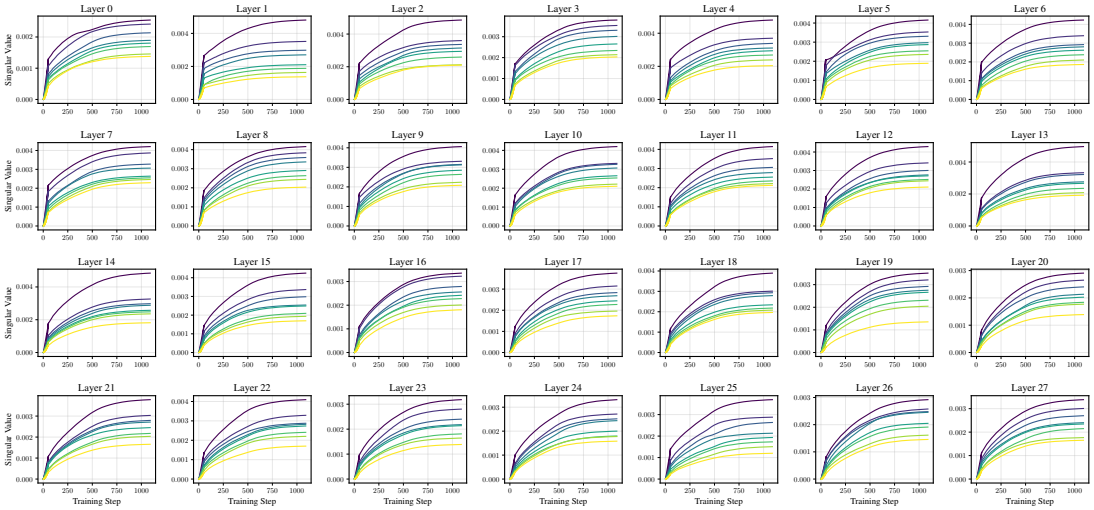
(a) Q matrix



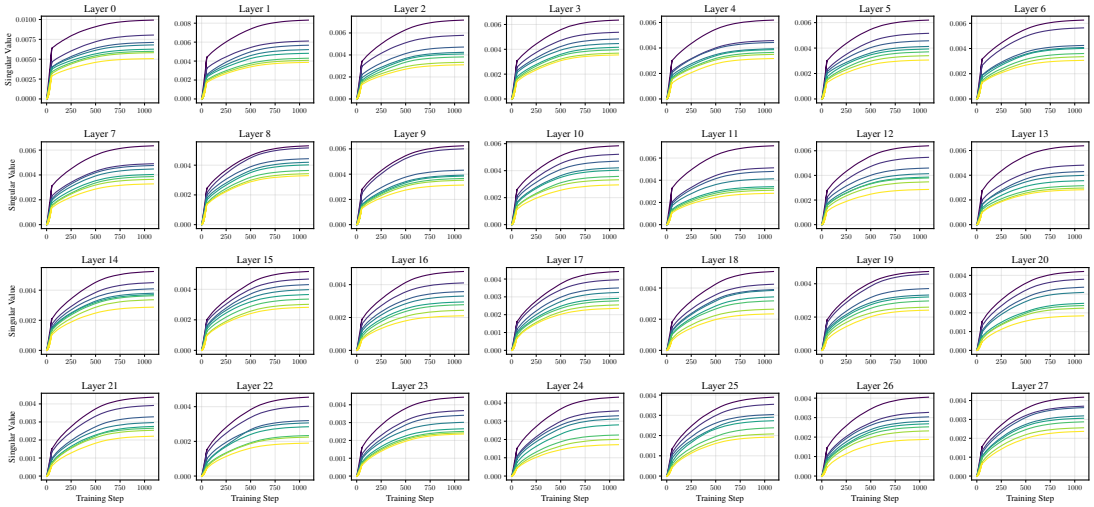
(b) V matrix

Figure 34: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with SOAP and UltraChat dataset.

LLaMA-3.2-3B with Shampoo optimizer.

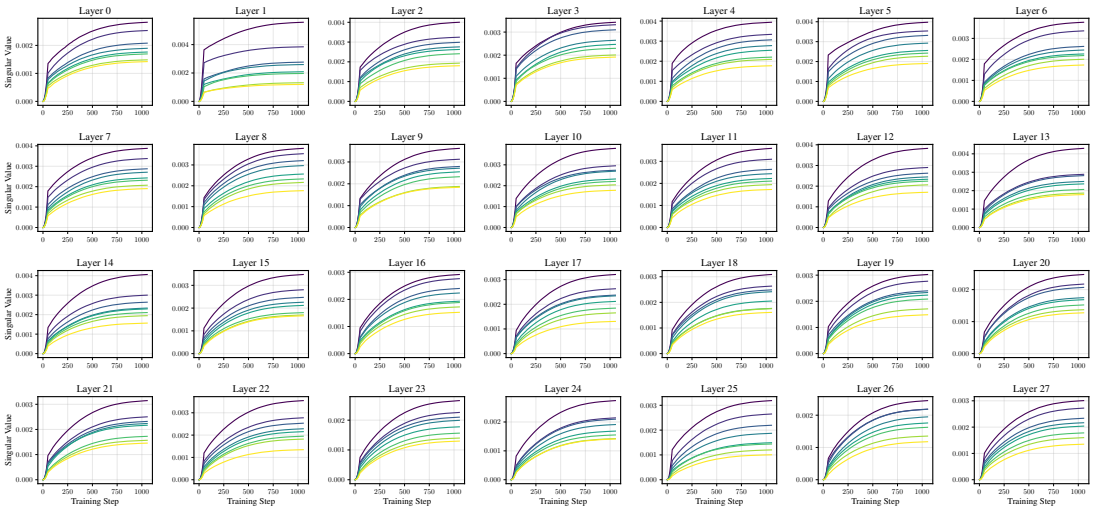


(a) Q matrix

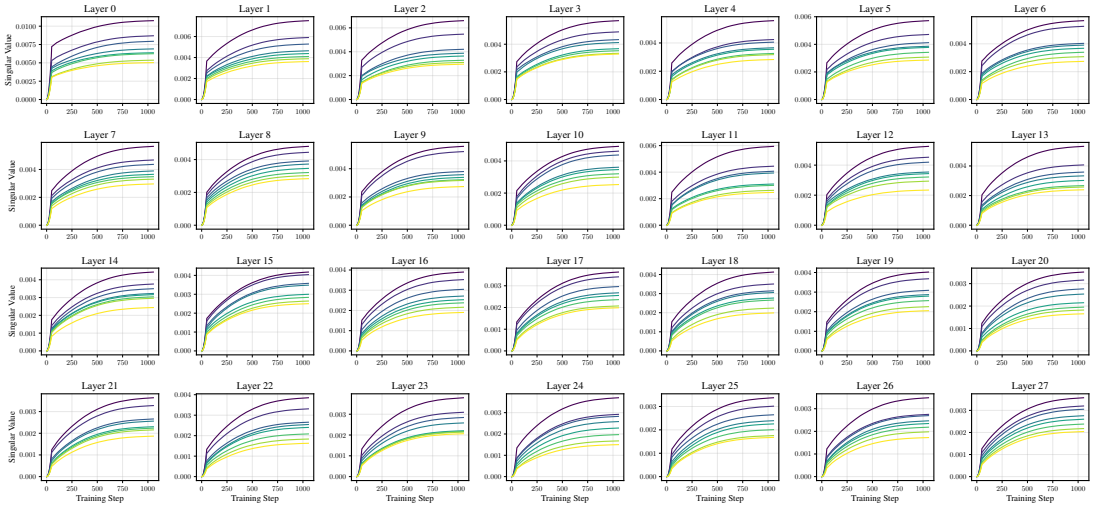


(b) V matrix

Figure 35: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Shampoo and Alpaca dataset.

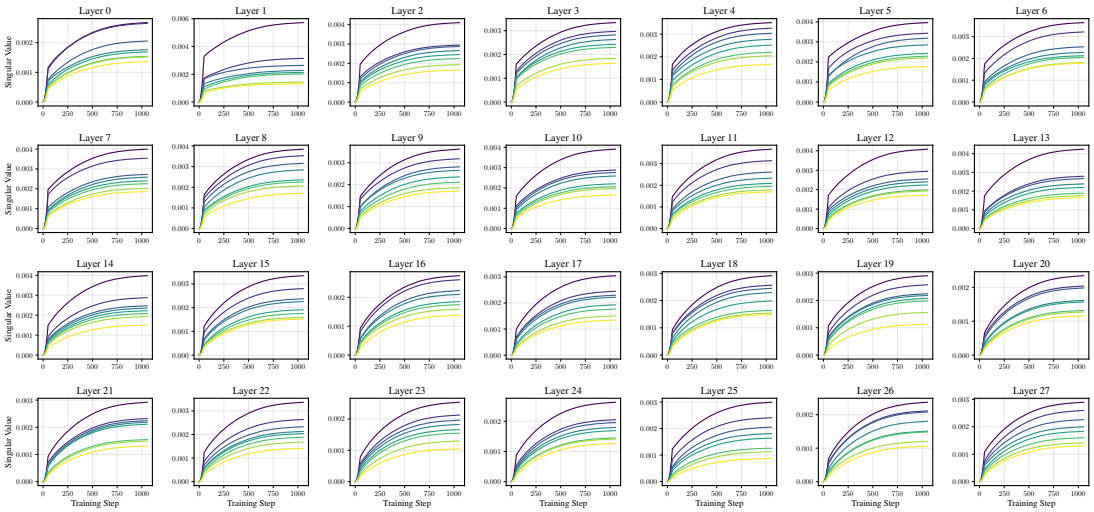


(a) Q matrix

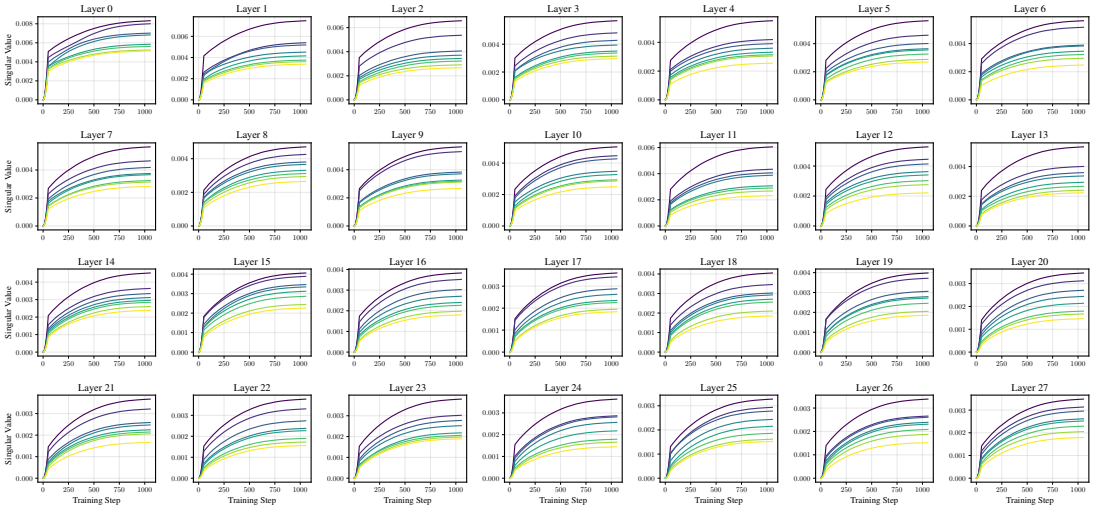


(b) V matrix

Figure 36: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Shampoo and OpenHermes 2.5 dataset.



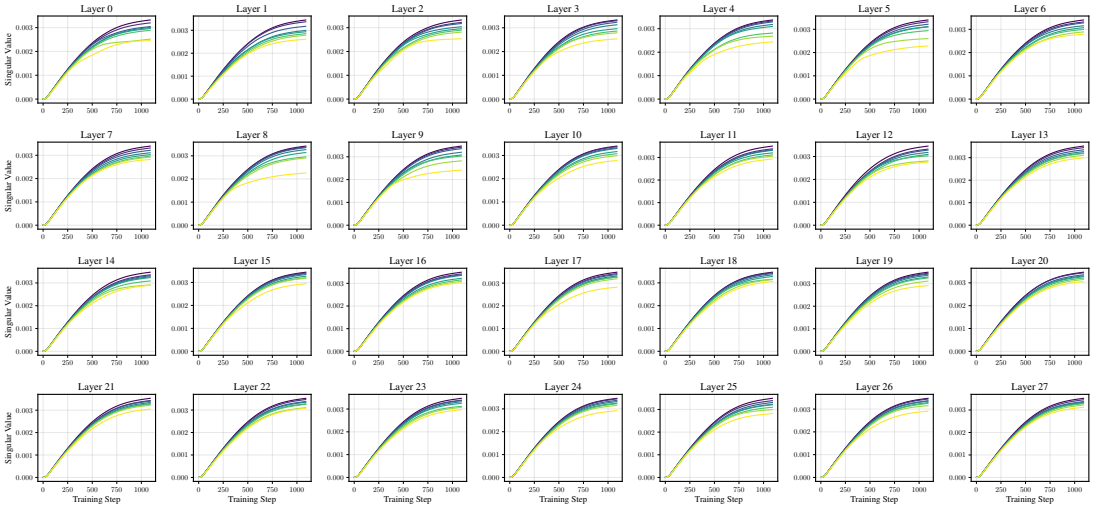
(a) Q matrix



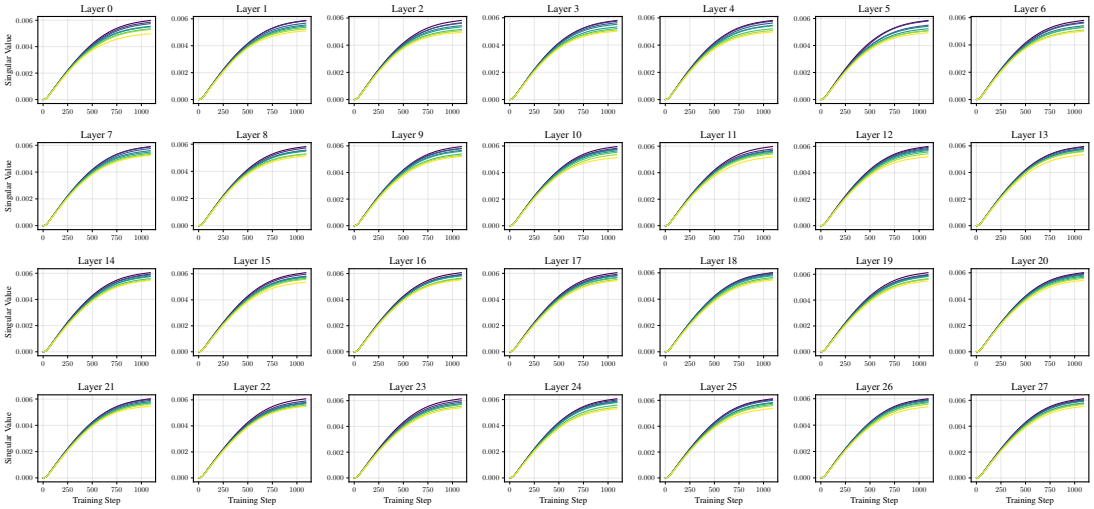
(b) V matrix

Figure 37: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Shampoo and UltraChat dataset.

LLaMA-3.2-3B with Dion optimizer.

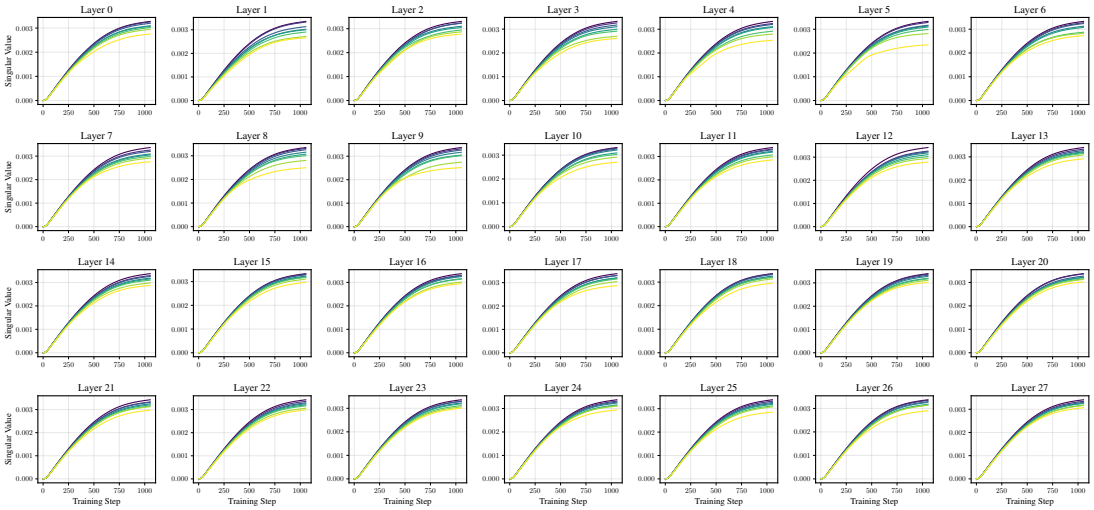


(a) Q matrix

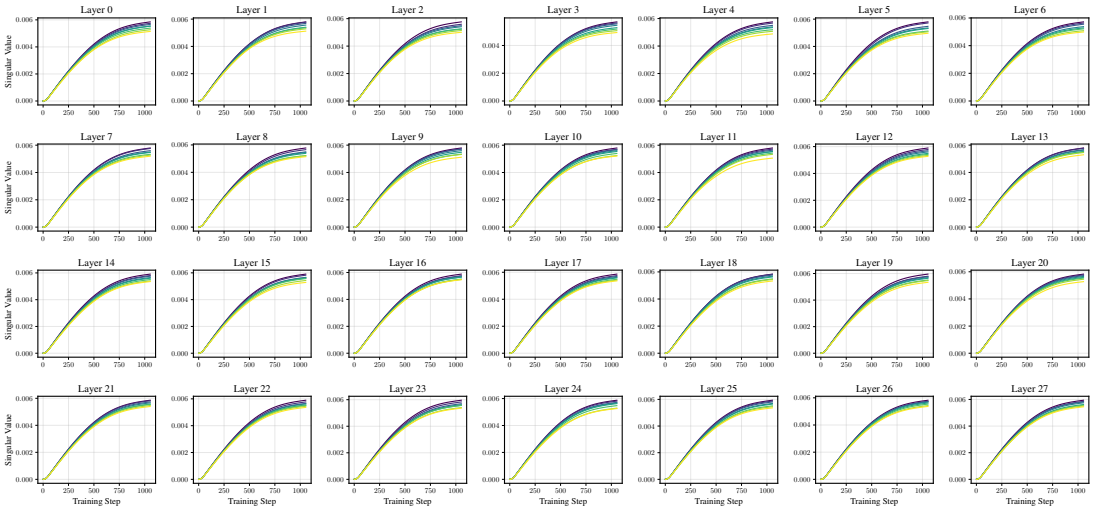


(b) V matrix

Figure 38: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Dion and Alpaca dataset.

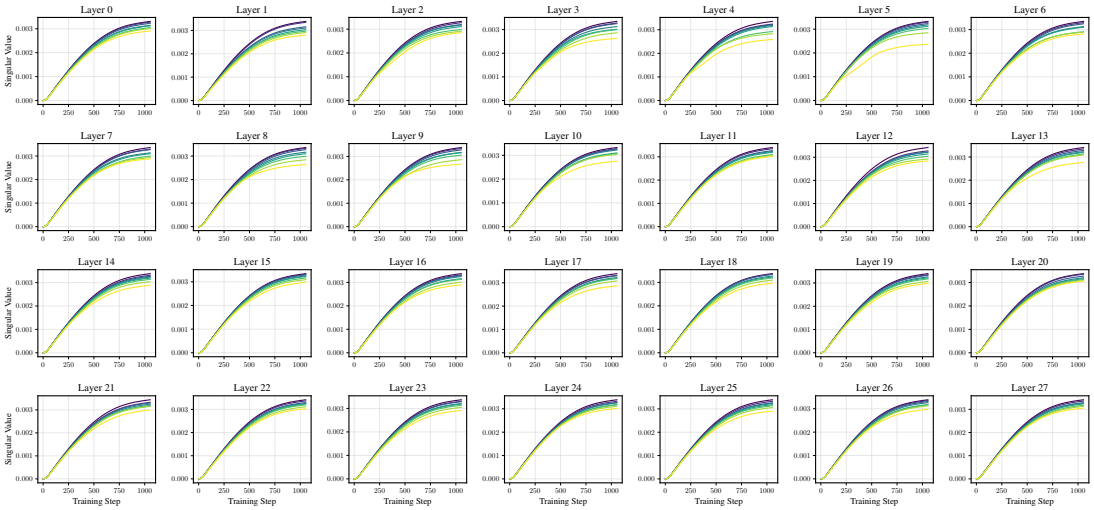


(a) Q matrix

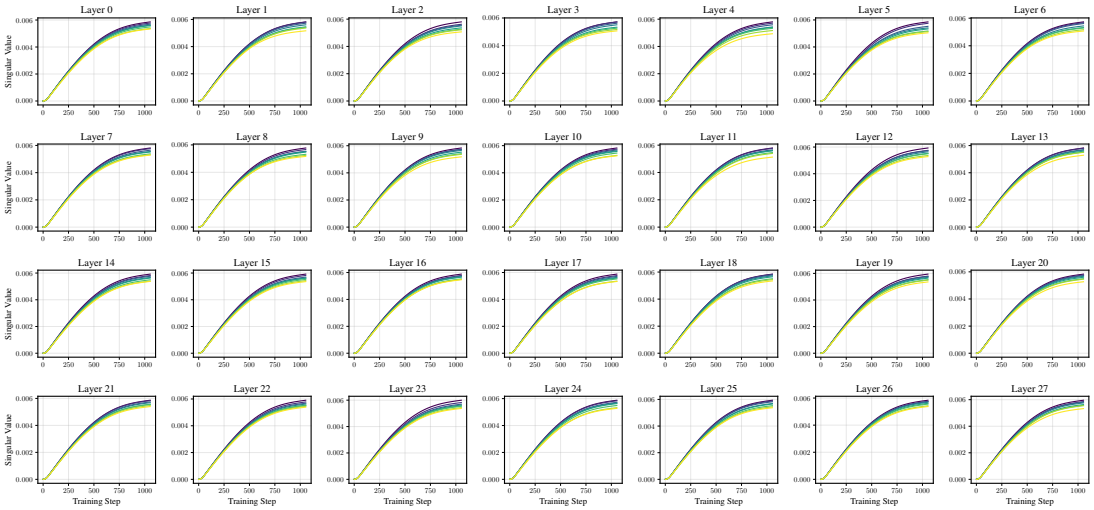


(b) V matrix

Figure 39: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Dion and OpenHermes 2.5 dataset.



(a) Q matrix



(b) V matrix

Figure 40: Singular values for all LoRA $\mathbf{AB}$ layers in LLaMA-3.2-3B trained with Dion and UltraChat dataset.