PoLAR: Polar-Decomposed Low-Rank Adapter Representation

Kai Lion [1] **Liang Zhang** [1] **Bingcong Li** [1 *] **Niao He** [1 *]

Abstract

We show that low-rank adaptation of large-scale models suffers from a low stable rank that is well below the linear algebraic rank of the subspace, degrading fine-tuning performance. To mitigate the underutilization of the allocated subspace, we propose PoLAR, a parameterization inspired by the polar decomposition that factorizes the low-rank update into two direction matrices constrained to Stiefel manifolds and an unconstrained scale matrix. Our theory shows that PoLAR yields an exponentially faster convergence rate on a canonical low-rank adaptation problem. Pairing the parameterization with Riemannian optimization leads to consistent gains on a commonsense reasoning benchmark with Llama-2-7B.

1. Introduction

Recent work attempts to overcome the low-rank constraint imposed by LoRA (Hu et al., 2022) while preserving its parameter-efficiency (Xia et al., 2024; Lialin et al., 2024; Zhao et al., 2024; Huang et al., 2025; Jiang et al., 2024). The underlying premise is that the low-rank nature of the adapter limits its expressiveness. However, this premise is at odds with recent theoretical results that LoRA can approximate any target transformer model reasonably well under mild assumptions (Zeng & Lee, 2024). Additionally, (Kalajdzievski, 2023) shows that raising the nominal rank does little to improve performance. Taken together, these findings suggest that the low-rank space offers sufficient expressiveness, but the classical low-rank adapter formulation struggles to fully utilize this potential.

Indeed, we observe comprehensive empirical evidence for this underutilization: when fine-tuning Llama-2-7B with LoRA, we find that the stable rank, a robust analogue of the matrix rank and measure of expressiveness, of the resulting update remains well below the linear algebraic rank. For some learned LoRA updates $\Delta\mathbf{W}$, the stable rank, defined as $\mathsf{sr}(\Delta\mathbf{W}) := \|\Delta\mathbf{W}\|_{\mathsf{F}}^2/\|\Delta\mathbf{W}\|_2^2$ (Rudelson & Vershynin, 2006), is as low as 1.06. This reveals that approximately a rank 1 subspace is utilized even if the LoRA rank is chosen as 32. Similar behaviors of low stable rank are consistently observed across layers and datasets; see Fig. 1a. Such deficiency results in a *diversity collapse* in the update directions among different neurons, where in extreme cases the updates for all neurons strongly align along a single direction (up to a sign flip); see Fig. 1b for an illustration and Fig. 1c for the directional diversity collapse when fine-tuning Llama-2-7B.

To address this pathology, we put forth PoLAR, a co-design of architecture and optimizer that mitigates the directional diversity collapse, as shown in Fig. 1d. On the *architecture side*, PoLAR facilitates effective exploitation of the linear algebraic rank by expressing the low-rank update as the product of two column-orthogonal direction matrices and a $r \times r$ scale matrix. On the *optimizer side*, we apply methods from Riemannian optimization (Boumal, 2023). Theoretically, we demonstrate that our co-design enables exponentially faster convergence than vanilla LoRA on a canonical problem. Our contribution is four-fold:

- Our empirical analysis demonstrates that the update matrices learned by LoRA have a stable rank far below their full linear algebraic rank, leading to a collapse in directional diversity and, in turn, preventing the adapters from fully realizing their expressiveness.
- We introduce PoLAR, an architecture-optimizer co-design that ensures diverse update directions by factoring the low-rank updates into column-orthogonal direction matrices and an arbitrary scale matrix. Riemannian optimization is then adopted for our PoLAR parameterization.
- On a matrix factorization prototype problem, we prove that our PoLAR parameterization achieves an *exponentially* faster convergence rate than vanilla LoRA.
- We evaluate PoLAR on a commonsense reasoning benchmark using Llama-2-7B and observe consistent performance gains.

Notation. Bold capital (lowercase) letters denote matrices (vectors); $(\cdot)^\top$ and $\|\cdot\|_{\mathsf{F}}$ refer to transpose and Frobenius

[*]Equal contribution [1]Department of Computer Science, ETH Zurich. Correspondence to: Kai Lion <kai.lion@inf.ethz.ch>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

norm of a matrix; $\|\cdot\|$ is the ℓ_2 (spectral) norm of a vector (matrix); $\sigma_i(\cdot)$ and $\lambda_i(\cdot)$ denote the i-th largest singular value and eigenvalue, respectively. For a matrix $\mathbf{X} \in \mathbb{R}^{r\times r}$, let $\mathsf{Skew}(\mathbf{X}) = \frac{1}{2}(\mathbf{X} - \mathbf{X}^\top)$ be its skew-symmetric part. The set of matrices with orthonormal columns, i.e., the Stiefel manifold, is denoted as $\mathsf{St}(m, r) := \{\mathbf{X} \in \mathbb{R}^{m\times r} : \mathbf{X}^\top\mathbf{X} = \mathbf{I}_r\}$. The set of $r \times r$ positive semi-definite (PSD) matrices is denoted as $\mathbb{S}^r_{\succeq 0} := \{\mathbf{X} \in \mathbb{R}^{r\times r} : \mathbf{X} \succeq 0\}$.

2. PoLAR: A Co-Design of Architecture and Optimizer

2.1. Overcoming Low Stable Rank with PoLAR

Given the pretrained weight (of a linear layer) $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$, LoRA learns an additive low-rank update $\Delta\mathbf{W} \in \mathbb{R}^{m\times n}$ with $\mathsf{rank}(\Delta\mathbf{W}) \le r$. The adapted weight is thus given by $\mathbf{W}_0 + \Delta\mathbf{W}$. In (Hu et al., 2022), the parameterization $\Delta\mathbf{W} = \mathbf{Z}_1\mathbf{Z}_2^\top$ with $\mathbf{Z}_1 \in \mathbb{R}^{m\times r}$ and $\mathbf{Z}_2 \in \mathbb{R}^{n\times r}$ is used.

While LoRA significantly enhances parameter efficiency as $(m + n)r \ll mn$, it turns out that it struggles to fully utilize the expressiveness of its parameterization. In particular, when fine-tuning Llama-2-7B with LoRA, we find that the stable rank $\mathsf{sr}(\Delta\mathbf{W})$ remains small on various datasets, oftentimes approaching 1 for many layers even with a reasonably large choice of $r = 32$; see Fig. 1a. Such a low stable rank translates to a loss in directional diversity of the neural updates. The directional diversity of an update matrix is measured by the average pairwise Euclidean distance of neurons when projected to the unit sphere. Note that this distance is within $[0, 2]$ with the lower and upper bounds attained by a pair of collinear neurons, pointing in the same or opposite directions, respectively. Consequently, a distribution of pairwise distances with most mass at the ends of the interval can be interpreted as evidence for low directional diversity. As observed in Fig. 1c, the LoRA update closely aligns along a single direction, with most neural updates being nearly collinear. Using the compact SVD $\Delta\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, it is possible to explain this observation and identify a low stable rank as a driver behind the lack of directional diversity. Here, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m\times r}$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{n\times r}$ have orthonormal columns and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ contains the singular values on its diagonal. Also, let $\mathbf{w}_i$ denote the i-th row of $\Delta\mathbf{W}$. With this notation, we can see that the direction of the LoRA update for the i-th neuron is approximated by:

$$\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} = \frac{1}{\sqrt{\sum_{j'=1}^r \sigma_{j'}^2\mathbf{U}_{ij'}^2}} \sum_{j=1}^r \sigma_j\mathbf{U}_{ij}\mathbf{v}_j^\top \overset{(a)}{\approx} \mathrm{sign}(\sigma_1\mathbf{U}_{i1})\mathbf{v}_1^\top, \forall i \in \{1, 2, \ldots, m\} \tag{1}$$

where (a) comes from $\mathsf{sr}(\Delta\mathbf{W}) \approx 1$. Equation (1) suggests that the weight update of all neurons tends to strongly align with the direction of the leading right singular vector up to a sign flip, causing the two-cluster pattern in Fig. 1c. Moreover, Fig. 1a demonstrates the wide applicability of this finding, as the LoRA updates across several datasets suffer from a very low stable rank. This pathology suggests that LoRA does not fully utilize its rank capacity, and we conjecture that a parameterization addressing this pathology increases the performance of LoRA. These insights lead to the following desiderata: we wish to learn $\mathcal{O}(r)$ roughly orthogonal directions whose contributions to the unit-norm neural update are roughly balanced, avoiding the collapse of directional diversity encountered in vanilla LoRA.

To this end, we advocate to incorporate orthogonality directly into the architecture. This can be achieved with the polar decomposition; see Appendix A.1 for a brief review. Applied to each low-rank factor, this decomposition yields $\mathbf{Z}_1 = \mathbf{X}\mathbf{\Theta}_1$ and $\mathbf{Z}_2 = \mathbf{Y}\mathbf{\Theta}_2$, separating them into PSD scale and column-orthogonal direction components with $\mathbf{X} \in \mathsf{St}(m, r)$, $\mathbf{Y} \in \mathsf{St}(n, r)$ and $\mathbf{\Theta}_i \in \mathbb{S}^r_{\succeq 0}$ for $i \in \{1, 2\}$. The desirable orthogonality is thus naturally enforced through the manifold structure of $\mathbf{X}$ and $\mathbf{Y}$. Moreover, rather than relying on two individual scale matrices, it is more convenient to learn a joint $\mathbf{\Theta} \in \mathbb{R}^{r\times r}$ matrix for the overall update, which amounts to merging the product $\mathbf{\Theta}_1\mathbf{\Theta}_2^\top$. These considerations give rise to our **Po**lar-decomposed **L**ow-rank **A**dapter **R**epresentation (PoLAR):

$$\Delta\mathbf{W} = \mathbf{X}\mathbf{\Theta}\mathbf{Y}^\top \tag{2}$$

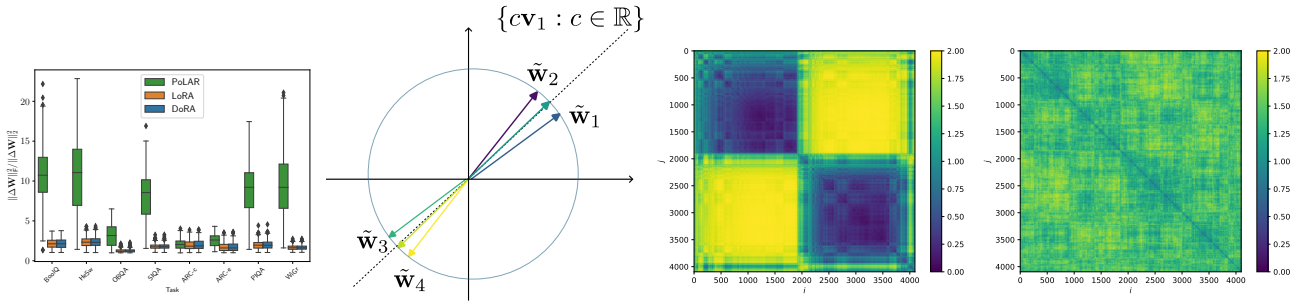
with $\mathbf{X} \in \mathsf{St}(m, r)$, $\mathbf{Y} \in \mathsf{St}(n, r)$, and $\mathbf{\Theta} \in \mathbb{R}^{r\times r}$. As a byproduct, PoLAR admits a natural interpretation in terms of a direction–magnitude decomposition. However, unlike alternative decompositions, such as the column-wise weight normalization used in DoRA (Salimans & Kingma, 2016; Liu et al., 2024), PoLAR enforces orthogonality, substantially increasing the stable rank (see Fig. 1a) and generating low-rank updates with more competitive performance.

2.2. Faster Optimization with PoLAR Parameterization

We now compare the convergence of PoLAR and LoRA on the problem of learning a low-rank adapter for a single layer with whitened data. As discussed in (Arora et al., 2018; Zhang & Pilanci, 2024; Li et al., 2024) and Appendix A.4, applying LoRA in this case is equivalent to a matrix factorization under the asymmetric Burer-Monteiro (BM) parameterization (Burer & Monteiro, 2003):

$$\min_{\mathbf{Z}_1\in\mathbb{R}^{m\times r}, \mathbf{Z}_2\in\mathbb{R}^{n\times r}} \frac{1}{2}\|\mathbf{Z}_1\mathbf{Z}_2^\top - \mathbf{A}\|_{\mathsf{F}}^2. \tag{3}$$

In problem (3), the matrix to be factorized (or the target matrix of LoRA) is denoted as $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{Z}_1, \mathbf{Z}_2$ represent the LoRA weights. In light of the low-rank setting,



(a) $\mathsf{sr}(\Delta\mathbf{W})$ of Llama-2-7B (b) diversity collapse (DC) (c) DC in LoRA (d) PoLAR solves DC

Figure 1: *(a)* $\mathsf{sr}(\Delta\mathbf{W})$ of Llama-2-7B low-rank updates fine-tuned on commonsense tasks with rank 32. *(b)* Illustration of directional diversity collapse (DC) of $\tilde{\mathbf{w}}_i = \mathbf{w}_i/\|\mathbf{w}_i\|_2$ where $\mathbf{w}_i$ denotes the i-th row of low-rank update $\Delta\mathbf{W}$. *(c) and (d)* Diversity of update directions of LoRA and PoLAR for a Llama-2-7B down-projection layer on dataset Social-IQA, respectively. Each pixel shows $\|\tilde{\mathbf{w}}_i - \tilde{\mathbf{w}}_j\|_2$; rows and columns are rearranged to reveal cluster patterns in both plots. Emergence of a cluster pattern is evidence for DC. The algebraic rank is 32 for both methods, yet the stable rank is 1.06 and 5.49 for LoRA and PoLAR, respectively. See also Section 2.1.

$r \ll \min\{m, n\}$ is assumed. Problem (3) has been widely adopted as a testbed for developing optimization schemes for LoRA; e.g., (Zhang & Pilanci, 2024) or (Li et al., 2024).

Our goal in this subsection is to understand the optimization dynamics of our PoLAR parameterization applied to the same one-layer setting, yielding the problem below

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{Y},\mathbf{\Theta}} \frac{1}{2}\|\mathbf{X}\mathbf{\Theta}\mathbf{Y}^\top - \mathbf{A}\|_{\mathsf{F}}^2, \\ &\text{s.t.} \quad \mathbf{X} \in \mathsf{St}(m, r),\ \mathbf{Y} \in \mathsf{St}(n, r),\ \mathbf{\Theta} \in \mathbb{R}^{r\times r}. \end{aligned} \tag{4}$$

Considering the sufficient expressiveness of LoRA (Zeng & Lee, 2024), we focus on the overparameterized regime for both (3) and (4), where $r > r_A := \mathsf{rank}(\mathbf{A})$. In this setting, zero loss is attainable. Let the compact SVD of $\mathbf{A}$ be $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{m\times r_A}$, $\mathbf{V} \in \mathbb{R}^{n\times r_A}$ and diagonal $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$. Without loss of generality, we assume $\sigma_1(\mathbf{\Sigma}) = 1$ and $\sigma_{r_A}(\mathbf{\Sigma}) = 1/\kappa$ where κ is the condition number. We also assume $m \geq n$, as one can transpose $\mathbf{A}$ if necessary.

Note that the semi-orthogonal low-rank factors live on Stiefel manifolds, requiring a treatment with Riemannian gradient descent (RGD). For technical simplicity, we consider a procedure that alternates between updating $\mathbf{\Theta}_t$ and $(\mathbf{X}_t, \mathbf{Y}_t)$. At iteration t, it starts by finding $\mathbf{\Theta}_t$ with gradient descent (GD) using learning rate $\gamma > 0$, i.e.,

$$\mathbf{\Theta}_t = (1-\gamma)\mathbf{\Theta}_{t-1} + \gamma\mathbf{X}_t^\top\mathbf{A}\mathbf{Y}_t. \tag{5a}$$

Setting $\gamma = 1$ significantly simplifies our analysis. With this value and the updated matrix $\mathbf{\Theta}_t$, the Riemannian gradient of $\mathbf{X}_t$ can be obtained via $\mathbf{E}_t = -(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{Y}_t\mathbf{\Theta}_t^\top$. Further involving polar retraction to remain on the manifold, the RGD update on $\mathbf{X}_t$ is given by

$$\mathbf{X}_{t+1} = (\mathbf{X}_t - \eta\mathbf{E}_t)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1/2}. \tag{5b}$$

We summarize the resulting RGD procedure in Alg. 2 (in Appendix) and establish its global convergence:

Theorem 2.1 (Global Convergence)**.** *Suppose that* $r_A \leq \frac{n}{2}$. *Let* $\rho := \min\{\frac{1}{m}, \frac{(r-r_A)^2}{rm}\}$. *W.h.p. over the random initialization of* $\mathbf{X}_0$ *and* $\mathbf{Y}_0$, *choosing the learning rates* $\eta = \mathcal{O}(\frac{(r-r_A)^2\rho}{r^2\kappa^2 m})$ *and* $\gamma = 1$, *The update procedure ensures* $\frac{1}{2}\|\mathbf{X}_T\mathbf{\Theta}_T\mathbf{Y}_T^\top - \mathbf{A}\|_{\mathsf{F}}^2 \leq \epsilon$ *for all* $T \geq \mathcal{O}\big(\frac{m^2r^3r_A\kappa^4}{\rho^2(r-r_A)^4} + \frac{m^2r^3\kappa^4}{\rho(r-r_A)^4}\log\frac{1}{\epsilon}\big)$.

Our rate compares favorably to previous results of GD in the overparameterized regime. In particular, (Xiong et al., 2024) show that overparameterization slows down GD, leading to an undesirable κ dependence with $\mathcal{O}\big(\max\{\kappa^{15}, \kappa^\kappa\}\log(1/\epsilon)\big)$. Our rates in Theorem 2.1 exponentially improve the κ dependence to a quartic one.

Our choice of an *unconstrained* $\mathbf{\Theta} \in \mathbb{R}^{r\times r}$ plays a crucial role for convergence. Empirical evidence in (Mishra et al., 2013) shows that substituting $\mathbf{\Theta}$ by a diagonal matrix $\mathbf{\Theta}^{\mathsf{d}}$, as done in AdaLoRA (Zhang et al., 2023), can adversely affect convergence, potentially because of the presence of non-strict saddles in the loss landscape (Levin et al., 2024). These spurious stationary points can be removed by a parameterization with positive-definite $\mathbf{\Theta}^{\mathsf{s}}$ (Levin et al., 2024). Our PoLAR parameterization poses no constraints on $\mathbf{\Theta}$. It thus avoids computational overheads associated with enforcing positive-definiteness (e.g., matrix exponentials), yet still ensures global convergence.

3. Practical PoLAR for Scalable Fine-Tuning

Although PoLAR increases expressiveness and accelerates convergence, optimizing under the manifold constraint of (2) with standard feasible methods does not scale to tasks such

Table 1: Performance on commonsense reasoning tasks with Llama-2-7B using PoLAR for different ranks in a single-task setup. HeSw refers to HellaSwag and WiGr to WinoGrande.

Rank	Adapter	BoolQ	PIQA	SIQA	HeSw	WiGr	ARC-e	ARC-c	OBQA	**Avg.**
4	LoRA	87.16	81.01	58.85	82.36	74.35	81.90	57.68	56.80	72.51
	DoRA	87.22	80.30	58.96	82.39	75.22	81.69	57.85	56.80	72.55
	PoLAR	87.49	82.59	59.31	81.23	81.77	81.61	56.31	55.80	**73.26**
32	LoRA	87.89	81.56	59.06	82.51	72.61	82.37	56.83	54.60	72.18
	DoRA	87.61	81.45	58.70	82.50	74.43	82.28	57.17	55.60	72.47
	PoLAR	88.13	82.64	60.03	83.12	82.00	81.99	56.14	55.60	**73.71**

Algorithm 1 PoLAR Fine-tuning

Input: Parameterize via (2); Initialize $\mathbf{X}_0, \mathbf{Y}_0$ uniformly random from Stiefel manifolds, set $\mathbf{\Theta}_0 = \mathbf{0}$, and denote λ regularization strength, ρ_t stateful gradient transformation (e.g., Adam)
for $t = 0, \dots, T-1$ **do**
$\mathbf{\Gamma}(\mathbf{X}_t) \leftarrow \psi(\mathbf{X}_t)\mathbf{X}_t + \lambda\nabla\mathcal{N}(\mathbf{X}_t)$
$\mathbf{\Gamma}(\mathbf{Y}_t) \leftarrow \psi(\mathbf{Y}_t)\mathbf{Y}_t + \lambda\nabla\mathcal{N}(\mathbf{Y}_t)$
$\mathbf{X}_{t+1} \leftarrow \mathbf{X}_t - \eta_t\rho_t(\mathbf{\Gamma}(\mathbf{X}_t))$
$\mathbf{Y}_{t+1} \leftarrow \mathbf{Y}_t - \eta_t\rho_t(\mathbf{\Gamma}(\mathbf{Y}_t))$
$\mathbf{\Theta}_{t+1} \leftarrow \mathbf{\Theta}_t - \eta_t\rho_t(\nabla_{\mathbf{\Theta}_t}\mathcal{L}(\mathbf{X}_t, \mathbf{\Theta}_t, \mathbf{Y}_t))$
end for

as LLM fine-tuning. Every retraction back onto the Stiefel manifold requires a matrix inversion or SVD (see (5b)), operations whose sequential nature limits GPU parallelism and becomes a runtime bottleneck (Sun et al., 2024). To sidestep these issues, we take inspiration from the recently proposed *landing algorithm* which completely eschews retractions, producing iterates that are not necessarily on the manifold, but provably *land* on it as training proceeds (Ablin & Peyré, 2022; Gao et al., 2022). At iteration t, we update $\mathbf{X}_t$ and $\mathbf{Y}_t$ with the landing field. Taking the update on $\mathbf{X}_t$ as an example, we use

$$\mathbf{\Gamma}(\mathbf{X}_t) := \underbrace{\psi(\mathbf{X}_t)\mathbf{X}_t}_{\text{Riemannian grad.}} + \underbrace{\lambda\nabla\mathcal{N}(\mathbf{X}_t)}_{\text{Infeasibility penalty}} \tag{6}$$

as a drop-in replacement for the Riemannian update described in Section 2.2. The first component in (6) is the standard Riemannian gradient for matrices on the Stiefel manifold with $\psi(\mathbf{X}) := \mathsf{Skew}(\nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}, \mathbf{\Theta}, \mathbf{Y})\mathbf{X}^\top)$. The second component in (6) is given by the gradient of the infeasibility penalty, $\mathcal{N}(\mathbf{X}) := \|\mathbf{X}^\top\mathbf{X} - \mathbf{I}_r\|_{\mathsf{F}}^2$, which attracts the iterate towards the Stiefel manifold, making retraction obsolete. The landing field computation involves only matrix multiplications, eliminating the need for sequential retraction routines that do not map well to GPU parallelism. The complete PoLAR fine-tuning procedure is summarized in Alg. 1 where $\lambda > 0$ is a tunable hyperparameter.

4. Experiments

We now evaluate the PoLAR fine-tuning procedure on commonsense reasoning tasks which test how well LLMs can mimic human-like understanding. We report the accuracy based on multiple-choice log-likelihood evaluation. PoLAR delivers the highest mean accuracy across tasks, outperforming both LoRA and DoRA (Table 1). Whereas the gains of DoRA and LoRA appear to be mostly flat going from $r = 4$ to $r = 32$, PoLAR's accuracy increases with larger rank. This is consistent with our conjecture that PoLAR counteracts directional-diversity collapse by exploiting the allocated rank more effectively. In Table 2 of the Appendix, we compare our method PoLAR with a procedure that brings our method closer to AdaLoRA (Zhang et al., 2023), which uses a SVD-type parameterization. We observe that the PoLAR parameterization based on the Riemannian gradient outperforms the diagonal parameterization based on the Euclidean gradient.

5. Conclusion

Low-rank adaptation has become the workhorse for efficient fine-tuning, yet our analysis revealed that the classical BM parameterization in LoRA often underutilizes its allocated subspace: the stable rank of the learned updates collapses, limiting expressive power. Building on this empirical finding, as well as theoretical insights from a canonical low-rank optimization problem, we introduced PoLAR, a reparameterization inspired by the polar decomposition that is coupled with a landing field optimization procedure. We empirically show that PoLAR delivers superior performance on a commonsense reasoning benchmark with Llama-2-7B, providing evidence that maintaining a higher stable rank translates into richer, more task-aligned updates. In practice, PoLAR's reliance on nothing more than matrix multiplications implies that it maps cleanly onto GPU hardware, offering a drop-in replacement for existing LoRA pipelines.

References

Ablin, P. and Peyré, G. Fast and Accurate Optimization on the Orthogonal Manifold without Retraction. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, January 2022.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2018.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J. Z., Foster, C., Hsu, J., Jaiswal, M., Lee, W. Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K. A., Winata, G. I., Yvon, F., and Zou, A. Lessons from the Trenches on Reproducible Evaluation of Language Models, May 2024. arXiv:2405.14782.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proc. AAAI Conf. Artif. Intel.*, November 2019.

Björck, A. and Golub, G. H. Numerical Methods for Computing Angles between Linear Subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Boumal, N. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, February 2003. ISSN 1436-4646. doi: 10.1007/s10107-002-0352-8. URL https://doi.org/10.1007/s10107-002-0352-8.

Chikuse, Y. *Statistics on Special Manifolds*, volume 174. Springer Science & Business Media, 2012.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proc. Conf. Assoc. Comput. Linguist. Meet.*, May 2019.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. arXiv:1803.05457.

Gao, B., Vary, S., Ablin, P., and Absil, P.-A. Optimization Flows Landing on the Stiefel Manifold, July 2022. arXiv:2202.09058.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. JHU press, 2013.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

Huang, Q., Ko, T., Zhuang, Z., Tang, L., and Zhang, Y. HiRA: Parameter-Efficient Hadamard High-Rank Adaptation for Large Language Models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2025.

Jiang, T., Huang, S., Luo, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., Zhang, Q., Wang, D., and Zhuang, F. MoRA: High-Rank Updating for Parameter-Efficient Fine-Tuning, May 2024. arXiv:2405.12130.

Kalajdzievski, D. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA, November 2023. arXiv:2312.03732.

Levin, E., Kileel, J., and Boumal, N. The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes. *Mathematical Programming*, March 2024.

Li, B., Zhang, L., Mokhtari, A., and He, N. On the crucial role of initialization for matrix factorization. *arXiv preprint arXiv:2410.18965*, 2024.

Lialin, V., Muckatira, S., Shivagunde, N., and Rumshisky, A. ReLoRA: High-Rank Training Through Low-Rank Updates. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 32100–32121, July 2024.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Mishra, B., Meyer, G., Bonnabel, S., and Sepulchre, R. Fixed-Rank Matrix Factorizations and Riemannian Low-Rank Optimization. *Computational Statistics*, 29:591–621, April 2013.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.

Rudelson, M. and Vershynin, R. Sampling from Large Matrices: An Approach through Geometric Functional Analysis. *Journal of the ACM*, 54(4), December 2006.

Rudelson, M. and Vershynin, R. Smallest Singular Value of a Random Rectangular Matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM*, 64(9):99–106, 2019.

Salimans, T. and Kingma, D. P. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense Reasoning about Social Interactions. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4463–4473, November 2019.

Sun, Y., Chen, S., Garcia, A., and Shahrampour, S. Retraction-Free Decentralized Non-convex Optimization with Orthogonal Constraints, December 2024. arXiv:2405.11590.

Vershynin, R. Introduction to the Non-Asymptotic Analysis of Random Matrices. *arXiv:1011.3027*, 2010.

Xia, W., Qin, C., and Hazan, E. Chain of LoRA: Efficient Fine-tuning of Language Models via Residual Learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

Xiong, N., Ding, L., and Du, S. S. How Over-Parameterization Slows Down Gradient Descent in Matrix Sensing: The Curses of Symmetry and Initialization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proc. Conf. Assoc. Comput. Linguist. Meet.*, May 2019.

Zeng, Y. and Lee, K. The Expressive Power of Low-Rank Adaptation. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

Zhang, F. and Pilanci, M. Riemannian Preconditioned LoRA for Fine-Tuning Foundation Models. In *Proc. Int. Conf. on Machine Learning (ICML)*, June 2024.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

A. Useful Facts

A.1. Polar Decomposition

We provide the definition for the polar decomposition below and refer the interested reader to Section 9.4.3 of (Golub & Van Loan, 2013) for further details.

Definition A.1 (Polar Decomposition)**.** The (right) polar decomposition of a matrix $\mathbf{X} \in \mathbb{R}^{m\times r}$ with $m \geq r$ is defined as $\mathbf{X} = \mathbf{U}\mathbf{P}$ where $\mathbf{U} \in \mathbb{R}^{m\times r}$ has orthonormal columns and $\mathbf{P} \in \mathbb{R}^{r\times r}$ is positive semi-definite.

When $r = 1$, the polar decomposition in Definition A.1 reduces to the familiar magnitude-direction decomposition of vectors. In general, the polar decomposition can be viewed as an extension to matrices, where $\mathbf{U}$ represents the directional component and $\mathbf{P}$ captures the magnitude.

A.2. Angles Between Subspaces

Angles between two subspaces are known as principal angles. Suppose that $n \geq p$ and $n \geq q$, and let $\mathcal{U}$ and $\mathcal{V}$ be two linear subspaces of dimension $n \times p$ and $n \times q$. Then the principle angles $\theta_i \in [0, \frac{\pi}{2}], \forall i \leq \min\{p, q\}$ is defined as

$$\theta_1 = \max\left\{ \arccos \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{u}\|\|\mathbf{v}\|} \middle| \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V} \right\}$$
$$\theta_i = \max\left\{ \arccos \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\|\mathbf{u}\|\|\mathbf{v}\|} \middle| \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}, \mathbf{u} \perp \mathbf{u}_j, \mathbf{v} \perp \mathbf{v}_j, \forall j \in 1, \ldots, i-1 \right\}, \forall i \neq 1.$$

There is a well-known relation between principal angles and SVD. Let $\mathbf{U}$ and $\mathbf{V}$ be basis of $\mathcal{U}$ and $\mathcal{V}$ respectively. It can be seen that all the singular values of $\mathbf{U}^\top\mathbf{V}$ belong to $[0, 1]$. Moreover, the principle angles defined above are just the arc-cosine of these singular values (Björck & Golub, 1973). For convenience of this work, *we refer to the singular values of* $\mathbf{U}^\top\mathbf{V}$ *as principal angles (instead of the arc-cosine of them).* If the basis $\mathbf{U}$ and $\mathbf{V}$ are both from $\mathsf{St}(m, r)$, we sometimes use the term "alignment", where we say $\mathbf{U}$ and $\mathbf{V}$ are aligned if all the singular values of $\mathbf{U}^\top\mathbf{V}$ are 1; or in other words, they share the same column space.

The principal angles are also related to the geodesic distance on Grassmann manifolds. Oftentimes, people use the term *chordal distance* to refer to $d(\mathbf{U}, \mathbf{V}) = \sqrt{\sum_i \sin^2\theta_i}$, where θ_i are principal angles between two subspaces spanned by $\mathbf{U}$ and $\mathbf{V}$. The square of the chordal distance coincides with our notation $\mathsf{Tr}(\mathbf{I} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$, where $\mathbf{\Phi}_t$ is defined in Section 2.2.

A.3. Other Useful Lemmas

Lemma A.2. *Given a PSD matrix* $\mathbf{A}$*, we have that* $(\mathbf{I} + \mathbf{A})^{-1} \succeq \mathbf{I} - \mathbf{A}$.

Proof. Simply diagonalizing the LHS and RHS, and using $1/(1 + \lambda) \geq 1 - \lambda, \forall\lambda \geq 0$ gives the result. □

Lemma A.3. *Suppose that* $\mathbf{X} \in \mathsf{St}(m, r)$*,* $\mathbf{U} \in \mathsf{St}(m, r_A)$ *and* $r_A \leq r$*. Let* $\mathbf{U}_\perp \in \mathbb{R}^{m\times(m-r_A)}$ *be the orthogonal complement of* $\mathbf{U}$*. Denote* $\mathbf{\Phi} = \mathbf{U}^\top\mathbf{X}$ *and* $\mathbf{\Omega} = \mathbf{U}_\perp^\top\mathbf{X}$*. It is guaranteed that* $\sigma_i^2(\mathbf{\Phi}) + \sigma_i^2(\mathbf{\Omega}) = 1$ *holds for* $i \in \{1, 2, \ldots, r\}$.

Proof. We have that

$$\begin{aligned}\mathbf{I}_r = \mathbf{X}^\top\mathbf{X} = \mathbf{X}^\top\mathbf{I}_m\mathbf{X} &= \mathbf{X}^\top[\mathbf{U}, \mathbf{U}_\perp]\begin{bmatrix}\mathbf{U}^\top \\ \mathbf{U}_\perp^\top\end{bmatrix}\mathbf{X} \\ &= \mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Omega}^\top\mathbf{\Omega}.\end{aligned} \tag{7}$$

Equation (7) indicates that $\mathbf{\Psi}^\top\mathbf{\Psi}$ and $\mathbf{\Phi}^\top\mathbf{\Phi}$ commute, i.e.,

$$\begin{aligned}(\mathbf{\Phi}^\top\mathbf{\Phi})(\mathbf{\Omega}^\top\mathbf{\Omega}) = (\mathbf{\Phi}^\top\mathbf{\Phi})(\mathbf{I}_r - \mathbf{\Phi}^\top\mathbf{\Phi}) &= \mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{\Phi} \\ &= (\mathbf{I}_r - \mathbf{\Phi}^\top\mathbf{\Phi})(\mathbf{\Phi}^\top\mathbf{\Phi}) = (\mathbf{\Omega}^\top\mathbf{\Omega})(\mathbf{\Phi}^\top\mathbf{\Phi}).\end{aligned}$$

The commutativity shows that the eigenspaces of $\mathbf{\Phi}^\top\mathbf{\Phi}$ and $\mathbf{\Omega}^\top\mathbf{\Omega}$ coincide. As a result, we have again from (7) that $\sigma_i^2(\mathbf{\Phi}) + \sigma_i^2(\mathbf{\Omega}) = 1$ for $i \in \{1, 2, \ldots, r\}$. The proof is thus completed. □

Lemma A.4. *Suppose that* $\mathbf{P}$ *and* $\mathbf{Q}$ *are* $m \times m$ *diagonal matrices, and their diagonal entries are non-negative. Let* $\mathbf{S}$ *be a PD matrix of* $m \times m$ *with smallest eigenvalue* $\lambda_{\min}$*, then we have that*

$$\mathsf{Tr}(\mathbf{PSQ}) \geq \lambda_{\min} \mathsf{Tr}(\mathbf{PQ}).$$

Proof. Let p_i and q_i be the (i,i)-th entry of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Then we have that

$$\mathsf{Tr}(\mathbf{PSQ}) = \sum_i p_i \mathbf{S}_{i,i} q_i \geq \lambda_{\min} \sum_i p_i q_i = \lambda_{\min} \mathsf{Tr}(\mathbf{PQ}) \tag{8}$$

where the inequality above comes from the positive definiteness of $\mathbf{S}$, i.e., $\mathbf{S}_{i,i} = \mathbf{e}_i^\top \mathbf{S} \mathbf{e}_i \geq \lambda_{\min}, \forall i$. □

Lemma A.5. *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *be a matrix with full column rank and* $\mathbf{B} \in \mathbb{R}^{n \times p}$ *be a non-zero matrix. Let* $\sigma_{\min}(\cdot)$ *be the smallest non-zero singular value. Then it holds that* $\sigma_{\min}(\mathbf{AB}) \geq \sigma_{\min}(\mathbf{A}) \sigma_{\min}(\mathbf{B})$.

Proof. Using the min-max principle for singular values,

$$\begin{aligned}
\sigma_{\min}(\mathbf{AB}) &= \min_{\|\mathbf{x}\|=1, \mathbf{x} \in \text{ColSpan}(\mathbf{B})} \|\mathbf{ABx}\| \\
&= \min_{\|\mathbf{x}\|=1, \mathbf{x} \in \text{ColSpan}(\mathbf{B})} \left\| \mathbf{A} \frac{\mathbf{Bx}}{\|\mathbf{Bx}\|} \right\| \cdot \|\mathbf{Bx}\| \\
&\overset{(a)}{=} \min_{\|\mathbf{x}\|=1, \|\mathbf{y}\|=1, \mathbf{x} \in \text{ColSpan}(\mathbf{B}), \mathbf{y} \in \text{ColSpan}(\mathbf{B})} \|\mathbf{Ay}\| \cdot \|\mathbf{Bx}\| \\
&\geq \min_{\|\mathbf{y}\|=1, \mathbf{y} \in \text{ColSpan}(\mathbf{B})} \|\mathbf{Ay}\| \cdot \min_{\|\mathbf{x}\|=1, \mathbf{x} \in \text{ColSpan}(\mathbf{B})} \|\mathbf{Bx}\| \\
&\geq \min_{\|\mathbf{y}\|=1} \|\mathbf{Ay}\| \cdot \min_{\|\mathbf{x}\|=1, \mathbf{x} \in \text{ColSpan}(\mathbf{B})} \|\mathbf{Bx}\| \\
&= \sigma_{\min}(\mathbf{A}) \sigma_{\min}(\mathbf{B})
\end{aligned}$$

where (a) is by changing of variables, i.e., $\mathbf{y} = \mathbf{Bx}/\|\mathbf{Bx}\|$. □

Lemma A.6 (Theorem 2.2.1 of (Chikuse, 2012))**.** *If* $\mathbf{Z} \in \mathbb{R}^{m \times r}$ *has entries drawn iid from Gaussian distribution* $\mathcal{N}(0,1)$*, then* $\mathbf{X} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1/2}$ *is a random matrix uniformly distributed on* $\mathsf{St}(m,r)$.

Lemma A.7 ((Vershynin, 2010))**.** *] If* $\mathbf{Z} \in \mathbb{R}^{m \times r}$ *is a matrix whose entries are independently drawn from* $\mathcal{N}(0,1)$*. Then for every* $\tau \geq 0$*, with probability at least* $1 - \exp(-\tau^2/2)$*, we have*

$$\sigma_1(\mathbf{Z}) \leq \sqrt{m} + \sqrt{r} + \tau.$$

Lemma A.8 ((Rudelson & Vershynin, 2009))**.** *If* $\mathbf{Z} \in \mathbb{R}^{m \times r}$ *is a matrix whose entries are independently drawn from* $\mathcal{N}(0,1)$*. Suppose that* $m \geq r$*. Then for every* $\tau \geq 0$*, we have for some universal constants* $C_1 > 0$ *and* $C_2 > 0$ *that*

$$\mathbb{P}\Big(\sigma_r(\mathbf{Z}) \leq \tau(\sqrt{m} - \sqrt{r-1})\Big) \leq (C_1\tau)^{m-r+1} + \exp(-C_2 r).$$

Lemma A.9. *If* $\mathbf{U} \in \mathsf{St}(m, r_A)$ *is a fixed matrix,* $\mathbf{X} \in \mathsf{St}(m,r)$ *is uniformly sampled from* $\mathsf{St}(m,r)$ *using methods described in Lemma A.6, and* $r > r_A$*, then we have that with probability at least* $1 - \exp(-m/2) - (C_1\tau)^{r-r_A+1} - \exp(-C_2 d)$*,*

$$\sigma_{r_A}(\mathbf{U}^\top \mathbf{X}) \geq \frac{\tau(r - r_A + 1)}{6\sqrt{mr}}.$$

Proof. Since $\mathbf{X} \in \mathsf{St}(m,r)$ is uniformly sampled from $\mathsf{St}(m,r)$ using methods described in Lemma A.6, we can write $\mathbf{X} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1/2}$, where $\mathbf{Z} \in \mathbb{R}^{m \times r}$ has entries iid sampled from $\mathcal{N}(0,1)$. We thus have

$$\sigma_{r_A}(\mathbf{U}^\top \mathbf{X}) = \sigma_{r_A}\big(\mathbf{U}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1/2}\big).$$

Now consider $\mathbf{U}^\top \mathbf{Z} \in \mathbb{R}^{r_A \times r}$. It is clear that entries of $\mathbf{U}^\top \mathbf{Z}$ are also iid Gaussian random variables $\mathcal{N}(0,1)$. As a consequence of Lemma A.8, we have that w.p. at least $1 - (C_1\tau)^{r-r_A+1} - \exp(-C_2 r)$,

$$\sigma_{r_A}\big(\mathbf{U}^\top \mathbf{Z}\big) \geq \tau(\sqrt{r} - \sqrt{r_A - 1}).$$

We also have from Lemma A.7 that with probability at least $1 - \exp(-m/2)$

$$\sigma_1(\mathbf{Z}^\top \mathbf{Z}) = \sigma_1^2(\mathbf{Z}) \leq (2\sqrt{m} + \sqrt{r})^2.$$

Taking union bound, we have with probability at least $1 - \exp(-m/2) - (C_1\tau)^{r-r_A+1} - \exp(-C_2 r)$,

$$\sigma_{r_A}(\mathbf{U}^\top \mathbf{X}) \overset{(a)}{\geq} \frac{\sigma_{r_A}(\mathbf{U}^\top \mathbf{Z})}{\sigma_1(\mathbf{Z})} = \frac{\tau(\sqrt{r} - \sqrt{r_A - 1})}{2\sqrt{m} + \sqrt{r}} \geq \frac{\tau(r - r_A + 1)}{3\sqrt{m} \cdot 2\sqrt{r}} = \frac{\tau(r - r_A + 1)}{6\sqrt{mr}} \tag{9}$$

where (a) comes from Lemma A.5. □

Lemma A.10. *If $\mathbf{V} \in \mathsf{St}(n, r_A)$ is a fixed matrix, and $\mathbf{Y} \in \mathsf{St}(n, r)$ is uniformly sampled from $\mathsf{St}(n, r)$ using methods described in Lemma A.6. Suppose $r > r_A$. Then we have that with probability at least $1 - \exp(-n/2) - (C_1\tau)^{r-r_A+1} - \exp(-C_2 r)$,*

$$\sigma_{r_A}(\mathbf{V}^\top \mathbf{Y}) \geq \frac{\tau(r - r_A + 1)}{6\sqrt{nr}}.$$

Proof. The proof is omitted since it follows the same steps of Lemma A.9. □

A.4. Equivalence of Matrix Factorization and LoRA

Whitening data refers to transforming the data such that the empirical uncentered covariance matrix of the features is identity. Suppose $\mathbf{D} \in \mathbb{R}^{n \times N}$ holds N training examples in its columns with n features each, then whitened data refers to having $\mathbf{D}\mathbf{D}^\top = \mathbf{I}_n$. We follow standard arguments laid out in (Arora et al., 2018; Li et al., 2024) to show that low-rank adaptation on a linear model with whitened data and squared loss can be written as a matrix factorization problem. Consider the minimization of

$$L(\mathbf{X}, \mathbf{Y}) = \|(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D} - \mathbf{A}\|_\mathsf{F}^2$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ holds m labels for each example, $\mathbf{W}_0 \in \mathbb{R}^{m \times n}$ is the pre-trained weight, and $\mathbf{X} \in \mathbb{R}^{m \times r}$, $\mathbf{Y} \in \mathbb{R}^{n \times r}$ are the weights of LoRA. Rewriting $L(\mathbf{X}, \mathbf{Y})$ yields

$$\begin{aligned}
L(\mathbf{X}, \mathbf{Y}) &= \|(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D} - \mathbf{A}\|_\mathsf{F}^2 \\
&= \mathsf{Tr}(((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D} - \mathbf{A})((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D} - \mathbf{A})^\top) \\
&= \mathsf{Tr}((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D}\mathbf{D}^\top(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)^\top) - \mathsf{Tr}((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D}\mathbf{A}^\top) \\
&\qquad - \mathsf{Tr}(\mathbf{A}\mathbf{D}^\top(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)^\top) + \mathsf{Tr}(\mathbf{A}\mathbf{A}^\top) \\
&\overset{(a)}{=} \mathsf{Tr}((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)^\top) - \mathsf{Tr}((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)\mathbf{D}\mathbf{A}^\top) \\
&\qquad - \mathsf{Tr}(\mathbf{A}\mathbf{D}^\top(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top)^\top) + \mathsf{Tr}(\mathbf{A}\mathbf{A}^\top) \\
&\overset{(b)}{=} \mathsf{Tr}(((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top) - \mathbf{\Lambda})((\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top) - \mathbf{\Lambda})^\top) - \mathsf{Tr}(\mathbf{\Lambda}\mathbf{\Lambda}^\top) + \mathsf{Tr}(\mathbf{A}\mathbf{A}^\top)
\end{aligned}$$

where (a) uses the fact that the data is whitened and (b) defines $\mathbf{\Lambda} = \mathbf{A}\mathbf{D}^\top \in \mathbb{R}^{m \times n}$, i.e., the matrix to be factorized. Thus, we can write

$$\begin{aligned}
L(\mathbf{X}, \mathbf{Y}) &= \|(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top) - \mathbf{\Lambda}\|_\mathsf{F}^2 + c \\
&= \|\mathbf{X}\mathbf{Y}^\top - \mathbf{\Lambda}'\|_\mathsf{F}^2 + c
\end{aligned}$$

for $\mathbf{\Lambda}' := \mathbf{W}_0 - \mathbf{\Lambda}$ and constant $c := -\mathsf{Tr}(\mathbf{\Lambda}\mathbf{\Lambda}^\top) + \mathsf{Tr}(\mathbf{A}\mathbf{A}^\top)$. Using the same arguments, one can frame low-rank adaptation of a linear model with the PoLAR parameterization as the matrix factorization problem given in (4).

Algorithm 2 RGD for PoLAR parameterized (4)

Input: Learning rates η, γ; sample $\mathbf{X}_0$ and $\mathbf{Y}_0$ uniformly from $\mathsf{St}(m,r)$ and $\mathsf{St}(n,r)$, respectively.
for $t = 0, \dots, T-1$ **do**
  Find $\mathbf{\Theta}_t$ via (5a)
  Obtain Riemannian gradients $\mathbf{E}_t$ and $\mathbf{F}_t$
  $\mathbf{X}_{t+1} = (\mathbf{X}_t - \eta \mathbf{E}_t)(\mathbf{I}_r + \eta^2 \mathbf{E}_t^\top \mathbf{E}_t)^{-1/2}$
  $\mathbf{Y}_{t+1} = (\mathbf{Y}_t - \eta \mathbf{F}_t)(\mathbf{I}_r + \eta^2 \mathbf{F}_t^\top \mathbf{F}_t)^{-1/2}$
end for

B. Proof of Theorem 2.1

For ease of reference, we summarize the procedure described in Section 2.2 in Alg. 2. We update

$$\mathbf{X}_{t+1} = (\mathbf{X}_t - \eta \mathbf{E}_t)(\mathbf{I}_r + \eta^2 \mathbf{E}_t^\top \mathbf{E}_t)^{-1/2}. \tag{10}$$

Likewise, the Riemannian gradient of $\mathbf{Y}_t$ is $\mathbf{F}_t = -(\mathbf{I}_n - \mathbf{Y}_t \mathbf{Y}_t^\top)\mathbf{A}^\top \mathbf{X}_t \mathbf{\Theta}_t$, leading to the update

$$\mathbf{Y}_{t+1} = (\mathbf{Y}_t - \eta \mathbf{F}_t)(\mathbf{I}_r + \eta^2 \mathbf{F}_t^\top \mathbf{F}_t)^{-1/2}. \tag{11}$$

B.1. Riemannian Gradients of (4)

One can start with the Euclidean gradient with respect to $\mathbf{X}_t$ as $\tilde{\mathbf{E}}_t = (\mathbf{X}_t \mathbf{\Theta}_t \mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t \mathbf{\Theta}_t^\top = (\mathbf{X}_t \mathbf{X}_t^\top - \mathbf{I}_m)\mathbf{A}\mathbf{Y}_t \mathbf{\Theta}_t^\top$. Note that for $\mathbf{\Theta}_t = \mathbf{X}_t^\top \mathbf{A}\mathbf{Y}_t$ (i.e., $\gamma = 1$) the Euclidean gradient is skew-symmetric such that it is already contained in the tangent space of $\mathsf{St}(m,r)$ at $\mathbf{X}_t$ (i.e., $\mathbf{X}_t^\top \tilde{\mathbf{E}}_t + \tilde{\mathbf{E}}_t^\top \mathbf{X}_t = 0$), yielding equality between the Riemannian gradient $\mathbf{E}_t$ and Euclidean gradient $\tilde{\mathbf{E}}_t$. This equivalence holds regardless of whether the Euclidean or canonical metric is used. In other words, the Riemmanian gradient for $\mathbf{X}_t$ at iteration t is given by

$$\mathbf{E}_t = -(\mathbf{I}_m - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{A}\mathbf{Y}_t \mathbf{\Theta}_t^\top.$$

Similarly, one can obtain the Riemannian gradient for $\mathbf{Y}_t$ via

$$\mathbf{F}_t = -(\mathbf{I}_n - \mathbf{Y}_t \mathbf{Y}_t^\top)\mathbf{A}^\top \mathbf{X}_t \mathbf{\Theta}_t.$$

B.2. General Dynamics

Here we derive several equations that are useful *throughout this section.* Note that the choice of learning rate $\gamma = 1$ will be leveraged in some equations.

Dynamics on $\mathbf{X}_t$, $\mathbf{E}_t$, and $\mathbf{\Phi}_t$. From the updates in Alg. 2, it is straightforward to arrive

$$\begin{aligned}
\mathbb{R}^{r_A \times r} \ni \mathbf{U}^\top \mathbf{E}_t &= -\mathbf{U}^\top (\mathbf{I}_m - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{A}\mathbf{Y}_t \mathbf{\Theta}_t^\top \\
&= -\mathbf{U}^\top (\mathbf{I}_m - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{Y}_t \mathbf{\Theta}_t^\top \\
&= -(\mathbf{I}_{r_A} - \mathbf{\Phi}_t \mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t \mathbf{\Theta}_t^\top.
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
\mathbb{R}^{r \times r} \ni \mathbf{E}_t^\top \mathbf{E}_t &= \mathbf{\Theta}_t \mathbf{Y}_t^\top \mathbf{A}^\top (\mathbf{I}_m - \mathbf{X}_t \mathbf{X}_t^\top)^2 \mathbf{A}\mathbf{Y}_t \mathbf{\Theta}_t^\top \\
&= \mathbf{\Theta}_t \mathbf{Y}_t^\top \mathbf{A}^\top (\mathbf{I}_m - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{A}\mathbf{Y}_t \mathbf{\Theta}_t^\top \\
&= \mathbf{\Theta}_t \mathbf{\Psi}_t^\top \mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t \mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t \mathbf{\Theta}_t^\top.
\end{aligned}$$

Applying $\sigma_1(\mathbf{\Psi}_t) \le 1$ and $\sigma_1(\mathbf{A}) = \sigma_1(\mathbf{\Sigma}) = 1$ to the equation above, and leveraging $\gamma = 1$ in the update of $\mathbf{\Theta}_t$ (i.e., $\mathbf{\Theta}_t = \mathbf{X}_t^\top \mathbf{A}\mathbf{Y}_t = \mathbf{\Phi}_t^\top \mathbf{\Sigma}\mathbf{\Psi}_t$), we have that

$$\sigma_1(\mathbf{E}_t^\top \mathbf{E}_t) \le \sigma_1^4(\mathbf{\Sigma})\sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) = \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top). \tag{12}$$

And the dynamics on the alignment $\mathbf{\Phi}_t \in \mathbb{R}^{r_A \times r}$ can be written as

$$\begin{aligned}\mathbf{\Phi}_{t+1} &= \big[\mathbf{\Phi}_t + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Theta}_t^\top\big]\big(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big)^{-1/2} \\ &\overset{(a)}{=} \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Phi}_t\big(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big)^{-1/2}\end{aligned}$$

where (a) uses $\mathbf{\Theta}_t = \mathbf{X}_t^\top \mathbf{A}\mathbf{Y}_t = \mathbf{\Phi}_t^\top \mathbf{\Sigma}\mathbf{\Psi}_t$. Hence, we have that

$$\begin{aligned}&\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top \\ &= \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Phi}_t\big(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big)^{-1}\mathbf{\Phi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\big].\end{aligned} \tag{13}$$

Dynamics on $\mathbf{Y}_t$, $\mathbf{F}_t$, and $\mathbf{\Psi}_t$. From the updates in Alg. 2 and similar to the derivation above, we arrive at

$$\begin{aligned}\mathbb{R}^{r_A \times r} \ni \mathbf{V}^\top\mathbf{F}_t &= -\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{A}^\top\mathbf{X}_t\mathbf{\Theta}_t \\ &= -\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{X}_t\mathbf{\Theta}_t \\ &= -(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Theta}_t.\end{aligned}$$

Moreover, we also have

$$\begin{aligned}\mathbb{R}^{r \times r} \ni \mathbf{F}_t^\top\mathbf{F}_t &= \mathbf{\Theta}_t^\top\mathbf{X}_t^\top\mathbf{A}(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)^2\mathbf{A}^\top\mathbf{X}_t\mathbf{\Theta}_t \\ &= \mathbf{\Theta}_t^\top\mathbf{X}_t^\top\mathbf{A}(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{A}^\top\mathbf{X}_t\mathbf{\Theta}_t \\ &= \mathbf{\Theta}_t^\top\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Theta}_t.\end{aligned}$$

Applying $\sigma_1(\mathbf{\Phi}_t) \le 1$ and $\sigma_1(\mathbf{A}) = 1$ to the equation above, we have that

$$\sigma_1(\mathbf{F}_t^\top\mathbf{F}_t) \le \sigma_1^4(\mathbf{\Sigma})\sigma_1(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) = \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top). \tag{14}$$

The alignment $\mathbf{\Psi}_t = \mathbf{V}^\top\mathbf{Y}_t \in \mathbb{R}^{r_A \times r}$ can be tracked via

$$\begin{aligned}\mathbf{\Psi}_{t+1} &= \big[\mathbf{\Psi}_t + \eta(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Theta}_t\big]\big(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big)^{-1/2} \\ &\overset{(b)}{=} \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Psi}_t\big(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big)^{-1/2}\end{aligned}$$

where (b) uses $\mathbf{\Theta}_t = \mathbf{X}_t^\top \mathbf{A}\mathbf{Y}_t = \mathbf{\Phi}_t^\top \mathbf{\Sigma}\mathbf{\Psi}_t$. Finally, we have that

$$\begin{aligned}&\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top \\ &= \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Psi}_t\big(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big)^{-1}\mathbf{\Psi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\big].\end{aligned} \tag{15}$$

With these preparations, we are ready to prove our main results.

B.3. Initialization

Lemma B.1. *Suppose that $\mathbf{X}_0$ and $\mathbf{Y}_0$ are uniformly sampled from $\mathsf{St}(m, r)$ and $\mathsf{St}(n, r)$, respectively, using methods described in Lemma A.6. There exist universal constants c_1 and c_2 such that whp the following holds*

$$\begin{aligned}\sigma_{r_A}(\mathbf{\Phi}_0) = \sigma_{r_A}(\mathbf{U}^\top\mathbf{X}_0) &\ge \frac{r - r_A + 1}{\sqrt{c_1 mr}} \ge \frac{r - r_A}{\sqrt{c_1 mr}} \\ \sigma_{r_A}(\mathbf{\Psi}_0) = \sigma_{r_A}(\mathbf{V}^\top\mathbf{Y}_0) &\ge \frac{r - r_A + 1}{\sqrt{c_2 nr}} \ge \frac{r - r_A}{\sqrt{c_2 nr}}.\end{aligned}$$

Proof. The proofs, the constants c_1 and c_2, as well as the exact probability follow directly from Lemma A.9 and Lemma A.10. □

B.4. Increasing Alignment

The Lemma below that the alignment between $\mathbf{X}_t$ and $\mathbf{U}$ is non-decreasing over iterations. This geometric observation bears resemblance to the descent lemma in standard GD.

Lemma B.2 (Increasing Alignment)**.** *Let $\beta_t := \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ and $\delta_t := \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$, and suppose that the learning rates are chosen as $\eta < 1$ and $\gamma = 1$. If the following conditions are met,*

$$\frac{2(1-\eta^2\beta_t)\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big) \geq \eta\beta_t\mathsf{Tr}\big(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big)$$
$$\frac{2(1-\eta^2\delta_t)\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big) \geq \eta\delta_t\mathsf{Tr}\big(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big)$$

Alg. 2 guarantees that $\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) \geq \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ *and* $\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) \geq \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$.

Lemma B.2 is proved in this subsection, where the detailed proof is divided into two parts.

B.4.1. Increasing Alignment Between $\mathbf{X}_t$ and $\mathbf{U}$

Lemma B.3. *Consider Alg. 2 with $\eta < 1$ and $\gamma = 1$. Let $\beta_t := \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$. If it holds that*

$$\frac{2(1-\eta^2\beta_t)\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big) \geq \eta\beta_t\mathsf{Tr}\big(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big),$$

we have $\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) \geq \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$.

Proof. From (13), we have that

$$\begin{aligned}
&\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top \\
&= \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Phi}_t\big(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big)^{-1}\mathbf{\Phi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\big] \\
&\overset{(a)}{\succeq} \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Phi}_t\big(\mathbf{I}_r - \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big)\mathbf{\Phi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\big] \\
&\overset{(b)}{\succeq} (1-\eta^2\beta_t)\big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\big] \\
&\succeq (1-\eta^2\beta_t)\Big\{\mathbf{\Phi}_t\mathbf{\Phi}_t^\top + \eta(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top + \eta\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\Big\}
\end{aligned} \tag{16}$$

where (a) is by Lemma A.2; and (b) is by $\big(\mathbf{I}_r - \eta^2\mathbf{E}_t^\top\mathbf{E}_t\big) \succeq (1 - \eta^2\sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top))\mathbf{I}_r$ as a result of (12), and we write $\beta_t := \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ for convenience. We also dropped the fourth term $(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}_t\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ given its PSDness. Note that $\beta_t \in [0, 1]$.

Now let the EVD of $\mathbf{\Phi}_t\mathbf{\Phi}_t^\top = \mathbf{Q}_t\mathbf{\Lambda}_t\mathbf{Q}_t^\top$, where both $\mathbf{Q}_t$ and $\mathbf{\Lambda}_t$ are $r_A \times r_A$ matrices. Note that $\mathbf{0} \preceq \mathbf{\Lambda}_t \preceq \mathbf{I}_{r_A}$. Then we have that

$$\begin{aligned}
&\mathsf{Tr}\Big((\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\Big) \\
&= \mathsf{Tr}\Big(\mathbf{Q}_t(\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{Q}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{Q}_t\mathbf{\Lambda}_t\mathbf{Q}_t^\top\Big) \\
&= \mathsf{Tr}\Big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{Q}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{Q}_t\mathbf{\Lambda}_t\Big) \\
&\overset{(c)}{\succeq} \frac{\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\Big) \\
&= \frac{\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\Big)
\end{aligned} \tag{17}$$

where (c) is by Lemma A.4 and Lemma A.5. More precisely, the PSDness of $\mathbf{Q}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{Q}_t$ justifies the prerequisites for Lemma A.4, and then we use $\sigma_{r_A}(\mathbf{Q}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{Q}_t) \geq \sigma_{r_A}^2(\mathbf{\Sigma})\sigma_{r_A}^2(\mathbf{\Psi}_t) = \sigma_{r_A}^2(\mathbf{\Psi}_t)/\kappa^2$.

Taking trace on both sides of (16) and plugging (17) in, we arrive at

$$\frac{\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top)}{1-\eta^2\beta_t} \geq \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) + \frac{2\eta\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\Big). \tag{18}$$

Simplifying this inequality gives the results. □

B.4.2. Increasing Alignment Between $\mathbf{Y}_t$ and $\mathbf{V}$

We proceed by proving the analogue of Lemma B.3 for the alignment between $\mathbf{Y}_t$ and $\mathbf{V}$ by following the same steps.

Lemma B.4. *Consider Alg. 2 with $\eta < 1$ and $\gamma = 1$. Let $\delta_t := \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$. If it holds that*

$$\frac{2(1-\eta^2\delta_t)\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}\mathsf{Tr}\big((\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big) \geq \eta\delta_t\mathsf{Tr}\big(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big),$$

we have $\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) \geq \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$.

Proof. From (15), we have that

$$\begin{aligned}
&\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top \\
&= \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Psi}_t\big(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big)^{-1}\mathbf{\Psi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\big] \\
&\overset{(a)}{\succeq} \big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Psi}_t\big(\mathbf{I}_r - \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big)\mathbf{\Psi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\big] \\
&\overset{(b)}{\succeq} (1-\eta^2\delta_t)\big[\mathbf{I}_{r_A} + \eta(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\big]\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big[\mathbf{I}_{r_A} + \eta\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\big] \\
&\succeq (1-\eta^2\delta_t)\Big\{\mathbf{\Psi}_t\mathbf{\Psi}_t^\top + \eta(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top + \eta\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\Big\}
\end{aligned} \tag{19}$$

where (a) is by Lemma A.2; and (b) is by $\big(\mathbf{I}_r - \eta^2\mathbf{F}_t^\top\mathbf{F}_t\big) \succeq (1-\eta^2\sigma_1(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top))\mathbf{I}_r$ as a result of (14), and we write $\delta_t := \sigma_1(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$ for convenience. Note that $\delta_t \in [0,1]$.

Now let the SVD of $\mathbf{\Psi}_t\mathbf{\Psi}_t^\top = \mathbf{P}_t\tilde{\mathbf{\Lambda}}_t\mathbf{P}_t^\top$, where both $\mathbf{P}_t$ and $\tilde{\mathbf{\Lambda}}_t$ are $r_A \times r_A$ matrices. Note that $\mathbf{0} \preceq \tilde{\mathbf{\Lambda}}_t \preceq \mathbf{I}_{r_A}$. Then we have that

$$\begin{aligned}
&\mathsf{Tr}\Big((\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\Big) \\
&= \mathsf{Tr}\Big(\mathbf{P}_t(\mathbf{I}_{r_A}-\tilde{\mathbf{\Lambda}}_t)\mathbf{P}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{P}_t\tilde{\mathbf{\Lambda}}_t\mathbf{P}_t^\top\Big) \\
&= \mathsf{Tr}\Big((\mathbf{I}_{r_A}-\tilde{\mathbf{\Lambda}}_t)\mathbf{P}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{P}_t\tilde{\mathbf{\Lambda}}_t\Big) \\
&\overset{(c)}{\succeq} \frac{\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A}-\tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\Big) \\
&= \frac{\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\Big)
\end{aligned} \tag{20}$$

where (c) is by Lemma A.4 and Lemma A.5. More precisely, we use the PSDness of $\mathbf{P}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{P}_t$ for applying Lemma A.4, and then employ $\sigma_{r_A}(\mathbf{P}_t^\top\mathbf{\Sigma}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{\Sigma}\mathbf{P}_t) \geq \sigma_{r_A}^2(\mathbf{\Sigma})\sigma_{r_A}^2(\mathbf{\Phi}_t) = \sigma_{r_A}^2(\mathbf{\Phi}_t)/\kappa^2$.

Taking trace on both sides of (19) and plugging (20) in, we arrive at

$$\frac{\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top)}{1-\eta^2\delta_t} \geq \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) + \frac{2\eta\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}\mathsf{Tr}\Big((\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\Big). \tag{21}$$

Simplifying this inequality gives the results. □

B.5. Non-Increasing Misalignment

Misalignment of $\mathbf{X}_t$ refers to the principal angles between $\mathbf{X}_t$ and the basis of the orthogonal complement of $\mathbf{U}$. Similarly, one can define the misalignment of $\mathbf{Y}_t$.

B.5.1. NON-INCREASING MISALIGNMENT OF $\mathbf{X}_t$

Lemma B.5. *Denote the orthogonal complement of* $\mathbf{U}$ *as* $\mathbf{U}_\perp \in \mathbb{R}^{m\times(m-r_A)}$. *Define the* $(m-r_A)\times r$ *matrix* $\mathbf{\Omega}_t := \mathbf{U}_\perp^\top \mathbf{X}_t$ *to characterize the alignment of* $\mathbf{X}_t$ *and* $\mathbf{U}_\perp$. *Under the same setting of Lemma B.3, we have that* $\mathbf{\Omega}_{t+1}\mathbf{\Omega}_{t+1}^\top \preceq \mathbf{\Omega}_t\mathbf{\Omega}_t^\top$. *Moreover, if* $r_A \le \frac{m}{2}$, *it is guaranteed to have* $\sigma^2_{r_A}(\mathbf{\Phi}_{t+1}) \ge \sigma^2_{r_A}(\mathbf{\Phi}_t)$.

Proof. From update (5b), we have that

$$\begin{aligned}
\mathbf{\Omega}_{t+1} &= \mathbf{U}_\perp^\top(\mathbf{X}_t - \eta\mathbf{E}_t)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1/2}\\
&= \big(\mathbf{\Omega}_t + \eta\mathbf{U}_\perp^\top(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{Y}_t\mathbf{\Theta}_t^\top\big)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1/2}\\
&\overset{(a)}{=} \big(\mathbf{\Omega}_t - \eta\mathbf{U}_\perp^\top\mathbf{X}_t\mathbf{\Theta}_t\mathbf{\Theta}_t^\top\big)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1/2}\\
&= \mathbf{\Omega}_t\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t\mathbf{\Theta}_t^\top\big)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1/2}.
\end{aligned}$$

where in (a) we have used $\mathbf{U}_\perp^\top\mathbf{A} = \mathbf{0}$ and $\mathbf{\Theta}_t = \mathbf{X}_t^\top\mathbf{A}\mathbf{Y}_t$. With this, we can see that

$$\begin{aligned}
\mathbf{\Omega}_{t+1}\mathbf{\Omega}_{t+1}^\top &= \mathbf{\Omega}_t\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t\mathbf{\Theta}_t^\top\big)(\mathbf{I}_r + \eta^2\mathbf{E}_t^\top\mathbf{E}_t)^{-1}\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t\mathbf{\Theta}_t^\top\big)\mathbf{\Omega}_t^\top\\
&\preceq \mathbf{\Omega}_t\mathbf{\Omega}_t^\top
\end{aligned}$$

where the last inequality comes from the fact that the three matrices in between are all PSD and their largest eigenvalues are smaller than 1 given our choices of η. This gives the proof for the first part of this Lemma.

To show $\sigma^2_{r_A}(\mathbf{\Phi}_{t+1}) \ge \sigma^2_{r_A}(\mathbf{\Phi}_t)$, notice that given $2r_A \le m$, we have from Lemma A.3 that $\sigma^2_{r_A}(\mathbf{\Phi}_t) = 1 - \sigma^2_{r_A}(\mathbf{\Omega}_t)$ and $\sigma^2_{r_A}(\mathbf{\Phi}_{t+1}) = 1 - \sigma^2_{r_A}(\mathbf{\Omega}_{t+1})$. The conclusion is straightforward. □

B.5.2. NON-INCREASING MISALIGNMENT OF $\mathbf{Y}_t$

Lemma B.6. *Denote the orthogonal complement of* $\mathbf{V}$ *as* $\mathbf{V}_\perp \in \mathbb{R}^{n\times(n-r_A)}$. *Define the* $(n-r_A)\times r$ *matrix* $\tilde{\mathbf{\Omega}}_t := \mathbf{V}_\perp^\top\mathbf{Y}_t$ *to characterize the alignment of* $\mathbf{Y}_t$ *and* $\mathbf{V}_\perp$. *Under the same setting of Lemma B.4, we have that* $\tilde{\mathbf{\Omega}}_{t+1}\tilde{\mathbf{\Omega}}_{t+1}^\top \preceq \tilde{\mathbf{\Omega}}_t\tilde{\mathbf{\Omega}}_t^\top$. *Moreover, if* $r_A \le \frac{n}{2}$, *it is guaranteed to have* $\sigma^2_{r_A}(\mathbf{\Psi}_{t+1}) \ge \sigma^2_{r_A}(\mathbf{\Psi}_t)$.

Proof. From update (11), we have that

$$\begin{aligned}
\tilde{\mathbf{\Omega}}_{t+1} &= \mathbf{V}_\perp^\top(\mathbf{Y}_t - \eta\mathbf{F}_t)(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t)^{-1/2}\\
&= \big(\tilde{\mathbf{\Omega}}_t + \eta\mathbf{V}_\perp^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{A}^\top\mathbf{X}_t\mathbf{\Theta}_t\big)(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t)^{-1/2}\\
&= \big(\tilde{\mathbf{\Omega}}_t - \eta\mathbf{V}_\perp^\top\mathbf{Y}_t\mathbf{\Theta}_t^\top\mathbf{\Theta}_t\big)(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t)^{-1/2}\\
&= \tilde{\mathbf{\Omega}}_t\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t^\top\mathbf{\Theta}_t\big)(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t)^{-1/2}.
\end{aligned}$$

With this, we can see that

$$\begin{aligned}
\tilde{\mathbf{\Omega}}_{t+1}\tilde{\mathbf{\Omega}}_{t+1}^\top &= \tilde{\mathbf{\Omega}}_t\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t^\top\mathbf{\Theta}_t\big)(\mathbf{I}_r + \eta^2\mathbf{F}_t^\top\mathbf{F}_t)^{-1}\big(\mathbf{I}_r - \eta\mathbf{\Theta}_t^\top\mathbf{\Theta}_t\big)\tilde{\mathbf{\Omega}}_t^\top\\
&\preceq \tilde{\mathbf{\Omega}}_t\tilde{\mathbf{\Omega}}_t^\top
\end{aligned}$$

where the last inequality comes from the fact that the three matrices in between are all PSD and their largest eigenvalues are smaller than 1 given our choice of η.

To show $\sigma^2_{r_A}(\mathbf{\Psi}_{t+1}) \ge \sigma^2_{r_A}(\mathbf{\Psi}_t)$, notice that given $2r_A \le m$, we have from Lemma A.3 that $\sigma^2_{r_A}(\mathbf{\Psi}_t) = 1 - \sigma^2_{r_A}(\tilde{\mathbf{\Omega}}_t)$ and $\sigma^2_{r_A}(\mathbf{\Psi}_{t+1}) = 1 - \sigma^2_{r_A}(\tilde{\mathbf{\Omega}}_{t+1})$. The conclusion is straightforward. □

B.6. Convergence of $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ and $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$

B.6.1. DYNAMICS OF $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$

Lemma B.7. *Suppose that* $r_A \le \frac{n}{2}$, *and let* $\rho := \min\{\frac{1}{m}, \frac{(r-r_A)^2}{mr}\}$. *Choosing* $\eta = \mathcal{O}\big(\frac{\rho(r-r_A)^2}{r^2\kappa^2 m}\big)$ *and* $\gamma = 1$, *Alg. 2 guarantees that after at most* $T = \mathcal{O}\big(\frac{r_A r^3\kappa^4 m^2}{\rho^2(r-r_A)^4} + \frac{m^2r^3\kappa^4}{\rho(r-r_A)^4}\log\frac{1}{\epsilon}\big)$ *steps* $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \le \epsilon$.

Proof. By rewriting (18), we arrive at

$$\begin{aligned}&\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\ &\geq \frac{2\eta\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}(1-\eta^2\beta_t)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\big) - \eta^2\beta_t\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top).\end{aligned} \tag{22}$$

Based on (22), we discuss the convergence in three different regimes.

Phase I. $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \geq r_A - 0.5$**.** This is the initial phase, and the condition is equivalent to $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \leq 0.5$. For notational convenience, let the SVD of $\mathbf{\Phi}_t\mathbf{\Phi}_t^\top = \mathbf{Q}_t\mathbf{\Lambda}_t\mathbf{Q}_t^\top$. Given these conditions, it can be seen that $\sigma_{r_A}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) = \sigma_{r_A}(\mathbf{I}_{r_A} - \mathbf{\Lambda}_t) \geq 0.5$. Recalling that $\beta_t = \sigma_1(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \leq 1$, we can simplify (22) as

$$\begin{aligned}\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) &\geq \frac{2\eta\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) - \eta^2\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\ &\overset{(a)}{\geq} \frac{\eta\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}(1-\eta^2)\mathsf{Tr}(\mathbf{\Lambda}_t) - \eta^2\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\ &\overset{(b)}{\geq} \frac{\eta(r-r_A)^2}{\kappa^2 c_2 nr}(1-\eta^2)\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) - \eta^2\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)\end{aligned}$$

where (a) uses $\sigma_{r_A}(\mathbf{I}_{r_A} - \mathbf{\Lambda}_t) \geq 0.5$; (b) uses Lemmas B.6 and B.1, which jointly imply that $\sigma_{r_A}^2(\mathbf{\Psi}_t) \geq \sigma_{r_A}^2(\mathbf{\Psi}_0) \geq (r-r_A)^2/(c_2 nr)$ for some universal constant c_2 defined in Lemma B.1. Rearranging the terms, we arrive at

$$\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) \geq \left(1 + \frac{\eta(r-r_A)^2}{\kappa^2 c_2 nr}(1-\eta^2) - \eta^2\right)\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$$

which is linearly increasing once the term in parentheses is greater than 1. This amounts to choosing a small enough η, i.e., $\eta \leq \mathcal{O}\big(\frac{(r-r_A)^2}{\kappa^2 nr}\big)$.

Phase II. $0.5 < \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) < r_A - 0.5$**.** Suppose that $\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) \geq \rho$, for some $\rho > 0$ to be discussed shortly. Let $\eta = \mathcal{O}\big(\frac{\rho(r-r_A)^2}{r^2\kappa^2 m}\big)$ and $\eta \leq 0.5$, it is straightforward to have

$$\begin{aligned}\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) &\geq \frac{2\eta(r-r_A)^2}{\kappa^2 c_2 nr}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) - \eta^2 r_A \\ &\geq \frac{2\eta(r-r_A)^2}{\kappa^2 c_2 mr}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) - \eta^2 r \\ &\geq \mathcal{O}\Big(\frac{\rho^2(r-r_A)^4}{r^3\kappa^4 m^2}\Big) := \Delta_1.\end{aligned} \tag{23}$$

Note that the $\mathcal{O}(\cdot)$ notation ignores the dependence on constants including c_1 and c_2. This means that per step, $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ increases at least by Δ_1. Consequently, after at most $(r_A - 1)/\Delta_1 = \mathcal{O}\big(r_A r^3\kappa^4 m^2/(\rho^2(r-r_A)^4)\big)$ iterations, RGD leaves Phase II.

Next, we show that $\rho \geq \mathcal{O}(\min\{\frac{1}{m}, \frac{(r-r_A)^2}{mr}\})$. Notice that $\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) \geq \sum_{i=1}^{r_A}\sigma_i^2(\mathbf{\Phi}_t)(1-\sigma_i^2(\mathbf{\Phi}_t)) \geq \sigma_{r_A}^2(\mathbf{\Phi}_t)(1-\sigma_{r_A}^2(\mathbf{\Phi}_t)) \geq \mathcal{O}(\min\{\frac{1}{m}, \frac{(r-r_A)^2}{mr}\})$, where the last inequality comes from the facts that i) for $x \in [a,b]$ with $0 < a < 0.5 < b < 1$, the smallest value of $x(1-x)$ is $\min\{a(1-a), b(1-b)\}$; and, ii) $\sigma_{r_A}^2(\mathbf{\Phi}_t)$ belongs to interval $[a,b]$ with $a = \mathcal{O}\big(\frac{(r-r_A)^2}{mr}\big)$ and $b = \frac{r_A - 0.5}{r_A} = 1 - \frac{1}{2r_A} \leq 1 - \frac{1}{m}$. Lemmas B.5 and B.1 are adopted to calculate a, that is, $\sigma_{r_A}^2(\mathbf{\Phi}_t) \geq \sigma_{r_A}^2(\mathbf{\Phi}_0) = \mathcal{O}((r-r_A)^2/mr)$.

Phase III. $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \leq 0.5$**.** This is a regime near the optimum. An implication of this phase is that $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \geq r_A - 0.5$. Given that the singular values of $\mathbf{\Phi}_t\mathbf{\Phi}_t^\top$ belong to $[0,1]$, it can be seen that $\sigma_{r_A}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) = \sigma_{r_A}(\mathbf{\Lambda}_t) \geq 0.5$.

Together with $\beta_t \le 0.5$ in this scenario, we can simplify (22) as

$$\begin{aligned}
&\mathsf{Tr}(\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\
&\ge \frac{2\eta\sigma_{r_A}^2(\mathbf{\Psi}_t)}{\kappa^2}\left(1-\frac{\eta^2}{2}\right)\mathsf{Tr}\big((\mathbf{I}_{r_A}-\mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) - \eta^2\beta_t\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\
&\overset{(c)}{\ge} \frac{\eta(r-r_A)^2}{c_2 nr\kappa^2}\left(1-\frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Lambda}_t) - \eta^2\beta_t r_A \\
&= \frac{\eta(r-r_A)^2}{c_2 nr\kappa^2}\left(1-\frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) - \eta^2\beta_t r_A \\
&\overset{(d)}{\ge} \frac{\eta(r-r_A)^2}{c_2 nr\kappa^2}\left(1-\frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) - \eta^2 r_A\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)
\end{aligned}$$

where (c) uses $\sigma_{r_A}(\mathbf{\Lambda}_t) \ge 0.5$; and (d) comes from $\beta_t \le \mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$. This further implies that

$$\begin{aligned}
&\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \\
&\le -\frac{\eta(r-r_A)^2}{c_2 nr\kappa^2}\left(1-\frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top) + \eta^2 r_A\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top).
\end{aligned}$$

Reorganizing the terms, we arrive at

$$\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top) \le \left(1-\frac{\eta(r-r_A)^2}{c_2 nr\kappa^2}\left(1-\frac{\eta^2}{2}\right)+\eta^2 r_A\right)\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Phi}_t\mathbf{\Phi}_t^\top). \tag{24}$$

This indicates a linear rate until we achieve optimality once η is chosen sufficiently small.

Note that our choice of η ensures the conditions in Lemma B.3 are satisfied, indicating that an increase of $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ per iteration is guaranteed. This means that $\mathsf{Tr}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top)$ traverses Phase I, II, and III consecutively. Combining these three phases together gives the claimed complexity bound. □

B.6.2. Dynamics of $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$

Lemma B.8. *Suppose that $r_A \le \frac{n}{2}$, and let $\rho := \min\{\frac{1}{m}, \frac{(r-r_A)^2}{mr}\}$. Choosing $\eta = \mathcal{O}\big(\frac{\rho(r-r_A)^2}{r^2\kappa^2 m}\big)$ and $\gamma = 1$, Alg.2 guarantees that after at most $T = \mathcal{O}\big(\frac{r_A r^3\kappa^4 m^2}{\rho^2(r-r_A)^4} + \frac{m^2 r^3\kappa^4}{\rho(r-r_A)^4}\log\frac{1}{\epsilon}\big)$ steps $\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \le \epsilon$.*

Proof. By rewriting (21), we arrive at

$$\begin{aligned}
&\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\ge \frac{2\eta\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}(1-\eta^2\delta_t)\mathsf{Tr}\big((\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)\mathbf{\Psi}_t\mathbf{\Psi}_t^\top\big) - \eta^2\delta_t\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top).
\end{aligned} \tag{25}$$

Based on (25), we discuss the convergence in three different regimes.

Phase I. $\mathsf{Tr}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \ge r_A - 0.5$. This is the initial phase, and the condition is equivalent to $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \le 0.5$. For notational convenience let the SVD of $\mathbf{\Psi}_t\mathbf{\Psi}_t^\top = \mathbf{P}_t\tilde{\mathbf{\Lambda}}_t\mathbf{P}_t^\top$. Given these conditions, it can be seen that $\sigma_{r_A}(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) = \sigma_{r_A}(\mathbf{I}_{r_A}-\tilde{\mathbf{\Lambda}}_t) \ge 0.5$. Together with $\delta_t \le 1$ (recall that $\delta_t = \sigma_1(\mathbf{I}_{r_A}-\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$), we can simplify (25) as

$$\begin{aligned}
\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) &\ge \frac{2\eta\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A}-\tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\big) - \eta^2\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\overset{(a)}{\ge} \frac{\eta\sigma_{r_A}^2(\mathbf{\Phi}_t)}{\kappa^2}(1-\eta^2)\mathsf{Tr}(\tilde{\mathbf{\Lambda}}_t) - \eta^2\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\overset{(b)}{\ge} \frac{\eta(r-r_A)^2}{\kappa^2 c_1 mr}(1-\eta^2)\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) - \eta^2\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)
\end{aligned}$$

where (a) uses $\sigma_{r_A}(\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t) \geq 0.5$; (b) uses Lemmas B.5 and B.1, which jointly imply that $\sigma^2_{r_A}(\mathbf{\Phi}_t) \geq \sigma^2_{r_A}(\mathbf{\Phi}_0) \geq (r - r_A)^2/(c_1 mr)$ for some universal constant c_1 in defined in Lemma B.1. Rearranging the terms, we arrive at

$$\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) \geq \left(1 + \frac{\eta(r-r_A)^2}{\kappa^2 c_1 mr}(1-\eta^2) - \eta^2\right)\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$$

which is linearly increasing once the term in parentheses is greater than 1. This amounts to choosing a small enough η, i.e., $\eta \leq \mathcal{O}\big(\frac{(r-r_A)^2}{\kappa^2 mr}\big)$.

Phase II. $0.5 < \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) < r_A - 0.5$**.** Suppose that $\mathsf{Tr}\big((\mathbf{I}_{r_A} - \mathbf{\Lambda}_t)\mathbf{\Lambda}_t\big) \geq \rho$, for some ρ to be discussed shortly. Choosing $\eta \leq 0.5$, and $\eta = \mathcal{O}\big(\frac{\rho(r-r_A)^2}{r^2\kappa^2 m}\big)$, it is straightforward to have

$$\begin{aligned}
\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) &\geq \frac{2\eta(r-r_A)^2}{\kappa^2 c_1 mr}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\big) - \eta^2 r_A \\
&\geq \frac{2\eta(r-r_A)^2}{\kappa^2 c_1 mr}(1-\eta^2)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\big) - \eta^2 r \\
&\geq \mathcal{O}\Big(\frac{\rho^2(r-r_A)^4}{r^3\kappa^4 m^2}\Big) := \Delta_2.
\end{aligned} \tag{26}$$

Note that the $\mathcal{O}(\cdot)$ notation ignores the dependence on constants including c_1 and c_2. This means that per step, $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$ at least increases by Δ_2. Consequently, after at most $(r_A - 1)/\Delta_2 = \mathcal{O}(r_A r^3\kappa^4 m^2/\rho^2(r-r_A)^4)$ iterations, RGD leaves Phase II.

Next, we show that $\rho \geq \mathcal{O}(\min\{\frac{1}{n}, \frac{(r-r_A)^2}{nr}\}) \geq \mathcal{O}(\min\{\frac{1}{m}, \frac{(r-r_A)^2}{mr}\})$. Notice that $\mathsf{Tr}\big((\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\big) \geq \sum_{i=1}^{r_A}\sigma_i^2(\mathbf{\Psi}_t)(1-\sigma_i^2(\mathbf{\Psi}_t)) \geq \sigma^2_{r_A}(\mathbf{\Psi}_t)(1-\sigma^2_{r_A}(\mathbf{\Psi}_t)) \geq \mathcal{O}(\min\{\frac{1}{n}, \frac{(r-r_A)^2}{nr}\})$, where the last inequality comes from the facts that i) for $x \in [a,b]$ with $0 < a < 0.5 < b < 1$, the smallest value of $x(1-x)$ is $\min\{a(1-a), b(1-b)\}$; and, ii) $\sigma^2_{r_A}(\mathbf{\Psi}_t)$ belongs to interval $[a,b]$ with $a = \mathcal{O}(\frac{(r-r_A)^2}{nr})$ and $b = \frac{r_A - 0.5}{r_A} = 1 - \frac{1}{2r_A} \leq 1 - \frac{1}{n}$. Lemmas B.6 and B.1 are adopted to calculate a, that is, $a = \sigma^2_{r_A}(\mathbf{\Psi}_t) \geq \sigma^2_{r_A}(\mathbf{\Psi}_0) = \mathcal{O}((r-r_A)^2/nr)$.

Phase III. $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \leq 0.5$**.** This is a regime near the optimum. An implication of this phase is that $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \geq r_A - 0.5$. Given that the singular values of $\mathbf{\Psi}_t\mathbf{\Psi}_t^\top$ belong to $[0,1]$, it can be seen that $\sigma_{r_A}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) = \sigma_{r_A}(\tilde{\mathbf{\Lambda}}_t) \geq 0.5$. Together with $\delta_t \leq 0.5$ in this scenario, we can simplify (22) as

$$\begin{aligned}
&\mathsf{Tr}(\mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\geq \frac{2\eta\sigma^2_{r_A}(\mathbf{\Phi}_t)}{\kappa^2}\left(1 - \frac{\eta^2}{2}\right)\mathsf{Tr}\big((\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t)\tilde{\mathbf{\Lambda}}_t\big) - \eta^2\delta_t\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\overset{(c)}{\geq} \frac{\eta(r-r_A)^2}{c_1 mr\kappa^2}\left(1 - \frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A} - \tilde{\mathbf{\Lambda}}_t) - \eta^2\delta_t r_A \\
&= \frac{\eta(r-r_A)^2}{c_1 mr\kappa^2}\left(1 - \frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) - \eta^2\delta_t r_A \\
&\overset{(d)}{\geq} \frac{\eta(r-r_A)^2}{c_1 mr\kappa^2}\left(1 - \frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) - \eta^2 r_A\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)
\end{aligned}$$

where (c) comes from $\sigma_{r_A}(\tilde{\mathbf{\Lambda}}_t) \geq 0.5$, as well as $\sigma^2_{r_A}(\mathbf{\Phi}_t) \geq \sigma^2_{r_A}(\mathbf{\Phi}_0) \geq (r-r_A)^2/(c_1 mr)$; and (d) uses $\delta_t \leq \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$. This further implies that

$$\begin{aligned}
&\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) - \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \\
&\leq -\frac{\eta(r-r_A)^2}{c_1 mr\kappa^2}\left(1 - \frac{\eta^2}{2}\right)\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) + \eta^2 r_A\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top).
\end{aligned}$$

Reorganizing the terms, we arrive at

$$\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_{t+1}\mathbf{\Psi}_{t+1}^\top) \leq \left(1 - \frac{\eta(r-r_A)^2}{c_1 mr\kappa^2}\left(1 - \frac{\eta^2}{2}\right) + \eta^2 r_A\right)\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top).$$

This indicates a linear rate until we achieve optimality once η is chosen sufficiently small.

Note that our choice of η ensures the conditions in Lemma B.4 are satisfied. In other words, increasing $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$ across t is guaranteed. This means that $\mathsf{Tr}(\mathbf{\Psi}_t\mathbf{\Psi}_t^\top)$ will traverse Phase I, II, and III consecutively. Combining these three phases together gives the claimed complexity bound. □

B.7. Convergence of $\mathbf{\Theta}_t$

Lemma B.9. *Suppose that at iteration t, Alg. 2 with $\gamma = 1$ satisfies $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \le \rho_1$ and $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \le \rho_2$. It is guaranteed to have $f(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{\Theta}_t) = \mathcal{O}(\rho_1 + \rho_2)$.*

Proof. Recall that $\gamma = 1$ implies $\mathbf{\Theta}_t = \mathbf{X}_t^\top\mathbf{A}\mathbf{Y}_t$, we thus have that

$$\begin{aligned}
\|\mathbf{X}_t\mathbf{\Theta}_t\mathbf{Y}_t^\top - \mathbf{A}\|_\mathsf{F} &= \|\mathbf{X}_t\mathbf{X}_t^\top\mathbf{A}\mathbf{Y}_t\mathbf{Y}_t^\top - \mathbf{A}\|_\mathsf{F} \\
&= \|\mathbf{X}_t\mathbf{X}_t^\top\mathbf{A}\mathbf{Y}_t\mathbf{Y}_t^\top - \mathbf{A}\mathbf{Y}_t\mathbf{Y}_t^\top + \mathbf{A}\mathbf{Y}_t\mathbf{Y}_t^\top - \mathbf{A}\|_\mathsf{F} \\
&\le \|(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{I}_m)\mathbf{A}\mathbf{Y}_t\mathbf{Y}_t^\top\|_\mathsf{F} + \|\mathbf{A}(\mathbf{Y}_t\mathbf{Y}_t^\top - \mathbf{I}_n)\|_\mathsf{F} \\
&\overset{(a)}{\le} \|(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{I}_m)\mathbf{U}\|_\mathsf{F}\|\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}_t\mathbf{Y}_t^\top\| + \|\mathbf{U}\mathbf{\Sigma}\|\|\mathbf{V}^\top(\mathbf{Y}_t\mathbf{Y}_t^\top - \mathbf{I}_n)\|_\mathsf{F} \\
&\le \|\mathbf{\Sigma}\|\|(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{U}\|_\mathsf{F} + \|\mathbf{\Sigma}\|\|\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\|_\mathsf{F}
\end{aligned}$$

where (a) uses the compact SVD of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Now we have that

$$\begin{aligned}
\|(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{U}\|_\mathsf{F}^2 &= \mathsf{Tr}\Big(\mathbf{U}^\top(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{I}_m - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{U}\Big) \\
&= \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \le \rho_1.
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
\|\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\|_\mathsf{F}^2 &= \mathsf{Tr}\Big(\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}\Big) \\
&= \mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \le \rho_2.
\end{aligned}$$

Combining these inequalities, we have that

$$\|\mathbf{X}_t\mathbf{\Theta}_t\mathbf{Y}_t^\top - \mathbf{A}\|_\mathsf{F}^2 \overset{(b)}{\le} 2\|\mathbf{\Sigma}\|^2\|(\mathbf{I}_m - \mathbf{X}\mathbf{X}^\top)\mathbf{U}\|_\mathsf{F}^2 + 2\|\mathbf{\Sigma}\|^2\|\mathbf{V}^\top(\mathbf{I}_n - \mathbf{Y}_t\mathbf{Y}_t^\top)\|_\mathsf{F}^2 = \mathcal{O}(\rho_1 + \rho_2)$$

where (b) uses $(a+b)^2 \le 2a^2 + 2b^2$. This finishes the proof. □

B.8. Proof of Theorem 2.1

Proof. The proof is straightforward. We apply Lemma B.7 and Lemma B.8 to show that within $\mathcal{O}\big(\frac{m^2r^3r_A\kappa^4}{\rho^2(r-r_A)^4} + \frac{m^2r^3\kappa^4}{\rho(r-r_A)^4}\log\frac{1}{\epsilon}\big)$ iterations, we have $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Psi}_t\mathbf{\Psi}_t^\top) \le \epsilon$ and $\mathsf{Tr}(\mathbf{I}_{r_A} - \mathbf{\Phi}_t\mathbf{\Phi}_t^\top) \le \epsilon$. Then, Lemma B.9 is adopted to reach the conclusion. □

C. Additional Experiments

In Table 2, we perform an ablation on the parameterization and the gradient type on Gemma-2-2B.

D. Experimental Details

Experiments are performed on NVIDIA GH200 GPUs using *PyTorch* (Paszke et al., 2019).

We consider the following tasks: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019), ARC-e and ARC-c (Clark et al., 2018), and OpenbookQA (Mihaylov et al., 2018). To facilitate reproducibility, we use Eleuther-AI's *lm-evaluation-harness* (Biderman et al., 2024) and report the accuracy based on multiple-choice log-likelihood evaluation, i.e., we select the answer choice with the highest

Table 2: Accuracy of Gemma-2-2B with PoLAR on commonsense reasoning tasks for different gradient types and parameterizations. Rie. (Eucl.) refers to Riemannian (Euclidean) gradient.

Rank	Param. $\mathbf{\Theta}$	Grad.	BoolQ	PIQA	SIQA	HeSw	WiGr	ARC-e	ARC-c	OBQA	**Avg.**
4	$\mathrm{Diag}(r)$	Eucl.	86.24	81.50	58.70	79.88	78.69	78.75	52.39	56.80	71.62
	$\mathbb{R}^{r\times r}$	Rie.	86.48	81.66	58.90	79.69	80.03	81.78	54.69	56.40	**72.45**
32	$\mathrm{Diag}(r)$	Eucl.	87.03	81.39	60.08	81.73	77.51	79.21	55.29	56.60	72.35
	$\mathbb{R}^{r\times r}$	Rie.	87.28	81.61	59.72	81.40	77.74	81.78	54.86	57.80	**72.77**

conditional log-likelihood as the predicted answer. For datasets with answer choices of varying length (PIQA, ARC-e, ARC-c, OpenbookQA, and Hellaswag), we perform byte-length normalization of the log-likelihood scores to remove any bias due to the answer length.

For the results in Table 1, we train for 5 epochs on each task with batch size 128 and choose the learning rate within $\{4\times10^{-4}, 8\times10^{-4}, 4\times10^{-3}\}$. We tune $\lambda \in \{10^{-3}, 5\times10^{-3}\}$ for PoLAR and set $\alpha = 32$. We choose the combination that performs best on average throughout all datasets and report these.