On the Crucial Role of Initialization for Matrix Factorization

Bingcong Li BINGCONG.LI@INF.ETHZ.CH
ETH Zurich, Switzerland
Liang Zhang LIANG.ZHANG@INF.ETHZ.CH
ETH Zurich, Switzerland
Aryan Mokhtari MOKHTARI@AUSTIN.UTEXAS.EDU
University of Texas at Austin, USA
Niao He NIAO.HE@INF.ETHZ.CH
ETH Zurich, Switzerland

Abstract

This work revisits the classical low-rank matrix factorization problem and unveils the critical role of initialization in shaping convergence rates for such nonconvex and nonsmooth optimization. We introduce Nyström initialization, which significantly improves the global convergence of Scaled Gradient Descent (ScaledGD) in both symmetric and asymmetric matrix factorization tasks. Specifically, we prove that ScaledGD with Nyström initialization achieves quadratic convergence in cases where only linear rates were previously known. Finally, we equip low-rank adapters (LoRA) with Nyström initialization for practical merits. The effectiveness of the resultant approach, NoRA, is demonstrated on several representative tasks for finetuning large language models (LLMs).

1. Introduction

Compared with learning rates and descent directions, initialization has been a relatively overlooked aspect of optimization. In the widely studied smooth optimization literature [15, 43], as long as a suitable (small) learning rate is chosen, most of optimization algorithms such as gradient descent (GD) provably converge to a stationary point at the same rate, regardless of initialization. This work goes beyond stationary points and highlights the crucial role of initialization for global optimality of Burer-Monteiro factorization [4] – *the same algorithm can exhibit markedly different behaviors, such as linear vs. quadratic convergence, depending on initialization*.

We consider matrix factorization as a canonical example, where the goal is to solve symmetric problems (1a); or the asymmetric ones (1b)

$$\min_{\mathbf{X}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_\mathsf{F}^2, \qquad (1a) \qquad\qquad \min_{\mathbf{X},\mathbf{Y}} \|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|_\mathsf{F}^2. \qquad (1b)$$

While these classical problems can be solved, they are still challenging for optimization, because they are nonconvex, nonsmooth (albeit differentiable), non-coercive (for asymmetric problems), and do not satisfy PL condition [7]. Taking the asymmetric problem (1b) as an example, more refined settings can be categorized based on rank$(\mathbf{A})$ and r. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$. The problem is exact-parametrized (EP) if rank$(\mathbf{A}) = r$, over-parametrized (OP) once rank$(\mathbf{A}) < r$ and under-parametrized (UP) if rank$(\mathbf{A}) > r$. EP and UP are the main focus of this work.

The global convergence of EP has been established for GD, Alternative GD (AltGD) and ScaledGD [12, 25, 60, 66]. The most popular initialization is close to a saddle point $(\mathbf{0}, \mathbf{0})$, i.e.,

Table 1: Comparison of complexity for global optimality in (a)symmetric matrix factorization. Note that our bounds for UP depict the complexity to near optima. Works marked with * are designed for another setting (hence the comparison may not be fair).

setting	algorithm	reference	init.	rate
Asym EP	GD	[66]	small	$\mathcal{O}\big(\kappa^3 \log(1/\epsilon)\big)$
	AltGD	[60]	special	$\mathcal{O}\big(\kappa^2 \log(1/\epsilon)\big)$
	ScaledGD	[53]	local	$\mathcal{O}(\log(1/\epsilon))$
	ScaledGD	**Theorem 5**	Nyström	$\mathcal{O}(1)$
Asym UP	GD	[12]	small	asymptotic
	ScaledGD	**Theorem 6**	Nyström	$\mathcal{O}(1)$
Sym EP	GD*	[51]	small	$\mathcal{O}\big(\kappa^8 + \kappa^2 \log(1/\epsilon)\big)$
	ScaledGD(λ)*	[64]	small	$\mathcal{O}\big(\log^2 \kappa + \log(1/\epsilon)\big)$
	ScaledGD	**Theorem 3**	Nyström	$\mathcal{O}\big(\kappa^3 \sqrt{r} + \log\log(1/\epsilon)\big)$
Sym UP	ScaledGD	**Theorem 4**	Nyström	$\mathcal{O}(r/\epsilon \cdot \log(1/\epsilon))$

$\mathbf{X}_0 \sim \mathcal{N}(0, \zeta_x^2)$ and $\mathbf{Y}_0 \sim \mathcal{N}(0, \zeta_y^2)$ with small ζ_x^2 and ζ_y^2. This initialization prompts linear convergence, and it also holds on our main focus, ScaledGD [25, 53]. In this work, we show that the linear rate of ScaledGD can be improved to a quadratic one under the proposed *Nyström initialization*.

To the best of our knowledge, only an asymptotic global convergence of GD is established for UP in [12]. We prove that with Nyström initialization, ScaledGD converges in a linear rate to the nearing neighbor of a global optimum, and then exhibits a sublinear rate to more fine-grained neighboring area. The improved rates are compared with existing bounds in Tab. 1.

We further extend Nyström initialization to finetune LLMs with LoRA [22]. This is motivated by recent works that use insights from matrix factorization to augment LoRA [65, 68]. Compared with other LoRA initialization methods [5, 41, 59], our Nyström initialization for LoRA (i.e., NoRA) is more economical and better suits for deployment. Our contributions can be summarized as:

- **Faster convergence.** Nyström initialization is provably beneficial for ScaledGD. A quadratic rate can be established on EP, while a (sub)linear rate is obtained for UP until near optimal; see Tab. 1.
- **Role of initialization.** Our results unveil an intriguing phenomenon in nonconvex (nonsmooth) optimization: the behaviors of the same algorithm, reflected via a quadratic vs. linear rate, are critically determined by initialization.
- **Practical tools.** The power of Nyström initialization is further demonstrated on finetuning LLMs. The resultant approach, NoRA, effectively outperforms LoRA on several downstream tasks.

Notation. Bold lowercase (capital) letters denote column vectors (matrices); $(\cdot)^\top$, $(\cdot)^\dagger$ and $\|\cdot\|_\mathsf{F}$ refer to transpose, pseudo inverse, and Frobenius norm of a matrix; $\|\cdot\|$ is the ℓ_2 (spectrum) norm of a vector (matrix); $\sigma_i(\cdot)$ and $\lambda_i(\cdot)$ denote the i-th largest singular value and eigenvalue, respectively.

2. The power of initialization for symmetric matrix factorization

We start to examine the critical role of initialization on symmetric problems (1a). Within this section, we assume that $\mathbf{A} \in \mathbb{R}^{m \times m}$ is positive semidefinite (PSD), otherwise the asymmetric formulation in later sections can be employed. Let $r_A := \text{rank}(\mathbf{A})$ and denote the thin eigendecomposition as $\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{m \times r_A}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r_A \times r_A}$. We employ $\sigma_i(\cdot)$ to denote ith largest singular values. Without loss of generality, the largest and smallest singular values are assumed to be $\sigma_1(\mathbf{A}) = 1$ and $\sigma_{r_A}(\mathbf{A}) = 1/\kappa$ such that the condition number is κ.

ScaledGD as our optimizer. We focus on ScaledGD [53], which is often understood as a preconditioned version of GD; see more details in e.g., [25, 53]. Starting from $t = 0$ with a learning rate $\eta > 0$, ScaledGD updates $\mathbf{X}_t \in \mathbb{R}^{m \times r}$ via

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A})\mathbf{X}_t \cdot (\mathbf{X}_t^\top\mathbf{X}_t)^{-1}. \tag{2}$$

The inversion of $\mathbf{X}_t^\top\mathbf{X}_t$ is affordable given $r \ll m$. Small initialization is widely adopted, i.e., $[\mathbf{X}_0]_{ij} \sim \mathcal{N}(0, \zeta^2)$, where ζ^2 is small. Under such initialization, ScaledGD converges linearly for EP ($r = r_A$), yet no rate is established for UP ($r \leq r_A$) [25]; see more in Tab. 1.

2.1. Nyström initialization

To achieve an improved convergence rate, it is sufficient for EP and UP to ensure that the initialization satisfies two conditions: i) each column of $\mathbf{X}_0$ is in the column space of $\mathbf{A}$, and ii) $\mathbf{X}_0$ is full rank, i.e., $\text{rank}(\mathbf{X}_0) = r$. While the reasons will be uncovered analytically, a straightforward means to meet these requirements is through Nyström sketch [55], i.e.,

$$\textbf{Nyström initialization:} \quad \mathbf{X}_0 = \mathbf{A}\mathbf{\Omega}, \quad \text{where } [\mathbf{\Omega}]_{ij} \sim \mathcal{N}(0, \xi^2), \forall i, \forall j \tag{3}$$

where $\mathbf{\Omega} \in \mathbb{R}^{m \times r}$ is a Gaussian random matrix. From this initialization, it is not difficult to see that requirement i) is satisfied already. Although not stating explicitly, our theorems hold under requirement ii) or $\text{rank}(\mathbf{X}_0) = r$, which is confirmed in the lemma below.

Lemma 1 (Initialization for EP and UP) *There exists a universal constant $\tau > 0$ such that $\sigma_r(\mathbf{X}_0) \geq \xi\tau(\sqrt{r_A} - \sqrt{r-1})\sigma_{r_A}(\mathbf{A})$ is satisfied with high probability, i.e., rank$(\mathbf{X}_0) = r$ w.h.p.*

2.2. Nyström initialization for EP

We start with EP ($r_A = r$). Our first result is the implicit regularization of Nyström initialization.

Lemma 2 *If $\mathbf{X}_0$ is obtained through Nyström initialization (3), update (2) ensures that for all $t \geq 0$*

i) every column of $\mathbf{X}_t$ is in the column space of $\mathbf{A}$, and $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ for some $\mathbf{\Phi}_t \in \mathbb{R}^{r \times r}$; and,

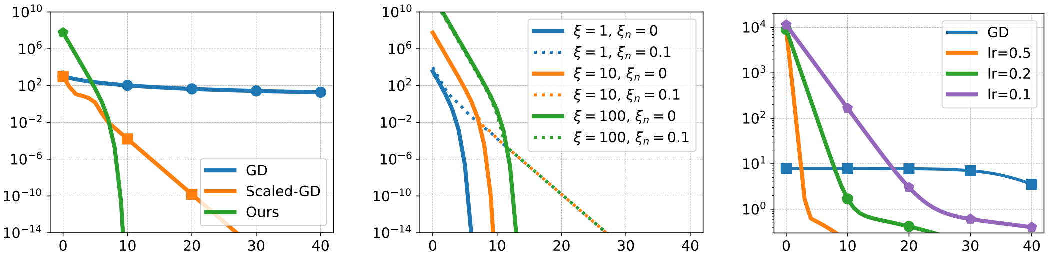
ii) the smallest eigenvalue of $\mathbf{X}_t\mathbf{X}_t^\top$ is bounded away from 0, that is, $\sigma_r(\mathbf{X}_{t+1}\mathbf{X}_{t+1}^\top) \geq (1-\eta)^{2t+2}\sigma_r(\mathbf{X}_0\mathbf{X}_0^\top) + (1-\eta)\sigma_r(\mathbf{A}) - (1-\eta)^{2t+3}\sigma_r(\mathbf{A})$.

Lemma 2 implies that full rankness of $\mathbf{X}_t$ over the trajectory $\text{rank}(\mathbf{X}_t) = r, \forall t$. This ensures an invertible $\mathbf{X}_t^\top\mathbf{X}_t$, that is, iteration (2) is always well-defined. The most important implication of Lemma 2 is the alignment of $\mathbf{X}_t$ with the directions of eigenvectors of $\mathbf{A}$, i.e., $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$. While we will expand this shortly, this alignment in directions enables us to establish a quadratic rate.

Theorem 3 *With Nyström initialization (3), ScaledGD in (2) has a two-phase convergence behavior:*

Phase 1 (linear convergence): Let $\eta = \mathcal{O}(\frac{1}{\kappa^3\|\mathbf{A}\|_\mathsf{F}})$, after $T_1 := \mathcal{O}(\kappa^3\sqrt{r}\log\kappa)$ iterations, ScaledGD ensures that $\|\mathbf{X}_{T_1}\mathbf{X}_{T_1}^\top - \mathbf{A}\|_\mathsf{F} \leq \mathcal{O}(1/\kappa^2)$; and,

Phase 2 (quadratic convergence): After Phase I, ScaledGD converges quadratically with $\eta = 0.5$. In particular, $\|\mathbf{X}_T\mathbf{X}_T^\top - \mathbf{A}\|_\mathsf{F} \leq \epsilon$ is ensured after $T = \mathcal{O}\big(\log\log(\frac{1}{\kappa\epsilon})\big)$ iterations.



(a) EP (b) EP (c) UP

Figure 1: Optimality error vs. iteration of different approaches. (a) GD, and ScaledGD with small / Nyström initialization. (b) solid lines show that our initialization is not sensitive to magnitude; and dotted lines illustrate that quadratic rate cannot be obtained after perturbing the initialization, i.e., $\mathbf{X}_0 = \mathbf{A}\mathbf{\Omega} + \mathbf{N}$, where $[\mathbf{N}]_{ij} \sim \mathcal{N}(0, \xi_n^2)$. (c) GD and ScaledGD with various η for UP.

The quadratic rate of ScaledGD is reflected in Fig. 1 (a) using synthetic data shown in Apdx. 10.1. Notably, the quadratic rate in Theorem 3 is achieved without Hessian on a nonconvex and nonsmooth problem. Moreover, there is no requirement on the magnitude of ξ – initialization does not need to be small. The convergence of ScaledGD under different ξs can be found in (the solid lines of) Fig. 1(b).

The critical role of initialization. As shown in Lemma 2, Nyström initialization aligns $\mathbf{X}_t$ with the directions of eigenvectors $\mathbf{Q}$, thereby eliminating the residual space, i.e., $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t = \mathbf{0}, \forall t$. This is in stark contrast with most of existing works on matrix factorization, where small initialization only ensures $\|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t\|_\mathsf{F} \to 0$ in a linear rate [12, 25, 66]. Getting rid of the residual space enables a quadratic rate of ScaledGD. This is illustrated in Fig. 1 (b), where small noise is injected to the residual space by slightly perturbing Nyström initialization. Reflected in the dotted lines, even if the earlier convergence does not differ with Nyström initialization, only a linear rate is observed.

2.3. Nyström initialization for UP

Next, we consider the case of UP of (1a), i.e., $r < r_A$. We will show that ScaledGD converges under weak optimality, that is, $\mathbf{X}^\top \mathbf{A}^\dagger \mathbf{X} - \mathbf{I}_r = \mathbf{0}$. Due to space limitation, we prove in Lemma 9 of Apdx. 7.3 that all global optima are also weak optima. Moreover, Lemma 10 shows that $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t = \mathbf{0}, \forall t$ also holds for UP, i.e., Nyström initialization eliminates the residual space. Building upon this, the convergence of ScaledGD can be established.

Theorem 4 *Depending on η, ScaledGD (2) with Nyström initialization (3) ensures that*

i) (Linear convergence to neighborhood of weak optima.) If one chooses $\eta \leq 1$, ScaledGD ensures that $\|\mathbf{X}_t^\top \mathbf{A}^\dagger \mathbf{X}_t - \mathbf{I}_r\|_\mathsf{F} \leq \mathcal{O}(\eta r) + \epsilon$ in $\mathcal{O}(\log \frac{1}{\epsilon})$ iterations; or,

ii) (Convergence to weak optima.) Let $\eta = \mathcal{O}(\epsilon / r)$, weak optimality is ensured by ScaledGD after $\mathcal{O}(\frac{r}{\epsilon} \log \frac{1}{\epsilon})$ iterations, i.e., $\|\mathbf{X}_t^\top \mathbf{A}^\dagger \mathbf{X}_t - \mathbf{I}_r\|_\mathsf{F} \leq \epsilon$.

Fig. 1 (c) illustrates the linear convergence. We also prove that at convergence ScaledGD ensures $\mathbf{X}_t$ to stay close to a global solution, and the distance is sublinear in r in Lemma 11 in appendix.

3. The power of initialization for asymmetric matrix factorization

This section demonstrates that the power of initialization is even more striking in asymmetric matrix factorization (1b), where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$. Moreover, denote $\text{rank}(\mathbf{A}) = r_A$ and the thin SVD be $\mathbf{A} := \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times r_A}$, $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$, and $\mathbf{V} \in \mathbb{R}^{n\times r_A}$.

Nyström initialization. We adopt an asymmetric manner to initialize $\mathbf{X}_0$ and $\mathbf{Y}_0$ for (1b), i.e.,

$$\textbf{Nyström initialization:}\quad \mathbf{X}_0 = \mathbf{A}\mathbf{\Omega}, \quad \mathbf{Y}_0 = \mathbf{0} \tag{4}$$

where $\mathbf{\Omega}$ is a Gaussian random matrix of $\mathbb{R}^{n\times r}$ with $[\mathbf{\Omega}]_{ij} \sim \mathcal{N}(0, \xi^2), \forall i, \forall j$.

Modified ScaledGD. To adapt to the non-invertible $\mathbf{Y}_0^\top\mathbf{Y}_0 = \mathbf{0}$ in Nyström initialization (4), we modify the first iteration of ScaledGD. More precisely, the updates are summarized below

$$\mathbf{X}_1 = \mathbf{X}_0, \text{ and } \mathbf{X}_{t+1} = \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t(\mathbf{Y}_t^\top\mathbf{Y}_t)^{-1}, \forall t \geq 1; \tag{5a}$$

$$\mathbf{Y}_{t+1} = \mathbf{Y}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})^\top\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t)^{-1}, \forall t \geq 0. \tag{5b}$$

3.1. Nyström initialization for EP

We start with EP, i.e., $r_A = r$ in (1b). The merit of Nyström initialization (4) is the elimination of residual space, i.e., $(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{X}_t = \mathbf{0}$ and $(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\mathbf{Y}_t = \mathbf{0}$ as shown in Lemma 14.

Theorem 5 (One-step convergence) *With $\eta = 1$ and Nyström initialization (4), the modified ScaledGD (5) guarantees $\mathbf{X}_1\mathbf{Y}_1^\top = \mathbf{A}$. In other words, global convergence is achieved in one step.*

3.2. Nyström initialization for UP

Lastly, we tackle the case of UP for the asymmetric problem (1b), where $r_A > r$. Here we consider generalized weak optimality, which is defined as $\mathbf{Y}^\top\mathbf{A}^\dagger\mathbf{X} - \mathbf{I}_r = \mathbf{0}$. Generalized weak optimality is satisfied by any global optimum as proved in Lemma 15. Now we are ready to show the convergence.

Theorem 6 *If $\eta = 1$, ScaledGD in (5) with Nyström initialization (4) ensures generalized weak optimality in one step, i.e., $\mathbf{Y}_1^\top\mathbf{A}^\dagger\mathbf{X}_1 - \mathbf{I}_r = \mathbf{0}$.*

The critical role of initialization. Through the theoretical analyses in the previous two sections, it is evident that the convergence of ScaledGD for matrix factorization is *highly dependent on the initialization*. Here is an intuitive, though not strictly rigorous, summary: Small initialization results in behaviors similar to first-order optimizers, i.e., linear convergence [25]. In contrast, the proposed Nyström initialization catalyzes quadratic rates and even one-step convergence, resembling the optimization trajectory of second-order approaches such as Newton's method [43].

4. NoRA: Nyström low rank adapters

We extend the benefit of Nyström initialization to another setting based on Burer-Monteiro factorization, i.e., low-rank adapters (LoRA) in finetuning deep neural networks [22]. Due to limited space, the detailed explanation of our methodology is deferred to Apdx. 6.3 and Apdx. 6.4, and here we only summarize the proposed approach:

- **Nyström LoRA (NoRA)**: Simply apply (4) on top of LoRA, that is, $\mathbf{X}_0 = \mathbf{A}\mathbf{\Omega}$ and $\mathbf{Y}_0 = \mathbf{0}$, where $\mathbf{A}$ is the pretrained weights.

Table 2: Test accuracy of various algorithms for commonsense reasoning on LLaMA2-7B.

LLaMA2-7B	BoolQ	PIQA	SIQA	HS	WG	ARC-e	ARC-c	OBQA	avg (↑)
LoRA†	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
LoRA-P	71.47	81.50	78.81	85.97	80.43	81.14	66.55	81.0	78.35
NoRA	71.16	83.08	79.53	85.90	81.85	80.64	66.13	81.80	78.76
NoRA+	70.52	81.94	79.07	87.66	82.24	82.70	67.06	80.2	**78.92**

- **Nyström preconditioned LoRA (NoRA+)**: this approach not only advances LoRA initialization with (4), but also leverages ScaledGD for optimization.

We note that ScaledGD has already been applied for LoRA training [68], which this approach is referred to as LoRA-P in our work (P for precondtioning). We will show that both LoRA and LoRA-P benefit from Nyström initialization.

Deployment efficiency. Nyström initialization is more economical than the SVD or QR used in other initialization methods such as PiSSA and OLoRA [5, 41]. Furthermore, NoRA requires no modification to the pretrained weights, making it an off-the-shelf solution without altering existing LoRA pipelines. We expand on this in Apdx. 6.4.

4.1. Numerical results for NoRA

The efficiency of proposed NoRA and NoRA+ is demonstrated using LLaMA2-7B [54]. Additional details on the datasets and experimental procedures can be found in Apdx. 10. We tackle commonsense reasoning following the setup in [23]. The rank of LoRA is chosen as 32. Numerical results on LLaMA2-7B are presented in Tab. 2. It is observed that LoRA is unstable, henceforth the results for LoRA are taken from [38]. This instability is not observed in the other approaches tested. NoRA and NoRA+ outperform LoRA and LoRA-P, demonstrating the efficiency of Nyström initialization.

Additional numerical results. More on NoRA on finetuning OPT [72] can be found in Apdx. 10.3, where the performance of NoRA is also compared with PiSSA and OLoRA [5, 41].

5. Concluding remarks and future directions

This work characterizes how initialization can crucially determine the convergence behavior of the same optimization algorithm on matrix factorization problems. We prove that Nyström initialization can significantly improve the complexity bounds of ScaledGD under a wide spectrum of settings; see details in Tab. 1. One of the key improvements is that Nyström initialization enables a quadratic convergence for exact-parametrized problems, whereas small initialization only guarantees a linear rate on ScaledGD. This performance gap calls for more careful investigation into the role of initialization in optimization. Additionally, the proposed Nyström initialization offers practical merits when applied on finetuning with LoRA, delivering deployment flexibility and promising numerical performance on large-scale problems.

Future work. While this work focuses on the impact of initialization in canonical matrix factorization problems, we believe our results extend to more complex settings, such as matrix sensing and tensor factorization, which are part of our future research plans. Additionally, in the context of Burer-Monteiro factorization with LoRA, our work suggests potential gains by exploiting priors embedded in pretrained weights. Investigating how to better uncover and utilize this hidden information is another attractive direction for future research.

References

[1] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2018.

[2] Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. *arXiv:2405.17604*, 2024.

[3] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proc. AAAI Conf. Artif. Intel.*, pages 7432–7439, 2020.

[4] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming*, 95(2):329–357, 2003.

[5] Kerim Büyükakyüz. OLoRA: Orthonormal low-rank adaptation of large language models. *arXiv:2406.01775*, 2024.

[6] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Long-LoRA: Efficient fine-tuning of long-context large language models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

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

[8] François Chollet. On the measure of intelligence. *arXiv:1911.01547*, 2019.

[9] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv:1905.10044*, 2019.

[10] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proc. Sinn und Bedeutung*, 23(2): 107–124, 2019.

[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[12] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[13] Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[14] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 1233–1242. PMLR, 2017.

[15] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[17] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[18] Yongchang Hao, Yanshuai Cao, and Lili Mou. FLORA: Low-rank adapters are secretly gradient compressors. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[19] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on lora finetuning dynamics. *arXiv:2406.08447*, 2024.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[21] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 2790–2799, 2019.

[22] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

[23] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[24] Uijeong Jang, Jason D Lee, and Ernest K Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[25] Xixi Jia, Hailin Wang, Jiangjun Peng, Xiangchu Feng, and Deyu Meng. Preconditioning matters: Fast global convergence of non-convex matrix factorization via scaled gradient descent. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2023.

[26] Kaiqi Jiang, Dhruv Malik, and Yuanzhi Li. How does adaptive optimization impact local neural network geometry? *Proc. Neural Information Processing Systems (NeurIPS)*, 36, 2023.

[27] Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5(3):723–744, 2023.

[28] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proc. Conf. Assoc. Comput. Linguist. Meet.*, pages 252–262, 2018.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2014.

[30] Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

[31] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.

[32] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proc. intl. conf. on Principles of Knowledge Representation and Reasoning*, 2012.

[33] Bingcong Li, Liang Zhang, and Niao He. Implicit regularization of sharpness-aware minimization for scale-invariant problems. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2024.

[34] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proc. Conf. Assoc. Comput. Linguist. Meet.*, pages 4582–4597, 2021.

[35] Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

[36] Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

[37] Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham Krishna Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. Svft: Parameter-efficient fine-tuning with singular vectors. *arXiv:2405.19597*, 2024.

[38] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

[40] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[41] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. *arXiv:2404.02948*, 2024.

[42] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. *arXiv:1809.02789*, 2018.

[43] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

[44] Mahdi Nikdan, Soroush Tabesh, and Dan Alistarh. RoSA: Accurate parameter-efficient fine-tuning via robust adaptation. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[45] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, 2016.

[46] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

[47] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[48] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv:1904.09728*, 2019.

[49] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-LoRA. *arXiv:2304.06027*, 2023.

[50] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

[51] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, pages 23831–23843, 2021.

[52] Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 10153–10161, 2021.

[53] Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22(150):1–63, 2021.

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[55] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Fixed-rank approximation of a positive-semidefinite matrix from streaming data. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[56] Aditya Vardhan Varre, Maria-Luiza Vladarean, Loucas Pillaud-Vivien, and Nicolas Flammarion. On the spectral bias of two-layer linear networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[57] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[58] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

[59] Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. *arXiv:2407.05000*, 2024.

[60] Rachel Ward and Tamara Kolda. Convergence of alternating gradient descent for matrix factorization. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 36, pages 22369–22382, 2023.

[61] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proc. Annual Conf. on Learning Theory (COLT)*, pages 3635–3673, 2020.

[62] Wenhan Xia, Chengwei Qin, and Elad Hazan. Chain of LoRA: Efficient fine-tuning of language models via residual learning. *arXiv:2401.04151*, 2024.

[63] Nuoya Xiong, Lijun Ding, and Simon S Du. 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.

[64] Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 38611–38654, 2023.

[65] Can Yaras, Peng Wang, Laura Balzano, and Qing Qu. Compressible dynamics in deep overparameterized low-rank learning & adaptation. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[66] Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, pages 1429–1439, 2021.

[67] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv:1905.07830*, 2019.

[68] Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[69] Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex matrix factorization. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, pages 5985–5996, 2021.

[70] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.

[71] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv:1810.12885*, 2018.

[72] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv:2205.01068*, 2022.

[73] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[74] Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Supplementary Document for "On the Crucial Role of Initialization for Matrix Factorization"

6. Missing details

6.1. More on related work

Theoretical merits of initialization. Despite the merits of initialization are widely recognized in practice [16, 20, 31], it is quite challenging to theoretically characterize the impact of initialization. In continuous optimization, it is shown that the magnitude of initialization relates to the implicit regularization of gradient flow on overparametrized problems [56, 61]. Initialization also impacts the robustness of ReLU neural networks [74]. Our work provides additional evidence on this regard, demonstrating that the same algorithm can have linear or quadratic rate under different initialization.

Matrix factorization from an optimization perspective. The goal of this work is to recap this classical problem and to unveil intriguing behaviors from an optimization perspective. This problem entails rich behaviors of loss landscape – nonconvexity, non-smooth and non-PL-ness. Recent works have examined the convergence of several classical algorithms, such as GD, AltGD and ScaledGD; see e.g., [12, 27, 60, 66] and Tab. 1. Most of previous works focus on small initialization, and some analysis techniques therein are tailored thus difficult to generalize. Our Nyström initialization enables us to derive faster convergence of ScaledGD in EP and UP settings within a unified framework. The initialization in AltGD [60] also adopts sketch, i.e., $\mathbf{X}_0 = \mathcal{O}(\mathbf{A}\mathbf{\Omega}_1/\sigma_1(\mathbf{A}))$ and $\mathbf{Y}_0 = \mathcal{O}(\sigma_1(\mathbf{A})\mathbf{\Omega}_2)$, where $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are Gaussian random matrices with small variance. Besides the requirement on the variance of $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ and the explicit need of $\sigma_1(\mathbf{A})$, this initialization cannot eliminate of residual space as ours in (4). Consequently, AltGD demands early stopping in EP, and little is known for UP.

Convergence of overparametrized matrix factorization problems. Consider again the asymmetric problem as an example, i.e., $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|^2$ with $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$. Overparametrization (OP) refers to the case where $\text{rank}(\mathbf{A}) \leq r$. Alternative gradient descent is considered in [60], where it obtains a linear convergence rate. The work of [27] shows that GD recovers singular values of $\mathbf{A}$ in a sequential manner in a different but related setting. The gradient flow on the extreme OP problem, where $r \geq \max\{m, n\}$, is studied in [52]. There are also papers [51, 63, 69] considering the matrix sensing problem, which partially relates to our problem when there are sufficient Gaussian measures. The work of [1] considers deeper problem (i.e., having more than 3 layers) while assuming $\mathbf{A}$ is full rank.

LoRA and parameter-efficient finetuning. LoRA [22] is a notable example of parameter-efficient finetuning (PEFT) approaches. The goal of PEFT is to reduce the resource requirement for finetuning LLMs on downstream tasks. Other commonly adopted PEFT methods include, e.g., adapters [21] and prefix tuning [34]. There are also various efforts to further enhance LoRA via adaptivity [70], chaining [36, 62], low-bit training [11, 35], modifications for long-sequences [6], weight decomposition [38], regularization [33], and combining with sparsity [44]. Additionally, there are several approaches aiming at further reducing the number of trainable parameters in LoRA; examples include [2, 13, 18, 30, 37, 73]. While originally designed for finetuning LLMs, LoRA also finds its applications in other domains, such as image generation [17] and continual learning [49].

LoRA initialization. When first proposed, LoRA initialization was largely overlooked. The work of [19] justifies that whether setting $\mathbf{X}_0$ or $\mathbf{Y}_0$ to be $\mathbf{0}$ from a stability perspective. Recent works [5, 41] observe a fundamental difference between initialization of LoRA and neural networks, emphasizing the availability of prior knowledge. These works experimentally demonstrate that pretrained model can serve as prior to guide the direction of adapters, and hence perform QR or SVD on the pretrained matrix and using (scaled) top-r singular vectors for LoRA initialization. Follow-up study [59] exploits stability for further improvement. However, these initialization methods are computationally expensive and lack flexibility for deployment. The proposed NoRA initialization overcomes these limitations.

6.2. LoRA for linear models as asymmetric matrix factorization

We argue that the asymmetric matrix factorization problem is equivalent to LoRA applied on linear models given a whitened dataset. The whitened dataset is widely adopted for theoretical analyses, and we refer to [1, 26, 65] for more details.

Assume that we have a pretrained (linear) model $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$. Applying LoRA on this layer with whitened data $\mathbf{B}$ is equivalent to solving the following problem

$$\frac{1}{2}\|(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top) - \mathbf{B}\|_{\mathsf{F}}^2. \tag{6}$$

It is clearly that this problem (6) is the same as (1b) by setting $\mathbf{A} = \mathbf{B} - \mathbf{W}_0$.

Unfortunately, initialization approaches in existing theoretical works have no support on the most widely adopted one in practice – either $\mathbf{X}_0$ or $\mathbf{Y}_0$ is chosen as $\mathbf{0}$ to preserve $\mathbf{W}_0 + \mathbf{X}_0\mathbf{Y}_0^\top = \mathbf{W}_0$. In this sense, our Nyström initialization in (4) is the first means of initialization that justifies one variable can be set to $\mathbf{0}$.

Additional similarities between LoRA and matrix factorization. LoRA and matrix factorization share similar mathematical properties. For example, they both have no spurious local minima [12, 14, 24]. There are also recent efforts using insights from matrix factorization to further improve LoRA; see e.g., [44, 65].

6.3. Nyström low rank adapters

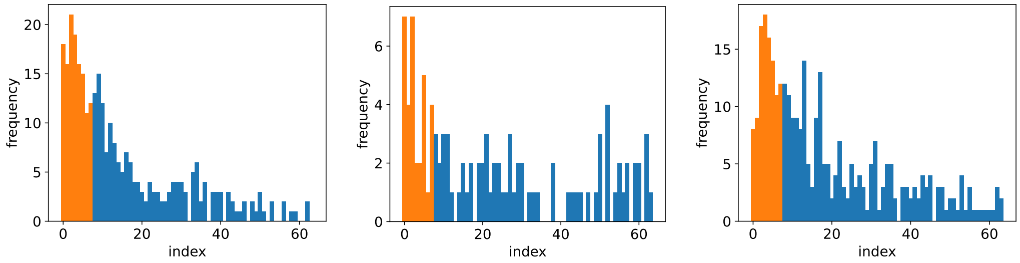
Our theoretical results highlight the merits of suitable initialization for matrix factorization (1b). One of the key insights is that the Burer-Monteiro factorization benefits from good directions of $\mathbf{X}_0$ and $\mathbf{Y}_0$ at initialization; cf. Lemmas 2 and 14. We term this as *directional alignment*. Here we extend the benefit of Nyström initialization to low-rank adapters (LoRA) in finetuning deep neural networks [22]. This is the full version of Sec. 4.

LoRA enhances parameter efficiency of finetuning by approximating the thought parameter-change $\Delta\mathbf{W} \in \mathbb{R}^{m\times n}$ via Burer-Monteiro factorization

$$\mathbf{W}_0 + \Delta\mathbf{W} \approx \mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top \tag{7}$$

where $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ is the pretrained weight (of a particular layer), and $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$ with $r \ll \min\{m, n\}$. A more detailed recap of LoRA can be found in Apdx. 6.1. Directional alignment can be achieved if singular vectors for $\Delta\mathbf{W}$ are leveraged to initialize $\mathbf{X}_0$ and $\mathbf{Y}_0$. While $\Delta\mathbf{W}$ is unavailable a priori, empirical wisdom suggests that there exist a set of well-performed adapters that lie in the column (row) span of the pretrained weight matrix [37], i.e., $\text{ColSpan}(\Delta\mathbf{W}) \subseteq \text{ColSpan}(\mathbf{W}_0)$ and $\text{RowSpan}(\Delta\mathbf{W}) \subseteq \text{RowSpan}(\mathbf{W}_0)$. In other words, $\mathbf{W}_0$ can be adopted as a suitable replacement of $\Delta\mathbf{W}$ for directional alignment.

Since directional alignment requires at most r directions for $\mathbf{X}_0$ (and $\mathbf{Y}_0$), the next question is how to identify them out of the m (or n) singular vectors of $\mathbf{W}_0$. This prompts us to examine the



(a) BoolQ (b) ReCoRD (c) RTE

Figure 2: Which singular values have the largest change after finetuning with LoRA of rank r? Orange: top-r singular values; blue: other singular values. Note that here we only plot the first 64 singular values as the rests rarely have sufficiently large change.

singular values that undergo the most significant change after LoRA finetuning. We evaluate LoRA on a few-shot learning task [40], with the detailed setup provided shortly in Apdx. 10.3. OPT-1.3B is chosen as the base model and LoRA rank is set to $r = 8$. We focus on the (index of) r singular values that exhibit the most significant changes after finetuning and summarize their frequencies in Fig. 2. It is observed that the top-r singular values tend to exhibit larger change, explaining the success of LoRA initialization approaches such as PiSSA and OLoRA [5, 41]. However, across all tested datasets, a substantial portion of non-top-r singular-values also demonstrate significant variation. Another observation from Fig. 2 is that the relative importance of singular directions is roughly proportional to singular values. Note that is akin to Nyström initialization, i.e., $\mathbb{E}[(\mathbf{W}_0\mathbf{\Omega})(\mathbf{W}_0\mathbf{\Omega})^\top] \propto \mathbf{W}_0\mathbf{W}_0^\top$.

6.4. More on NoRA and NoRA+

As discussed in Section 4, LoRA can significantly benefit from the aligned directions at initialization. Besides the theoretical benefits of applying Nyström initialization on ScaledGD (NoRA+), Nyström initialization can also be used directly with Adam (or AdamW), i.e., NoRA. There are several reasons for this. First, directional alignment from initialization is benefit to most optimizers. While our theoretical results focus on ScaledGD, we believe that the aligned directions also benefit GD. Despite the improvement may be less significant as in ScaledGD, we conjecture that the linear term in [66, Theorem 1.1] can be removed with Nyström initialization, because it can be roughly understood as the price for searching for proper directions. In other words, the benefits of Nyström initialization extend to other optimizers as well. Second, Adam also affords an explanation of preconditioning, and the preconditioner for $\mathbf{X}_t$ is also closely related to $\mathbf{Y}_t$. In other words, Adam shares similarities with ScaledGD in (5). These two reasons prompt the proposed NoRA, as summarized in Alg. 1. For NoRA+ in Alg. 2, we modify the vanilla ScaledGD iterations in (5) with two add-ons. First, a small parameter λ is introduced for numerical stability of matrix inversion. This is a standard practice for numerical optimizers such as Adam [29, 39]. Second, the gradient is normalized by the Frobenius norm of its preconditioner. The reason is that an optimal λ is difficult to tune as shown in [68], where they use λ from 10^{-6} to 100. With this normalizer, we can set $\lambda = 10^{-6}$ in all our experiments without any tuning. Moreover, this normalizer is useful to prevent the instability in earlier iterations due to the non-invertable $\mathbf{Y}_0 = \mathbf{0}$.

Deployment efficiency of NoRA. One benefit of NoRA (as well as NoRA+) is that it can be deployed jointly with adapters trained with LoRA – and hence there is no need to modify the current

Algorithm 1 NoRA for a specific LoRA layer

```
**Initialize:** $\xi$ – standard deviation of random matrix $\mathbf{\Omega}$
Set $\mathbf{X}_0$ and $\mathbf{Y}_0$ via Nyström initialization (4)
Standard training process
```

Algorithm 2 NoRA+ for a specific LoRA layer

```
**Initialize:** $\xi$ – standard deviation of random matrix $\mathbf{\Omega}$; $\lambda$ – numerical stability of matrix inversion
Set $\mathbf{X}_0$ and $\mathbf{Y}_0$ via Nyström initialization (4)
**for** $t = 0, \ldots, T-1$
    Get gradient $\mathbf{G}_{\mathbf{X}_t}$ and $\mathbf{G}_{\mathbf{Y}_t}$
    **if** $t > 0$ **then**
        $\mathbf{G}_{\mathbf{X}_t} \leftarrow \mathbf{G}_{\mathbf{X}_t}(\mathbf{Y}_t^\top\mathbf{Y}_t + \lambda\mathbf{I}_r)^{-1}/\|(\mathbf{Y}_t^\top\mathbf{Y}_t + \lambda\mathbf{I}_r)^{-1}\|_\mathsf{F}$
    **end if**
    $\mathbf{G}_{\mathbf{Y}_t} \leftarrow \mathbf{G}_{\mathbf{Y}_t}(\mathbf{X}_t^\top\mathbf{X}_t + \lambda\mathbf{I}_r)^{-1}/\|(\mathbf{X}_t^\top\mathbf{X}_t + \lambda\mathbf{I}_r)^{-1}\|_\mathsf{F}$
    Optimizer update
**end for**
```

pipeline for deployment. This is because both of NoRA and LoRA do not need to modify the trained parameters, and the finetuned model is just $\mathbf{W}_0 + \mathbf{X}_T\mathbf{Y}_T^\top$, where $\mathbf{W}_0$ is the pretrained model, and $\mathbf{X}_T$ and $\mathbf{Y}_T$ are finetuned adapter weights. On the contrary, other initialization approaches such as PiSSA and OLoRA [5, 41] are less efficient for using jointly with LoRA at deployment because both approaches modify the pretrained weight, so that the finetuned model becomes $\widehat{\mathbf{W}}_0 + \mathbf{X}_T\mathbf{Y}_T^\top$, where $\widehat{\mathbf{W}}_0 = \mathbf{W}_0 - \mathbf{X}_0\mathbf{Y}_0^\top$. The use of $\widehat{\mathbf{W}}_0$ comes from the fact that initialization in PiSSA and OLoRA does not satisfy $\mathbf{X}_0\mathbf{Y}_0^\top = \mathbf{0}$. Consequently, when deploying PiSSA jointly with LoRA, one needs to store both $\mathbf{W}_0$ (for LoRA) and $\widehat{\mathbf{W}}_0$ (for PiSSA), leading to reduced memory efficiency.

7. Missing proofs for symmetric settings

7.1. Initialization of EP and UP

7.1.1. PROOF OF LEMMA 1

Proof Let the thin eigenvalue decomposition of $\mathbf{A}$ be $\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{m \times r_A}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r_A \times r_A}$. We then have that

$$\mathbf{X}_0 = (\mathbf{Q}\mathbf{\Sigma})(\mathbf{Q}^\top\mathbf{\Omega}). \tag{8}$$

It is not hard to verify that the matrix $\mathbf{Q}^\top\mathbf{\Omega} \in \mathbb{R}^{r_A \times r}$ is also a Gaussian random matrix, where each entry follows $\mathcal{N}(0, \xi^2)$. Applying Lemma 20 on $\mathbf{Q}^\top\mathbf{\Omega}$, it can be seen that

$$\mathbb{P}\Big(\frac{\sigma_r(\mathbf{Q}^\top\mathbf{\Omega})}{\xi} \le \tau(\sqrt{r_A} - \sqrt{r-1})\Big) \le (C_1\tau)^{r_A - r + 1} + e^{-C_2 r_A} := \delta$$

where C_1 and C_2 are universal constants independent of r_A and r. This inequality shows that with probability at least $1 - \delta$, $\sigma_r(\mathbf{Q}^\top\mathbf{\Omega}) \ge \xi\tau(\sqrt{r_A} - \sqrt{r-1})$.

Note that inequality $\sigma_{\min}(\mathbf{C}\mathbf{D}) \ge \sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})$ holds given full column rank of $\mathbf{C}$. Applying it to (8), we have that

$$\begin{aligned}\sigma_r(\mathbf{X}_0) &\ge \sigma_{r_A}(\mathbf{Q}\mathbf{\Sigma})\sigma_r(\mathbf{Q}^\top\mathbf{\Omega}) = \sigma_{r_A}(\mathbf{A})\sigma_r(\mathbf{Q}^\top\mathbf{\Omega}) \\ &\overset{(a)}{\ge} \xi\tau(\sqrt{r_A} - \sqrt{r-1})\sigma_{r_A}(\mathbf{A})\end{aligned}$$

where (a) holds with probability at least $1 - \delta$. ■

7.2. Missing proofs for the symmetric and EP setting

In the EP setting, it is convenient to define

$$\mathbf{B}_t := \mathbf{\Phi}_t \mathbf{\Phi}_t^\top \tag{9}$$

where $\mathbf{\Phi}_t \in \mathbb{R}^{r \times r}$ comes from Lemma 2, i.e., $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$. The notation $\mathbf{B}_t$ will be used frequently in the proofs in this subsection. With the help of Lemma 2, $\mathbf{B}_t$ can be understood as the "core" part of $\mathbf{X}_t\mathbf{X}_t^\top$, because $\mathbf{X}_t\mathbf{X}_t^\top = \mathbf{Q}\mathbf{\Phi}_t\mathbf{\Phi}_t^\top\mathbf{Q}^\top = \mathbf{Q}\mathbf{B}_t\mathbf{Q}^\top$. Once proving Lemma 2, it allows us to study dynamics using a simpler but equivalent notion $\|\mathbf{B}_t - \mathbf{\Sigma}\|_\mathsf{F}$, i.e.,

$$\|\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A}\|_\mathsf{F} = \|\mathbf{Q}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{Q}^\top\|_\mathsf{F} = \|\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma}\|_\mathsf{F} = \|\mathbf{B}_t - \mathbf{\Sigma}\|_\mathsf{F}.$$

7.2.1. PROOF OF LEMMA 2

Proof The proof relies on $\mathbf{B}_t$ defined in (9). We will prove this lemma by induction. Since $\mathbf{X}_0 = \mathbf{A}\mathbf{\Omega}$ in Nyström initialization, we have that $\mathbf{\Phi}_0 = \mathbf{\Sigma}\mathbf{Q}^\top\mathbf{\Omega}$. Moreover, our base assumption $\sigma_r(\mathbf{B}_0) > 0$ is true because $\text{rank}(\mathbf{B}_0) = \text{rank}(\mathbf{X}_0\mathbf{X}_0^\top) = r$, which is the result of Lemma 1.

For induction, assume that $\mathbf{X}_t$ can be written as $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ with a full rank $\mathbf{\Phi}_t \in \mathbb{R}^{r\times r}$ at iteration t. By the update (2), we have that

$$\begin{aligned}
\mathbf{X}_{t+1} &= \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A})\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t)^{-1} \\
&= \mathbf{Q}\mathbf{\Phi}_t - \eta\mathbf{Q}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{Q}^\top\mathbf{Q}\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{\Phi}_t)^{-1} \\
&\overset{(a)}{=} \mathbf{Q}\Big(\mathbf{\Phi}_t - \eta(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}\Big) \\
&\overset{(b)}{=} \mathbf{Q}\underbrace{\Big((1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}\Big)}_{:=\mathbf{\Phi}_{t+1}}
\end{aligned} \tag{10}$$

where (a) uses $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_r$; and (b) uses $\mathbf{\Phi}_t$ is full rank (hence invertible). Note that $\mathbf{Q}$ and $\mathbf{A}$ share the same column space. This proves the first claim i) of this lemma.

Next we show that the smallest eigenvalue of $\mathbf{B}_{t+1}$ is bounded away from 0, or equivalently, $\mathbf{\Phi}_{t+1}$ is full rank. To start with, we have that from the expression of $\mathbf{\Phi}_{t+1}$ in (10),

$$\begin{aligned}
\mathbf{B}_{t+1} = \mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top &= (1-\eta)^2\mathbf{\Phi}_t\mathbf{\Phi}_t^\top + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}\mathbf{\Phi}_t^{-1}\mathbf{\Sigma} \\
&= (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma}.
\end{aligned} \tag{11}$$

Note that $\mathbf{B}_{t+1}$ is a PSD matrix by definition (hence the eigen values and singular values are the same). To see the smallest eigenvalue of $\mathbf{B}_{t+1}$ is lower bounded, we will apply Lemma 17 on (11) twice, i.e.,

$$\begin{aligned}
&\sigma_r(\mathbf{B}_{t+1}) \qquad\qquad (12)\\
&\overset{(c)}{\geq} 2\eta(1-\eta)\sigma_r(\boldsymbol{\Sigma}) + \sigma_r\Big((1-\eta)^2\mathbf{B}_t + \eta^2\boldsymbol{\Sigma}\mathbf{B}_t^{-1}\boldsymbol{\Sigma}\Big)\\
&\overset{(d)}{\geq} 2\eta(1-\eta)\sigma_r(\boldsymbol{\Sigma}) + (1-\eta)^2\sigma_r\big(\mathbf{B}_t\big)\\
&\overset{(e)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + 2\eta(1-\eta)\sigma_r(\boldsymbol{\Sigma})\frac{1-(1-\eta)^{2t+2}}{2\eta-\eta^2}\\
&\overset{(f)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + (1-\eta)\sigma_r(\boldsymbol{\Sigma}) - (1-\eta)^{2t+3}\sigma_r(\boldsymbol{\Sigma})
\end{aligned}$$

where (c) and (d) are because of Lemma 17; (e) is by unrolling $\sigma_r(\mathbf{B}_t)$ using (d); and (f) is by $\frac{2\eta}{2\eta-\eta^2} \geq 1$. Combining (10) and (12) concludes the induction. ■

7.2.2. PROOF OF THEOREM 3

Proof The proof is by combining Lemmas 7 and 8. ■

Lemma 7 (Phase I. Linear convergence near optimal.) *Let $\eta = \mathcal{O}(\frac{1}{\kappa^3\|\mathbf{A}\|_{\mathsf{F}}})$. After $\mathcal{O}(\kappa^3\sqrt{r}\log\kappa)$ iterations, ScaledGD (2) with Nyström initialization (3)) ensures that $\|\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A}\|_{\mathsf{F}} \leq \mathcal{O}(1/\kappa^2)$.*

Proof Subtracting $\boldsymbol{\Sigma}$ from both sides of (11), we can obtain that

$$\mathbf{B}_{t+1} - \boldsymbol{\Sigma} = (1-\eta)^2(\mathbf{B}_t - \boldsymbol{\Sigma}) - \eta^2\boldsymbol{\Sigma} + \eta^2\boldsymbol{\Sigma}\mathbf{B}_t^{-1}\boldsymbol{\Sigma}.$$

This implies that

$$\begin{aligned}
&\|\mathbf{B}_{t+1} - \boldsymbol{\Sigma}\|_{\mathsf{F}}\\
&\overset{(a)}{\leq} (1-\eta)^2\|\mathbf{B}_t - \boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\|\boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\|\boldsymbol{\Sigma}\mathbf{B}_t^{-1}\|_2\|\boldsymbol{\Sigma}\|_{\mathsf{F}}\\
&\overset{(b)}{\leq} (1-\eta)^2\|\mathbf{B}_t - \boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\|\boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\|\boldsymbol{\Sigma}\|_2\|\mathbf{B}_t^{-1}\|_2\|\boldsymbol{\Sigma}\|_{\mathsf{F}}\\
&\leq (1-\eta)\|\mathbf{B}_t - \boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\|\boldsymbol{\Sigma}\|_{\mathsf{F}} + \eta^2\frac{\sigma_1(\boldsymbol{\Sigma})\|\boldsymbol{\Sigma}\|_{\mathsf{F}}}{\sigma_r(\mathbf{B}_t)}
\end{aligned}$$

where (a) is by $\|\mathbf{MN}\|_{\mathsf{F}} \leq \|\mathbf{M}\|_2\|\mathbf{N}\|_{\mathsf{F}}$; and (b) follows from the sub-multiplicity of $\|\cdot\|_2$.

By Lemma 2, there exists $T_1 = \mathcal{O}(\frac{1}{\eta})$ such that $\sigma_r(\mathbf{B}_t) \geq \sigma_r(\boldsymbol{\Sigma})/3, \forall t \geq T_1$. To avoid such complexity, one can simply choose the step size to be $\eta_1 = 0.5$ until this is achieved, and then use other desirable step sizes. Alternatively, we can choose ξ in (3) sufficiently large such that

$\sigma_r(\mathbf{B}_0) \geq \sigma_r(\mathbf{\Sigma})/3$, i.e., $T_1 = 0$. Our proof below goes with the second method, i.e., $T_1 = 0$.

$$\begin{aligned}
&\|\mathbf{B}_{t+1} - \mathbf{\Sigma}\|_{\mathsf{F}} \\
&\leq (1-\eta)\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}} + \eta^2\|\mathbf{\Sigma}\|_{\mathsf{F}} + \eta^2 \frac{\sigma_1(\mathbf{\Sigma})\|\mathbf{\Sigma}\|_{\mathsf{F}}}{\sigma_r(\mathbf{B}_t)} \\
&\leq (1-\eta)\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}} + \eta^2\|\mathbf{\Sigma}\|_{\mathsf{F}} + 3\eta^2 \frac{\sigma_1(\mathbf{\Sigma})\|\mathbf{\Sigma}\|_{\mathsf{F}}}{\sigma_r(\mathbf{\Sigma})} \\
&\overset{(c)}{\leq} \eta\|\mathbf{\Sigma}\|_{\mathsf{F}} + 3\eta\kappa\|\mathbf{\Sigma}\|_{\mathsf{F}} + (1-\eta)^{t+1-T_1}\|\mathbf{B}_{T_1} - \mathbf{\Sigma}\|_{\mathsf{F}} \\
&= \eta\|\mathbf{A}\|_{\mathsf{F}} + 3\eta\kappa\|\mathbf{A}\|_{\mathsf{F}} + (1-\eta)^{t+1-T_1}\|\mathbf{B}_{T_1} - \mathbf{\Sigma}\|_{\mathsf{F}}
\end{aligned}$$

where (c) is by Lemma 16. From this inequality it is not difficult to see that once $\eta = \mathcal{O}(\frac{1}{\kappa^3\|\mathbf{A}\|_{\mathsf{F}}})$, one will have $\|\mathbf{B}_{t+1} - \mathbf{\Sigma}\|_{\mathsf{F}} \leq \mathcal{O}(1/\kappa^2)$ within the stated iterations. ■

Lemma 8 (Phase II. Quadratic convergence to global optima.) *If we choose $\eta = 0.5$, and suppose that after T_2 iterations, $\sigma_r(\mathbf{B}_{T_2}) \geq \sigma_r(\mathbf{\Sigma})/3$ and $\|\mathbf{B}_{T_2} - \mathbf{\Sigma}\|_{\mathsf{F}} \leq 2/(3\kappa^2)$ are satisfied. ScaledGD then ensures that for any $t \geq T_2$*

$$\|\mathbf{X}_{t+1}\mathbf{X}_{t+1}^\top - \mathbf{A}\|_{\mathsf{F}} = \|\mathbf{B}_{t+1} - \mathbf{\Sigma}_r\|_{\mathsf{F}} \leq \frac{4}{3\kappa^2}\frac{1}{2^{2^{t+1}}}.$$

Proof Let $\mathbf{C}_t = \mathbf{\Sigma}^{-1}\mathbf{B}_t$. We can rewrite (11) as

$$\mathbf{C}_{t+1} = (1-\eta)^2\mathbf{C}_t + 2\eta(1-\eta)\mathbf{I}_r + \eta^2\mathbf{C}_t^{-1}.$$

Subtracting $\mathbf{I}_r$ and rearranging it, we arrive at

$$\mathbf{C}_{t+1} - \mathbf{I}_r = (1-2\eta)(\mathbf{C}_t - \mathbf{I}_r) + \eta^2\mathbf{C}_t^{-1}(\mathbf{C}_t - \mathbf{I}_r)^2.$$

By choosing $\eta = 0.5$, we have that

$$\mathbf{C}_{t+1} - \mathbf{I}_r = \frac{1}{4}\mathbf{C}_t^{-1}(\mathbf{C}_t - \mathbf{I}_r)^2.$$

Multiplying both sides with $\mathbf{\Sigma}$, we have that

$$\begin{aligned}
\mathbf{B}_{t+1} - \mathbf{\Sigma} &= \frac{1}{4}\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma}(\mathbf{C}_t - \mathbf{I}_r)(\mathbf{C}_t - \mathbf{I}_r) \\
&= \frac{1}{4}\mathbf{\Sigma}\mathbf{B}_t^{-1}(\mathbf{B}_t - \mathbf{\Sigma})\mathbf{\Sigma}^{-1}(\mathbf{B}_t - \mathbf{\Sigma}).
\end{aligned}$$

This implies that

$$\begin{aligned}
\|\mathbf{B}_{t+1} - \mathbf{\Sigma}\|_{\mathsf{F}} &\leq \frac{1}{4}\|\mathbf{\Sigma}\|_2\|\mathbf{B}_t^{-1}\|_2\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}}\|\mathbf{\Sigma}^{-1}\|_2\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}} \\
&\overset{(a)}{\leq} \frac{3}{4}\frac{\sigma_1(\mathbf{\Sigma})}{\sigma_r^2(\mathbf{\Sigma})}\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}}^2 \overset{(b)}{=} \frac{3\kappa^2}{4}\|\mathbf{B}_t - \mathbf{\Sigma}\|_{\mathsf{F}}^2
\end{aligned}$$

where (a) is by Lemma 2, i.e., once $\sigma_r(\mathbf{B}_{T_2}) \geq \sigma_r(\mathbf{\Sigma})/3$, $\sigma_r(\mathbf{B}_t) \geq \sigma_r(\mathbf{\Sigma})/3$ holds for all $t \geq T_2$; and (b) is by $\sigma_1(\mathbf{\Sigma}) = 1$ and $\sigma_r(\mathbf{\Sigma}) = 1/\kappa$.

Finally, applying Lemma 18, it can be seen that long as $\|\mathbf{B}_{T_2} - \mathbf{\Sigma}\|_{\mathsf{F}} \leq \frac{2}{3\kappa^2}$, a quadratic rate can be established. And this condition is satisfied from Lemma 7. ■

7.3. Missing proofs for the symmetric and UP setting

We start with some notation that would be helpful for this subsection. Let the thin eigenvalue decomposition of $\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{m\times r_A}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$.

In Lemma 10 we will prove that we can always write $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ if we employ Nyström initialization and ScaledGD in (2), where $\mathbf{\Phi}_t \in \mathbb{R}^{r_A\times r}$. We also denote $\mathbf{\Theta}_t := \mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}$, where the invertibility of $(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)$ will become clear in the proof.

Lastly, let $\mathbf{B}_t := \mathbf{\Phi}_t^\top\mathbf{\Sigma}^{-1}\mathbf{\Phi}_t$. Note that $\mathbf{B}_t \in \mathbb{R}^{r\times r}$ and $\mathbf{B}_t = \mathbf{X}_t^\top\mathbf{A}^\dagger\mathbf{X}_t$.

7.3.1. HOW GOOD IS WEAK OPTIMALITY

Lemma 9 *All global optimal solutions to (1a) are also weakly optimal.*

Proof We start with rewriting $\mathbf{A}$,

$$\mathbf{A} = [\mathbf{Q}_1, \mathbf{Q}_2]\begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_2\end{bmatrix}\begin{bmatrix}\mathbf{Q}_1^\top\\ \mathbf{Q}_2^\top\end{bmatrix} = \mathbf{Q}_1\mathbf{\Sigma}_1\mathbf{Q}_1^\top + \mathbf{Q}_2\mathbf{\Sigma}_2\mathbf{Q}_2^\top \tag{13}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{m\times r}$ and $\mathbf{Q}_2 \in \mathbb{R}^{m\times(r_A-r)}$ are the first r and other columns of $\mathbf{Q}$, respectively; and $\mathbf{\Sigma}_1 \in \mathbb{R}^{r\times r}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{(r-r_A)\times(r-r_A)}$ are diagonal matrices formed by the first r and the rest diagonal entries of $\mathbf{\Sigma}$.

It is not difficult to see that the optimal solution of (1a) is $\mathbf{X}_* = \mathbf{Q}_1\mathbf{\Sigma}_1^{1/2}\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{r\times r}$ is any unitary matrix that accounts for rotation. Note that the pseudo-inverse of $\mathbf{A}$ can be written as $\mathbf{A}^\dagger = \mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{Q}^\top$. Plugging $\mathbf{X}_*$ into the definition of weak optimality, we arrive at

$$\mathbf{X}_*^\top\mathbf{A}^\dagger\mathbf{X}_* = \mathbf{U}\Sigma_1^{1/2}\mathbf{Q}_1^\top(\mathbf{Q}_1\mathbf{\Sigma}_1^{-1}\mathbf{Q}_1^\top + \mathbf{Q}_2\mathbf{\Sigma}_2^{-1}\mathbf{Q}_2^\top)\mathbf{Q}_1\mathbf{\Sigma}_1^{1/2}\mathbf{U}^\top \overset{(a)}{=} \mathbf{I}_r$$

where in (a) we use the facts $\mathbf{Q}_1^\top\mathbf{Q}_1 = \mathbf{I}_r$ and $\mathbf{Q}_1^\top\mathbf{Q}_2 = \mathbf{0}_{r\times(r_A-r)}$. This concludes the proof. ■

7.3.2. ELIMINATING RESIDUAL SPACE

Lemma 10 *If the update (2) is equipped with Nyström initialization (3), one can write $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t, \forall t$ for some $\mathbf{\Phi}_t \in \mathbb{R}^{r_A\times r}$.*

Proof The proof is based on induction. First we have that $\mathbf{X}_0 = \mathbf{A}\mathbf{\Omega} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{\Omega}$. It is clear that $\mathbf{\Phi}_0 = \mathbf{\Sigma}\mathbf{Q}^\top\mathbf{\Omega}$. Now suppose that one can write $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$, following the update (2), it is not hard to see that

$$\begin{aligned}\mathbf{\Phi}_{t+1} &= \mathbf{\Phi}_t - \eta\big(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma}\big)\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}\\ &= (1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\underbrace{\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}}_{:=\mathbf{\Theta}_t}.\end{aligned} \tag{14}$$

The variable $\mathbf{\Theta}_t \in \mathbb{R}^{r_A\times r}$ can be roughly viewed as a pseudo-inverse of $\mathbf{\Phi}_t^\top$ because $\mathbf{\Phi}_t^\top\mathbf{\Theta}_t = \mathbf{I}_r$. We note that the invertibility of $(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)$ will become clear in Lemma 12. ■

7.3.3. PROOF OF THEOREM 4

Proof Using $\mathbf{\Phi}_t^\top \mathbf{\Theta}_t = \mathbf{I}_r$, definition of $\mathbf{B}_t$ (at the start of Apdx. 7.3), and the update of $\mathbf{\Phi}_{t+1}$ in (14), it is not difficult to verify that

$$\mathbf{B}_{t+1} = (1-\eta)^2 \mathbf{B}_t + 2\eta(1-\eta)\mathbf{I}_r + \eta^2 \mathbf{\Theta}_t^\top \mathbf{\Sigma} \mathbf{\Theta}_t. \tag{15}$$

Subtracting $\mathbf{I}_r$ on both sides of (15), we can get

$$\mathbf{B}_{t+1} - \mathbf{I}_r = (1-\eta)^2 (\mathbf{B}_t - \mathbf{I}_r) - \eta^2 \mathbf{I}_r + \eta^2 \mathbf{\Theta}_t^\top \mathbf{\Sigma} \mathbf{\Theta}_t.$$

This ensures that

$$\begin{aligned}
&\|\mathbf{B}_{t+1} - \mathbf{I}_r\|_{\mathsf{F}} \\
&\leq (1-\eta)^2 \|\mathbf{B}_t - \mathbf{I}_r\|_{\mathsf{F}} + \eta^2 \sqrt{r} + \eta^2 \|\mathbf{\Theta}_t^\top \mathbf{\Sigma} \mathbf{\Theta}_t\|_{\mathsf{F}} \\
&\leq (1-\eta)^2 \|\mathbf{B}_t - \mathbf{I}_r\|_{\mathsf{F}} + \eta^2 \sqrt{r} + \eta^2 \frac{r}{\sigma_r(\mathbf{B}_t)}
\end{aligned}$$

where the last inequality is because of Lemma 13. Suppose that $\eta \leq 2/3$, from Lemma 12, one can see that there exists a time T_1 such that $\sigma_r(\mathbf{B}_t) \geq 1/3$. We assume $T_1 = 0$ following the same argument (i.e., initialized large with large ξ) as previous proofs. With these arguments, we obtain that

$$\begin{aligned}
&\|\mathbf{B}_{t+1} - \mathbf{I}_r\|_{\mathsf{F}} \\
&\leq (1-\eta) \|\mathbf{B}_t - \mathbf{I}_r\|_{\mathsf{F}} + \eta^2 \sqrt{r} + 3r\eta^2 \\
&\leq \eta\sqrt{r} + 3\eta r + (1-\eta)^{t+1-T_1} \|\mathbf{B}_{T_1} - \mathbf{I}_r\|_{\mathsf{F}} \\
&\leq \eta\sqrt{r} + 3\eta r + (1-\eta)^{t+1-T_1} \|\mathbf{B}_{T_1} - \mathbf{I}_r\|_{\mathsf{F}}.
\end{aligned} \tag{16}$$

This implies a linear rate, i.e, $\|\mathbf{B}_{t+1} - \mathbf{I}_r\|_{\mathsf{F}} \leq \mathcal{O}(\eta r) + \epsilon$ if $\eta = \mathcal{O}(1)$ with sufficient iterations.

Inequality (16) also implies that choosing $\eta = \mathcal{O}(\epsilon/r)$, $\|\mathbf{B}_{t+1} - \mathbf{I}_r\|_{\mathsf{F}} \to 0$ at a rate of $\mathcal{O}(\frac{r}{\epsilon} \log \frac{1}{\epsilon})$. The proof is thus completed. ∎

7.3.4. GLOBAL BEHAVIOR OF SCALEDGD UNDER NYSTRÖM INITIALIZATION

Lemma 11 *Let $\mathbf{Q}_1$ be the first r column on $\mathbf{Q}$, and $\mathbf{\Sigma}_1$ be the top-left $r \times r$ sub-block of $\mathbf{\Sigma}$. Denote an optimal solution to (1a) as $\mathbf{X}_* = \mathbf{Q}_1 \mathbf{\Sigma}_1^{1/2}$. ScaledGD ensures that*

$$\lim_{t \to \infty} \|\mathbf{X}_t - \mathbf{X}_*\|_{\mathsf{F}} \leq \mathcal{O}(r^{3/4}).$$

Proof We start with notation. Let

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{bmatrix}, \qquad \mathbf{\Phi}_t = \begin{bmatrix} \mathbf{M}_t \\ \mathbf{N}_t \end{bmatrix}, \tag{17}$$

where $\mathbf{\Sigma}_1 \in \mathbb{R}^{r \times r}$ is the learnable eigenvalues, while $\mathbf{\Sigma}_2 \in \mathbb{R}^{(r_A - r) \times (r_A - r)}$ are the unlearnable eigenvalues, and $\mathbf{M}_t \in \mathbb{R}^{r \times r}$ and $\mathbf{N}_t \in \mathbb{R}^{(r_A - r) \times r}$. Ideally at global convergence, we hope that $\mathbf{M}_t \to \mathbf{\Sigma}_1^{1/2}$ up to rotation; while $\mathbf{N}_t \to \mathbf{0}$.

We consider a scenario with $t \to \infty$, i.e., $\mathbf{B}_t = \mathbf{I}_r$. Using (17) to rewrite $\mathbf{B}_t = \mathbf{I}_r$, we have that

$$\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1} \mathbf{M}_t + \mathbf{N}_t^\top \mathbf{\Sigma}_2^{-1} \mathbf{N}_t = \mathbf{I}_r. \tag{18}$$

The above equation implies that

$$\begin{aligned} \text{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1} \mathbf{M}_t) &= \text{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2} \mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t) \\ &= \|\mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t\|_{\mathsf{F}}^2 \overset{(a)}{\leq} r \end{aligned} \tag{19}$$

where (a) is by (18) and Lemma 19.

Since we hope $\mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t \to \mathbf{I}_r$, we have that

$$\begin{aligned} & \|\mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t - \mathbf{I}_r\|_{\mathsf{F}}^2 \\ &= \text{Tr}\Big((\mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t - \mathbf{I}_r)^\top (\mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t - \mathbf{I}_r) \Big) \\ &= \text{Tr}\big(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2} \mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t\big) + \text{Tr}(\mathbf{I}_r) - 2\text{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}) \\ &\overset{(a)}{\leq} \text{Tr}\big(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2} \mathbf{\Sigma}_1^{-1/2} \mathbf{M}_t\big) + \text{Tr}(\mathbf{I}_r) + 2r^{3/2} \\ &\overset{(b)}{\leq} 2r + 2r^{3/2} \end{aligned} \tag{20}$$

where (a) is because that i) for any $r \times r$ matrix $\mathbf{C}$, we have that $\text{Tr}(\mathbf{C}) \geq r \min_i \mathbf{C}_{ii} \geq -r\|\mathbf{C}\|_{\mathsf{F}}$, ii) take $\mathbf{C} = \mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}$ and then apply (19); and (b) is by (19).

To bound $\mathbf{N}_t$, it can be seen that

$$\frac{1}{\sigma_{r+1}(\mathbf{A})} \text{Tr}\big(\mathbf{N}_t^\top \mathbf{N}_t\big) \leq \text{Tr}\big(\mathbf{N}_t^\top \mathbf{\Sigma}_2^{-1} \mathbf{N}_t\big) \overset{(c)}{\leq} r \tag{21}$$

where (c) is by applying Lemma 19 on (18). This suggests that $\|\mathbf{N}_t\|_{\mathsf{F}} \leq \sqrt{r\sigma_{r+1}(\mathbf{A})}$.

Lastly, note that $\mathbf{X}_*$ can be written as $\mathbf{X}_* = \mathbf{Q}[\mathbf{\Sigma}_1, \mathbf{0}]^\top$ and $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$. Using this fact and combining (20) and (21), we have that

$$\begin{aligned} \|\mathbf{X}_t - \mathbf{X}_*\|_{\mathsf{F}}^2 &= \|\mathbf{M}_t - \mathbf{\Sigma}^{1/2}\|_{\mathsf{F}}^2 + \|\mathbf{N}_t\|_{\mathsf{F}}^2 \\ &= \|\mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{-1/2}\mathbf{M}_t - \mathbf{I}_r)\|_{\mathsf{F}}^2 + \|\mathbf{N}_t\|_{\mathsf{F}}^2 \\ &\leq \sigma_1(\mathbf{\Sigma}^{1/2})^2 \|\mathbf{\Sigma}^{-1/2}\mathbf{M}_t - \mathbf{I}_r\|_{\mathsf{F}}^2 + \|\mathbf{N}_t\|_{\mathsf{F}}^2 \\ &= \mathcal{O}(r^3/2) \end{aligned} \tag{22}$$

where we used $\sigma_1(\mathbf{\Sigma}) = 1$ and $\sigma_{r+1}(\mathbf{\Sigma}) \leq 1$. The proof is thus completed. ■

7.3.5. USEFUL LEMMAS FOR SYMMETRIC UP PROBLEMS

It is clear that $\mathbf{B}_t$ is symmetric by definition, i.e., $\mathbf{B}_t = \mathbf{\Phi}_t^\top \mathbf{\Sigma}^{-1} \mathbf{\Phi}_t$. This enables us to give a lower bound on $\sigma_r(\mathbf{B}_t)$ using Lemma 17.

Lemma 12 *$\sigma_r(\mathbf{B}_t)$ is lower bounded by*

$$\sigma_r(\mathbf{B}_{t+1}) \geq (1-\eta) - (1-\eta)^{2t+3} + (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0).$$

Proof Given the definition of $\mathbf{B}_t$, it is not difficult to see that $\mathbf{B}_t$ is PSD for all t. We can then apply Lemma 17 on (15) to arrive at

$$\begin{aligned}
&\sigma_r(\mathbf{B}_{t+1}) \\
&\geq 2\eta(1-\eta) + \sigma_r\big((1-\eta)^2\mathbf{B}_t + \eta^2\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t\big) \\
&\geq 2\eta(1-\eta) + (1-\eta)^2\sigma_r\big(\mathbf{B}_t\big) \\
&\overset{(a)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + 2\eta(1-\eta)\frac{1-(1-\eta)^{2t+2}}{2\eta-\eta^2} \\
&\overset{(b)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + (1-\eta) - (1-\eta)^{2t+3}
\end{aligned}$$

where (a) uses Lemma 16 to unroll $\sigma_r(\mathbf{B}_t)$; and (b) is because $\frac{2\eta}{2\eta-\eta^2} \geq 1$. ■

Lemma 13 *Let $\mathbf{\Theta}_t$ and $\mathbf{B}_t$ defined the same as those in Section 2.3. It is guaranteed to have that*

$$\|\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t\|_\mathsf{F} \leq \frac{r}{\sigma_r(\mathbf{B}_t)}.$$

Proof Using the inequality $\|\mathbf{A}^\top\mathbf{A}\|_\mathsf{F} \leq \|\mathbf{A}\|_\mathsf{F}^2$, we have that

$$\|\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t\|_\mathsf{F} = \|\mathbf{\Theta}_t^\top\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Theta}_t\|_\mathsf{F} \leq \|\mathbf{\Sigma}^{1/2}\mathbf{\Theta}_t\|_\mathsf{F}^2. \tag{23}$$

Now let $\mathbf{E}_t := \mathbf{\Sigma}^{1/2}\mathbf{\Theta}_t$ and $\mathbf{F}_t := \mathbf{\Sigma}^{-1/2}\mathbf{\Phi}_t$. Since we have that $\mathbf{F}_t^\top\mathbf{E}_t = \mathbf{I}_r$, we have that

$$\|\mathbf{F}_t^\top\mathbf{E}_t\|_\mathsf{F} = \|\mathbf{I}_r\|_\mathsf{F} = \sqrt{r}.$$

Since we also have that

$$\sqrt{r} = \|\mathbf{F}_t^\top\mathbf{E}_t\|_\mathsf{F} \overset{(a)}{\geq} \sigma_r(\mathbf{F}_t)\|\mathbf{E}_t\|_\mathsf{F} \overset{(b)}{=} \sqrt{\sigma_r(\mathbf{B}_t)}\|\mathbf{E}_t\|_\mathsf{F} \tag{24}$$

where (a) does not hold true for general two matrices $\mathbf{E}_t$ and $\mathbf{F}_t$. Here it holds because $\mathbf{E}_t$ and $\mathbf{F}_t$ share the same column space and row space and both of them have rank r, which implies that $\langle Null(\mathbf{F}), [\mathbf{E}_t]_i\rangle = \mathbf{0}, \forall i$ ($[\mathbf{E}_t]_i$ is the ith column of $\mathbf{E}_t$). And (b) is because $\mathbf{F}_t^\top\mathbf{F}_t = \mathbf{B}_t$, which means that the singular values of $\mathbf{F}_t$ are just square root of eigenvalues of $\mathbf{B}_t$. This implies that $\|\mathbf{E}_t\|_\mathsf{F} \leq \sqrt{r}/\sqrt{\sigma_r(\mathbf{B}_t)}$. Combining this inequality with (23), we have that

$$\|\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t\|_\mathsf{F} \leq \|\mathbf{\Theta}_t^\top\mathbf{\Sigma}^{1/2}\|_\mathsf{F}^2 = \|\mathbf{E}_t\|_\mathsf{F}^2 \leq \frac{r}{\sigma_r(\mathbf{B}_t)}.$$

The proof is thus completed. ■

8. Missing proofs for asymmetric setting

8.1. Missing proofs for asymmetric and EP setting

8.1.1. Elimination of residual space

Lemma 14 *The modified ScaledGD update* (5) *under Nyström initialization* (4) *ensures that* $\mathbf{X}_t = \mathbf{U}\mathbf{\Phi}_t$ *and* $\mathbf{Y}_t = \mathbf{V}\mathbf{\Psi}_t$*,* $\forall t \geq 0$ *for some* $\mathbf{\Phi}_t \in \mathbb{R}^{r\times r}$ *and* $\mathbf{\Psi}_t \in \mathbb{R}^{r\times r}$*.*

Proof The proof is finished by induction. From our Nyström initialization, one has that $\mathbf{\Psi}_0 = \mathbf{0}$ and $\mathbf{\Phi}_0 = \mathbf{\Sigma}\mathbf{V}^\top\mathbf{\Omega}$. Now assume that one can write $\mathbf{X}_t = \mathbf{U}\mathbf{\Phi}_t$ and $\mathbf{Y}_t = \mathbf{V}\mathbf{\Psi}_t$ for some iteration t. We will show that $\mathbf{X}_{t+1} = \mathbf{U}\mathbf{\Phi}_{t+1}$ and $\mathbf{Y}_{t+1} = \mathbf{V}\mathbf{\Psi}_{t+1}$ under iteration (5). Let us start with $\mathbf{X}_{t+1}$. Note that if $t = 0$, $\mathbf{X}_1 = \mathbf{U}\mathbf{\Phi}_1$ is trivial. We only focus on $t \geq 1$, where we have

$$\begin{aligned}
\mathbf{X}_{t+1} &= \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t(\mathbf{Y}_t^\top\mathbf{Y}_t)^{-1} \\
&= \mathbf{U}\mathbf{\Phi}_t - \eta(\mathbf{U}\mathbf{\Phi}_t\mathbf{\Psi}_t^\top\mathbf{V}^\top - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)\mathbf{V}\mathbf{\Psi}_t(\mathbf{\Psi}_t^\top\mathbf{V}^\top\mathbf{V}\mathbf{\Psi}_t)^{-1} \\
&= \mathbf{U}\mathbf{\Phi}_t - \eta\mathbf{U}(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top - \mathbf{\Sigma})\mathbf{\Psi}_t(\mathbf{\Psi}_t^\top\mathbf{\Psi}_t)^{-1} \\
&= \mathbf{U}\underbrace{\Big(\mathbf{\Phi}_t - \eta(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top - \mathbf{\Sigma})\mathbf{\Psi}_t(\mathbf{\Psi}_t^\top\mathbf{\Psi}_t)^{-1}\Big)}_{:=\mathbf{\Phi}_{t+1}}.
\end{aligned}$$

Note that the invertible of $(\mathbf{\Psi}_t^\top\mathbf{\Psi}_t)$ will be proved through an asymmetric-to-symmetric reduction.

Step 1. Positive definiteness of $\mathbf{\Phi}_t\mathbf{\Psi}_t^\top$. We will first show that $\mathbf{\Phi}_t\mathbf{\Psi}_t^\top$ is symmetric and positive definite (PD) for any $t \geq 1$. From the proof of Theorem 5, it can be seen that $\mathbf{\Phi}_1\mathbf{\Psi}_1^\top = \eta\mathbf{\Sigma}$ is symmetric and PD. This means that the base case of induction holds. Now suppose that $\mathbf{\Phi}_t\mathbf{\Psi}_t^\top$ is symmetric and PD at iteration t. Based on the updates of ScaledGD, we can write the iteration as

$$\mathbf{\Phi}_{t+1} = (1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\mathbf{\Psi}_t^{-\top} \tag{25a}$$

$$\mathbf{\Psi}_{t+1} = (1-\eta)\mathbf{\Psi}_t + \eta\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}. \tag{25b}$$

This gives that

$$\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top = (1-\eta)^2\mathbf{\Phi}_t\mathbf{\Psi}_t^\top + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top)^{-1}\mathbf{\Sigma}. \tag{26}$$

The symmetry of $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$ directly follows from (26). For the positive definiteness of $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$, we can apply Lemma 17 to get

$$\lambda_{\min}(\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top) \geq (1-\eta)^2\lambda_{\min}(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top) + 2\eta(1-\eta)\lambda_{\min}(\mathbf{\Sigma}) + \eta^2\lambda_{\min}(\mathbf{\Sigma}(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top)^{-1}\mathbf{\Sigma}) > 0.$$

This concludes the PD of $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$.

Step 2. Define $\mathbf{B}_t := \mathbf{\Phi}_t\mathbf{\Psi}_t^\top$, then (26) can be rewritten as

$$\mathbf{B}_{t+1} = (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma} \tag{27}$$

which is exactly same iteration as (11) for the symmetric EP case. Based on the results from Step 1, that is, $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$ is symmetric and PD, we can apply the same analysis steps for symmetric EP problems to show that $\mathbf{B}_t$ is rank r We do not repeat for conciseness. This implies the full rankness of $\mathbf{\Phi}_t$ and $\mathbf{\Psi}_t$. ■

8.1.2. PROOF OF THEOREM 5

Proof Based on the initialization (4) and iteration (5), we can obtain that

$$\mathbf{\Phi}_1 = \mathbf{\Phi}_0 \tag{28a}$$

$$\begin{aligned}\mathbf{\Psi}_1 &= \mathbf{V}^\top \mathbf{Y}_1 = \mathbf{0} - \eta \mathbf{V}^\top (\mathbf{0} - \mathbf{A})^\top \mathbf{U}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0)^{-1} \\ &= \eta \mathbf{V}^\top \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0)^{-1} \\ &= \eta \mathbf{\Sigma}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{\Phi}_0)^{-1} \\ &= \eta \mathbf{\Sigma}\mathbf{\Phi}_0^{-\top}.\end{aligned} \tag{28b}$$

This ensures that

$$\mathbf{\Phi}_1 \mathbf{\Psi}_1^\top = \eta \mathbf{\Sigma}.$$

Choosing $\eta = 1$ completes the proof. ■

8.2. Missing proofs for asymmetric and UP setting

8.2.1. HOW GOOD IS WEAK OPTIMALITY?

Lemma 15 *Every global optimum for* (1b) *is also weakly optimal.*

Proof We start with rewriting the SVD of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ as

$$\mathbf{A} = [\mathbf{U}_1, \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^\top \\ \mathbf{V}_2^\top \end{bmatrix} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^\top + \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2^\top \tag{29}$$

where $\mathbf{U}_1 \in \mathbb{R}^{m \times r}$ and $\mathbf{U}_2 \in \mathbb{R}^{m \times (r_A - r)}$ are the first r and other columns of $\mathbf{U}$, respectively; $\mathbf{\Sigma}_1 \in \mathbb{R}^{r \times r}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{(r - r_A) \times (r - r_A)}$ are diagonal matrices formed by the first r and rest diagonal entries of $\mathbf{\Sigma}$; and $\mathbf{V}_1 \in \mathbb{R}^{n \times r}$ and $\mathbf{V}_2 \in \mathbb{R}^{n \times (r_A - r)}$ are the first r and other columns of $\mathbf{V}$.

It is not hard to see that the optimal solutions of (1a) are $\mathbf{X}_* = \mathbf{U}_1 \mathbf{\Sigma}_1^{1/2} \mathbf{Q}$ and $\mathbf{Y}_* = \mathbf{V}_1 \mathbf{\Sigma}_1^{1/2} \mathbf{Q}^{-\top}$, where $\mathbf{Q} \in \mathbb{R}^{r \times r}$ is any invertible matrix. Using these notation, we have that

$$\begin{aligned}\mathbf{Y}_*^\top \mathbf{A}^\dagger \mathbf{X}_* &= \mathbf{Q}^{-1} \mathbf{\Sigma}_1^{1/2} \mathbf{V}_1^\top (\mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \mathbf{U}_1^\top + \mathbf{V}_2 \mathbf{\Sigma}_2^{-1} \mathbf{U}_2^\top) \mathbf{U}_1 \mathbf{\Sigma}_1^{1/2} \mathbf{Q} \\ &\overset{(a)}{=} \mathbf{I}_r\end{aligned}$$

where in (a) we use the facts $\mathbf{U}_1^\top \mathbf{U}_1 = \mathbf{I}_r$ and $\mathbf{U}_1^\top \mathbf{U}_2 = \mathbf{0}_{r \times (r_A - r)}$. This concludes the proof. ■

8.2.2. PROOF OF THEOREM 6

Proof The update in (5) ensures that

$$\mathbf{\Phi}_1 = \mathbf{\Phi}_0, \tag{30}$$

$$\begin{aligned}\mathbf{\Psi}_1 &= \mathbf{V}^\top \mathbf{Y}_1 = \mathbf{0} - \eta \mathbf{V}^\top (\mathbf{0} - \mathbf{A})^\top \mathbf{U}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0)^{-1} \\ &= \eta \mathbf{V}^\top \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_0)^{-1} \\ &= \eta \mathbf{\Sigma}\mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{\Phi}_0)^{-1} \\ &\overset{(a)}{:=} \eta \mathbf{\Sigma}\mathbf{\Theta}_0\end{aligned} \tag{31}$$

where in (a) we define $\mathbf{\Theta}_t := \mathbf{\Phi}_t (\mathbf{\Phi}_t^\top \mathbf{\Phi}_t)^{-1}$.

From the definition of generalized weak optimality, we can see that

$$\begin{aligned}\mathbf{Y}_1^\top \mathbf{A}^\dagger \mathbf{X}_1 &= \mathbf{\Psi}_1^\top \mathbf{V}^\top \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \mathbf{U}\mathbf{\Phi}_1 = \mathbf{\Psi}_1^\top \mathbf{\Sigma}^{-1}\mathbf{\Phi}_1 \\ &= \eta \mathbf{\Theta}_0^\top \mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Phi}_0 = \eta \mathbf{I}_r\end{aligned}$$

This means that when $\eta = 1$, UP achieves generalized weak optimality in one step. ■

9. Other useful lemmas

Lemma 16 *Let $A_{t+1} = (1-\theta)A_t + \beta$ with some $\alpha \in (0,1)$ and $\beta \geq 0$, then we have*

$$A_{t+1} = (1-\theta)^{t+1} A_0 + \beta \frac{1-(1-\theta)^{t+1}}{\theta} \leq (1-\theta)^{t+1} A_0 + \frac{\beta}{\theta}.$$

Proof The proof can be completed by simply unrolling A_{t+1} and using the fact $1+\alpha+\alpha^2+\ldots+\alpha^t \leq \frac{1}{1-\alpha}$. ■

Lemma 17 *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$ are positive semi-definite matrices, we have $\lambda_{\min}(\mathbf{A}+\mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B})$.*

Proof The smallest eigenvalue of $\mathbf{A}+\mathbf{B}$ can be expressed as

$$\lambda_{\min}(\mathbf{A}+\mathbf{B}) = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top(\mathbf{A}+\mathbf{B})\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \min_{\mathbf{x}_1\neq\mathbf{0}, \mathbf{x}_1=\mathbf{x}_2} \frac{\mathbf{x}_1^\top \mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^\top\mathbf{x}_1} + \frac{\mathbf{x}_2^\top \mathbf{B}\mathbf{x}_2}{\mathbf{x}_2^\top\mathbf{x}_2}. \tag{32}$$

On the other hand, we also have that

$$\lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) = \min_{\mathbf{x}_1\neq\mathbf{0}, \mathbf{x}_2\neq\mathbf{0}} \frac{\mathbf{x}_1^\top \mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^\top\mathbf{x}_1} + \frac{\mathbf{x}_2^\top \mathbf{B}\mathbf{x}_2}{\mathbf{x}_2^\top\mathbf{x}_2}. \tag{33}$$

Because (32) is a constrained version of the minimization problem (33), they share the same objective but (32) has shrinked feasible region. It is not difficult to see that $\lambda_{\min}(\mathbf{A}+\mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B})$. The proof is thus completed. ■

Lemma 18 *Consider a sequence $\{A_t\}_t$ with $A_t \geq 0, \forall t$. If there exists α such that $A_{t+1} \leq \alpha A_t^2$ and $A_0 \leq \frac{1}{2\alpha}$, A_t converges to 0 at a quadratic rate, i.e.,*

$$A_{t+1} \leq \frac{1}{\alpha} \frac{1}{2^{2^{t+1}}}.$$

Proof Unrolling A_{t+1}, we get that

$$A_{t+1} \leq \alpha A_t^2 \leq \alpha^3 A_{t-1}^4 \leq \alpha^7 A_{t-2}^8 \leq \frac{1}{\alpha}(\alpha A_0)^{2^{t+1}} \leq \frac{1}{\alpha}\frac{1}{2^{2^{t+1}}}.$$

The proof is thus completed. ■

Lemma 19 *For PSD matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} + \mathbf{B} = \mathbf{I}_r$, then we have Tr$(\mathbf{A}) \leq r$ and Tr$(\mathbf{B}) \leq r$.*

Proof The proof is straightforward and is omitted here. ■

Lemma 20 ([46]) *Let $\mathbf{\Omega}$ be an $d \times r$ matrix with $d \geq r$. The entries of $\mathbf{\Omega}$ are drawn independently from $\mathcal{N}(0,1)$. Then for every $\tau > 0$, we have that*

$$\mathbb{P}\big(\sigma_r(\mathbf{\Omega}) \leq \tau(\sqrt{d} - \sqrt{r-1})\big) \leq (C_1\tau)^{d-r+1} + e^{-C_2 d}.$$

where C_1 and C_2 are universal constants independent of d and r.

10. Missing experimental details

10.1. Details for problems with synthetic data

This subsection contains the detailed setup for the problems with synthetic data in Fig. 1. Recall that here we focus on symmetric problem for EP and UP.

For EP in Figs. 1 (a) and (b), we choose the PSD matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ in the following manner. We set $m = 1000$ and $r = r_A = 20$. The non-zero singular values are set as $\{1.0, 0.99, 0.98, \ldots, 0.82, 0.01\}$, where we intentionally set $\sigma_{r_A} = 0.01$ to enlarge the condition number. We choose the step size of GD as 0.01 to avoid divergence. The learning rate for ScaledGD is 0.5.

For UP in Fig. 1 (c), we choose PSD matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ in the following manner. We set $m = 1000$ and $r_A = 40$. The singular values of $\mathbf{A}$ are $\{1.0, 0.99, 0.98, \ldots, 0.45, 0.44, 0.05, 0.025, 0.01\}$. We choose $r = 20$ to ensure the UP nature of this problem.

10.2. Datasets

The evaluation of NoRA and NoRA+ is carried out on commonly adopted datasets in the literature. The experiments are conducted on NVIDIA H100 GPUs.

GLUE benchmark. GLUE is designed to provide general-purpose evaluation of language understanding [58]. Those adopted in our work include SST-2 (sentiment analysis, [50]), RTE[1] (inference). These datasets are released under different permissive licenses.

[1] https://paperswithcode.com/dataset/rte

SuperGLUE benchmark. SuperGLUE [57] is another commonly adopted benchmark for language understanding, and it is more challenging compared with GLUE. The considered datasets include CB (inference, [10]), ReCoRD (question answering [71]), WSC (coreference resolution, [32]), BoolQ (question answering, [9]), and MiltiRC (question answering, [28]). These datasets are released under different permissive licenses.

Commonsense reasoning. These datasets are a collection tasks that require commonsense reasoning to answer. The considered datasets include WinoGrande [47], PIQA [3], SOCIAL-I-QA (SIQA) [48], HellaSwag [67], ARC-easy, ARC-challenge [8] and OpenbookQA [42]. These datasets are released under different permissive licenses.

Additional datasets. We also use SQuAD (question answering [45]) in our experiments, which is released under license CC BY-SA 4.0.

10.3. Few-shot learning with OPT-1.3B

Our evaluation starts with a few-shot learning task following [40]. The objective is to rapidly adapt a language model with a small training set. The datasets for this experiment are drawn from GLUE and SuperGLUE benchmarks [57, 58]. Consistent with [40], we randomly sample 1,000 data points for training and another 1,000 for testing.

We embrace OPT-1.3B as our base model [72]. The rank of LoRA is set to 8. We compare the proposed NoRA and NoRA+ with LoRA, prefix tuning [34], OLoRA [5], and PiSSA [41]. Note that the latter two serve alternative methods for initializing LoRA. Adam is adopted as the optimizer.

For this experiment, we first search for the best hyperparameters, e.g., batchsizes, for LoRA. The same batchsize is applied for other tested algorithms as well, but we search additionally for the best learning rate. This ensures that different algorithms see the same amount of data, while still have their best performed learning rate. The hyperparameters adopted are searched over values in Table 3. Adam is adopted for optimization.

Table 3: Hyperparameters used for few-shot learning with OPT-1.3B.

Hyperparameters	Values
LoRA r	8
LoRA α	16
LoRA module	q_proj, v_proj
# epochs	5
batchsize	2, 4, 8
learning rate	1×10^{-5}, 5×10^{-5}, 1×10^{-4}
NoRA ξ	0.05, 0.1, 0.2

The performance of different algorithms is summarized in Tab. 4. It is evident that OLoRA, PiSSA, NoRA and NoRA+ all outperform LoRA because their initialization strategies have provided more favorable directions for optimization. Among these initialization approaches, NoRA and NoRA+ have the best average accuracy, with absolute improvement over LoRA by 1.8 and 1.9.

Table 4: Performance of NoRA and NoRA+ for few-shot learning with OPT-1.3B.

OPT-1.3B	SST-2	WSC	BoolQ	CB	RTE	ReCoRD	MultiRC	SQuAD	avg (↑)
Prefix	92.9$_{\pm 0.9}$	59.6$_{\pm 1.6}$	73.1$_{\pm 2.3}$	71.6$_{\pm 2.9}$	65.2$_{\pm 2.6}$	69.7$_{\pm 1.0}$	64.4$_{\pm 3.2}$	82.2$_{\pm 1.4}$	72.3
LoRA	93.1$_{\pm 0.2}$	59.1$_{\pm 2.0}$	70.6$_{\pm 5.2}$	72.6$_{\pm 3.7}$	69.1$_{\pm 4.7}$	70.8$_{\pm 1.0}$	68.0$_{\pm 1.4}$	81.9$_{\pm 1.8}$	73.2
OLoRA	92.7$_{\pm 0.5}$	60.0$_{\pm 2.3}$	70.9$_{\pm 3.1}$	80.3$_{\pm 2.7}$	69.7$_{\pm 1.0}$	71.3$_{\pm 1.2}$	66.7$_{\pm 0.9}$	80.0$_{\pm 1.4}$	74.0
PiSSA	92.7$_{\pm 0.6}$	60.6$_{\pm 3.7}$	70.4$_{\pm 0.7}$	78.0$_{\pm 7.2}$	70.4$_{\pm 2.8}$	70.9$_{\pm 1.2}$	67.9$_{\pm 2.1}$	82.1$_{\pm 0.4}$	74.1
NoRA	93.4$_{\pm 0.7}$	60.6$_{\pm 3.8}$	73.2$_{\pm 0.6}$	79.2$_{\pm 5.2}$	72.0$_{\pm 1.3}$	71.3$_{\pm 1.0}$	68.5$_{\pm 1.2}$	81.8$_{\pm 0.7}$	**75.0**
NoRA+	93.2$_{\pm 0.5}$	61.2$_{\pm 0.6}$	72.9$_{\pm 1.3}$	79.5$_{\pm 5.8}$	72.4$_{\pm 3.6}$	71.5$_{\pm 0.9}$	68.4$_{\pm 1.2}$	82.0$_{\pm 0.9}$	**75.1**

10.4. Commonsense reasoning with LLaMA2

The base model considered is LLaMA2-7B. The experimental setup and choices of hyperparameter follow [38]. Training data are merged from 8 datasets listed in Tab. 2. The test sets remain separate for individual evaluation. The hyperparameters are summarized in Table 5.

Table 5: Hyperparameters used for commonsense reasoning with LLaMA2-7B.

Hyper-parameters	Values
LoRA r (rank)	32
LoRA α	64
LoRA module	q_proj, k_proj, v_proj, up_proj, down_proj
epoch	3
learning rate	3×10^{-4}
batchsize	16
cutoff length	256
NoRA ξ	0.02, 0.05, 0.1