On the Crucial Role of Initialization for Matrix Factorization

Bingcong Li[1], Liang Zhang[1], Aryan Mokhtari[2], Niao He[1]
[1]ETH Zurich, [2]University of Texas at Austin
{bingcong.li,liang.zhang,niao.he}inf.ethz.ch, mokhtari@austin.utexas.edu

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. Furthermore, we extend this initialization to low-rank adapters (LoRA) commonly used for finetuning foundation models. Our approach, NoRA, i.e., LoRA with Nyström initialization, demonstrates superior performance in various downstream tasks in large language and diffusion models.

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 [1, 2], as long as a suitable (small) learning rate is chosen, most of optimization algorithms such as 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 a class of Burer-Monteiro factorization [3] – *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 i) symmetric problems, $\min_{\mathbf{X}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_{\mathsf{F}}^2$; and ii) asymmetric ones, $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|_{\mathsf{F}}^2$. While these classical problems can be handled via various approaches, they are notoriously challenging for optimization, since they are nonconvex, nonsmooth (albeit differentiable), non-coercive (for asymmetric problems), and do not satisfy Polyak-Lojasiewicz (PL) condition [4]. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ (or $\mathbf{A} \in \mathbb{R}^{m\times m}$) for asymmetric (symmetric) problems, $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$. Building on the relation of $\mathrm{rank}(\mathbf{A})$ and r, we can categorize matrix factorization into three setups: exact-parametrized ($\mathrm{rank}(\mathbf{A})=r$), over-parametrized ($\mathrm{rank}(\mathbf{A})<r$), and under-parametrized ($\mathrm{rank}(\mathbf{A})>r$).

The asymmetric problem ii) is thoroughly explored in the literature. For the exact- and over-parametrized cases, global convergence has been established for GD, Alternating GD (AltGD), and ScaledGD [5–9], where most of them admit a linear rate. Regarding under-parametrized settings, only asymptotic global convergence of GD is established in [5] to the best of our knowledge. Common to above algorithms is the small initialization with $\mathbf{X}_0 \sim \mathcal{N}(0, \zeta_x^2)$ and $\mathbf{Y}_0 \sim \mathcal{N}(0, \zeta_y^2)$ for some sufficiently small ζ_x^2 and ζ_y^2. However, such initialization results in unfavorable performance both theoretically and empirically, partly because of the need of escaping from a saddle point $(\mathbf{0}, \mathbf{0})$.

This work proposes *Nyström initialization* to effectively bypass the aforementioned saddle point. More importantly, it significantly enhances the global convergence rates when applied on top of ScaledGD. In the exact- and over-parametrized settings, Nyström initialization boosts ScaledGD to converge at a *quadratic* rate (i.e., $\mathcal{O}(\log\log(1/\epsilon))$) on symmetric problems and enables a *one-step* convergence for asymmetric problems. For the more challenging case with under-parametrization, we prove that with our Nyström initialization, ScaledGD converges at a linear rate to the neighbor of a global optimum on symmetric problems, and then exhibits a sublinear rate to a more fine-grained

Second Conference on Parsimony and Learning (CPAL 2025).

Table 1: Comparison of complexity for global optimality in (a)symmetric matrix factorization in various settings. Here, EP, OP, and UP are abbreviations for exact-, over- and under- parametrization. See more explanations in the footnote.

setting		alg.	ref.	init.	rate
Asymmetric	EP	GD	[6]	small	$\mathcal{O}\big(\kappa^3 \log(1/\epsilon)\big)$
		AltGD	[7]	special	$\mathcal{O}\big(\kappa^2 \log(1/\epsilon)\big)$
		ScaledGD	[9]	local	$\mathcal{O}(\log(1/\epsilon))$
		ScaledGD	**Theorem 3**	Nyström	$\mathcal{O}(1)$
	OP	AltGD	[7]	special	$\mathcal{O}\big(\kappa^2 \log(1/\epsilon)\big)$
		ScaledGD	**Theorem 6**	Nyström	$\mathcal{O}(1)$
	UP	GD	[5]	small	asymptotic
		ScaledGD	**Theorem 4**	Nyström	$\mathcal{O}(1)$
Symmetric	EP	GD*	[10]	small	$\mathcal{O}\big(\kappa^8 + \kappa^2 \log(1/\epsilon)\big)$
		ScaledGD	**Theorem 1**	Nyström	$\mathcal{O}\big(\kappa^3\sqrt{r} + \log\log(1/\epsilon)\big)$
	OP	GD*	[10]	small	$\mathcal{O}\big(\kappa^8 + \kappa^6 \log(\kappa/\epsilon)\big)$
		ScaledGD	**Theorem 5**	Nyström	$\mathcal{O}\big(\kappa^3\sqrt{r} + \log\log(1/\epsilon)\big)$
	UP	ScaledGD	**Theorem 2**	Nyström	$\mathcal{O}(r/\epsilon \cdot \log(1/\epsilon))$

neighboring area. Overall, Nyström initialization enables us to improve existing rates in exact-, over-, and under-parametrized settings; see more detailed comparisons in Tab. 1.[1]

Our results highlight that the convergence of ScaledGD is *critically determined by the initialization.* Taking symmetric and exact-parametrized problems as an example, our quadratic rate slows down to a linear one when adopting either small initialization or slightly perturbed Nyström initialization.

After demonstrating the theoretical merits of Nyström initialization, we further extend its applications to another scenario with Burer-Monteiro factorization, in the context of LoRA for finetuning deep neural networks [11]. This is motivated by the fact that asymmetric matrix factorization is equivalent to LoRA applied on linear models with whitened data [12, 13], and is in line with several recent works that take insights from matrix factorization to improve LoRA [14, 15]. Compared with existing strategies for initializing LoRA [16–18], our Nyström initialization for LoRA (abbreviated as NoRA) is more economical and aligns better with existing deployment pipelines. The effectiveness of NoRA is demonstrated on downstream tasks from various domains, through both diffusion and large language models (LLMs). In a nutshell, our contributions can be summarized as:

- **Faster rates.** Nyström initialization is provably beneficial to ScaledGD. For symmetric problems, it catalyzes not only the first *quadratic rate* in exact- and over- parameterized settings, but also a (sub)linear rate for under-parametrization where only asymptotic results were known. It also allows more remarkable improvement on asymmetric problems; see details in Tab. 1. Moreover, these improved rates are obtained through a unified analysis framework.
- **Critical role of initialization.** Our theoretical results convey an intriguing message for nonconvex (nonsmooth) optimization: the behaviors of the same algorithm, whether converging at a quadratic or linear rate, are critically determined by initialization.
- **Practical implications.** We further illustrate the power of Nyström initialization for finetuning diffusion and large language models (LLMs). The resultant approach, NoRA, effectively improves the performance of LoRA on several representative tasks.

[1] In Tab. 1, ϵ is the prescribed optimality error, and κ denotes the condition number of $\mathbf{A}$. The bound for UP depicts the complexity to near optima. The "special" initialization in AltGD is still a small initialization, but with more careful designs that will be clear in Sec. 3.1. Works marked with * are designed for another setting (hence the comparison may not be fair).

1.1. Related works

We only streamline results on the convergence of matrix factorization under the broad umbrella of quartic optimization, where the objective is to minimize a forth-order polynomial. Other closely related topics, such as LoRA variants, can be found in Apdx. A.2.

Quartic optimization. Matrix factorization problems considered in this work are classical examples of forth-order growth functions. It involves a complex landscape characterized by nonconvexity, nonsmoothness, and the absence of PL condition. Similar to other works listed in Tab. 1, the goal of this work is to unveil intriguing behaviors from an optimization perspective. Recent works have examined the convergence of several algorithms, such as GD, AltGD, and ScaledGD [5–8, 19] in exact- and over- parametrized settings. Most of them admit linear convergence with different dependences on the condition number of the factorized matrix $\mathbf{A}$. GD for matrix square root problems is studied in [20]. Another closely related setting is matrix sensing; see e.g., [10, 21–24]. Linear rates are obtained for problems with exact- and over- parametrization, despite some of them demand early stopping. Similar to matrix factorization, not too much is known for under-parametrization. There are other approaches to tackle general forth-order growth optimization. For example, relative smoothness is considered in [25]; adaptive step sizes induced by fine-grained geometry are studied in [26]. The work of [27] also copes with such problems but requires convexity of the objective.

2. The power of initialization for symmetric matrix factorization

We start to examine the critical role of initialization on symmetric matrix factorization problems

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times r}} \frac{1}{4}\|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_{\mathsf{F}}^2. \tag{1}$$

Within this section, we assume that $\mathbf{A} \in \mathbb{R}^{m\times m}$ is positive semidefinite (PSD), otherwise one can employ the asymmetric formulation as in later sections. Problem (1) also closely links with matrix sensing, particularly under a sufficient number of Gaussian measurements [23]. From an optimization perspective, problem (1) is nonconvex and has no global Lipschitz gradient [4, 28].

Notationally, let $r_A := \text{rank}(\mathbf{A})$ and further denote the compact eigendecomposition as $\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{m\times r_A}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$. Since PSD matrices share the same eigen and singular values, we employ $\sigma_i(\cdot)$ to denote both in this section. Without loss of generality, we assume that the largest and smallest singular values are $\sigma_1(\mathbf{A}) = 1$ and $\sigma_{r_A}(\mathbf{A}) = 1/\kappa$ such that the condition number is κ.

ScaledGD as our optimizer. We investigate the power of initialization on ScaledGD [9], a preconditioned version of GD; see detailed discussions in e.g., [8, 9]. Starting from $t = 0$ with a learning rate $\eta > 0$, the update of ScaledGD is given by

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

The inversion of the $r \times r$ matrix $\mathbf{X}_t^\top\mathbf{X}_t$ is computationally feasible in the low-rank setting with $r \ll m$. Small initialization is widely adopted, i.e., $[\mathbf{X}_0]_{ij} \sim \mathcal{N}(0, \zeta^2)$, where ζ is a sufficiently small positive number. Under such initialization, ScaledGD converges linearly for exact-parametrization ($r = r_A$), yet less is known for under- and over-parametrization; see more in Tab. 1. Next, we show that a simple yet effective initialization can provoke faster convergence of ScaledGD.

2.1. Nyström initialization

To improve the convergence rates, it is essential to ensure that the initialization satisfies two conditions for exact- and under-parametrized problems[2]: 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., $\text{rank}(\mathbf{X}_0) = r$. The analytical rationale will be elucidated in the subsequent sections. A straightforward means to meet these conditions is via Nyström sketch [29]

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

[2]For the ease of presentation, the over-parametrized setting is considered in the appendix.

where $\mathbf{\Omega} \in \mathbb{R}^{m\times r}$ is a Gaussian random matrix. From this initialization, it is not difficult to see that condition i) is satisfied already. Our next lemma shows that the condition ii) holds w.h.p.

Lemma 1 (Initialization for exact- and under- parametrization)**.** *For some universal constant $\tau > 0$, $\sigma_r(\mathbf{X}_0) \geq \xi\tau(\sqrt{r_A} - \sqrt{r-1})\sigma_{r_A}(\mathbf{A})$ is satisfied with high probability, i.e., rank$(\mathbf{X}_0) = r$ w.h.p.*

The detailed expression for this "high probability" in Lemma 1 can be found in Apdx. B.1.1. Note that there is a "w.h.p." over initialization in most of our results. This refers to that rank$(\mathbf{X}_0) = r$ is needed for exact- and under-parametrized settings, and rank$(\mathbf{X}_0) = r_A$ is needed when over-parametrized.

2.2. Nyström initialization in the exact-parametrized setting

We start with Nyström initialization for exact-parametrized problems, i.e., $r_A = r$. Our first result dives into the implicit regularization induced by the ScaledGD under the proposed initialization.

Lemma 2. *If $\mathbf{X}_0$ is obtained by Nyström initialization (3) and rank$(\mathbf{X}_0) = r$ is satisfied, ScaledGD in (2) ensures that for all $t \geq 0$*

- *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,*
- *the smallest eigenvalue of $\mathbf{X}_t\mathbf{X}_t^\top$ satisfies that*

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

Lemma 2 implies the full rankness of $\mathbf{X}_t$ over the trajectory, i.e., rank$(\mathbf{X}_t) =$ rank$(\mathbf{\Phi}_t) = r, \forall t$. This ensures an invertible preconditioner $\mathbf{X}_t^\top\mathbf{X}_t$. In other words, iteration (2) is well-defined. The most important implication of Lemma 2 is the alignment of $\mathbf{X}_t$ with the directions of eigenvectors of $\mathbf{A}$, that is, $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$. This can be equivalently understood as the elimination of the residual space, i.e., $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t = \mathbf{0}, \forall t$. While we will expand this discussion shortly, this alignment in directions enables us to establish a quadratic rate for ScaledGD.

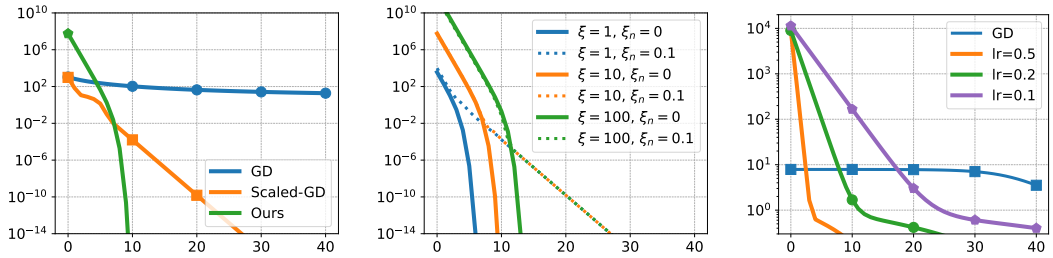
Theorem 1. *ScaledGD (2) with Nyström initialization (3) has a two-phase behavior w.h.p. over initialization.*

- *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 achieved after $T = \mathcal{O}\big(\log\log(\frac{1}{\kappa\epsilon})\big)$ iterations.*

Theorem 1 establishes that ScaledGD attains global optimality of (1) within $\mathcal{O}(\kappa^3\sqrt{r}\log\kappa + \log\log\frac{1}{\kappa\epsilon})$ iterations. ScaledGD linearly converges to a local region satisfying $\|\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A}\|_\mathsf{F} \leq \mathcal{O}(\frac{1}{\kappa^2})$, after which a quadratic rate can be granted. This is, to the best of our knowledge, the first quadratic rate for problem (1). Interestingly, it is achieved without requiring (exact) Hessian on a nonconvex and nonsmooth problem. A graphical illustration of this quadratic rate can be found in Fig. 1 (a) using data detailed in Apdx. E.1, where ScaledGD with Nyström initialization outperforms linearly converging algorithms such as GD and ScaledGD with small initialization. Moreover, it is worth emphasizing that Theorem 1 has no requirement on the magnitude of Nyström initialization – it does not need ξ in (3) to be small, avoiding escaping from the stationary point $\mathbf{0}$.

The critical role of initialization. As shown in Lemma 2, Nyström initialization aligns $\mathbf{X}_t$ to 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 [5, 6, 8], where small initialization only guarantees that $\|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t\|_\mathsf{F}$ converges to 0 at a linear rate. By getting rid of the residual space via Nyström initialization, ScaledGD can achieve a quadratic rate. We graphically illustrate this point in Fig. 1 (b), where we perturb Nyström initialization slightly to inject noise into the residual space. Reflected in the dotted lines, even if the noise is so small such that the earlier convergence does not differ from Nyström initialization, only a linear rate is observed for perturbed initialization.

Extensions to the case of over-parametrization. Nyström initialization is further extended to cope with over-parametrized case ($r > r_A$) in Apdx. B.4. We slightly modify ScaledGD by substituting the possibly non-invertible $(\mathbf{X}_t^\top\mathbf{X}_t)^{-1}$ in (2) with $(\mathbf{X}_t^\top\mathbf{X}_t)^\dagger$; see (26). Unlike previous works [21, 24],



(a) exact-parametrization (b) exact-parametrization (c) under-parametrization

Figure 1: Optimality error vs. iteration of ScaledGD under Nyström initialization in different settings. (a) Comparison of GD, and ScaledGD with small / Nyström initialization (ours). (b) Solid lines show that our initialization is not sensitive to magnitude of ξ; and dotted lines illustrate that quadratic convergence 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) Comparison of ScaledGD under Nyström initialization with various η.

our modification requires no damping parameters thanks to our Nyström initialization. This leads to, as far as we know, the first quadratic rate for over-parametrized problems. Numerical experiments on over-parametrized problems are provided in Fig. 4 in appendix to validate the quadratic rate.

2.3. Nyström initialization in the under-parametrized setting

Next, we consider the under-parametrized case of (1), i.e., $r < r_A$. To the best of our knowledge, only asymptotic convergence is established for GD on such problems [5]. This is partially because that even the local PL condition is challenging to be verified. With Nyström initialization, we will show that ScaledGD converges under a slightly weaker criterion.

Definition 1 (Weak optimality)**.** *Matrix $\mathbf{X} \in \mathbb{R}^{m\times r}$ is weakly optimal to (1) if $\mathbf{X}^\top \mathbf{A}^\dagger \mathbf{X} - \mathbf{I}_r = \mathbf{0}$.*

Our first result characterizes that all global optima are also weakly optimal. In other words, if weak optimality is ensured, this algorithm has a chance to reach a global optimum as well.

Lemma 3. *All globally optimal solutions to (1) are also weakly optimal.*

We then focus on the convergence of ScaledGD to weak optimality. In the case of under-parametrization, Nyström initialization also aligns $\mathbf{X}_t$ to the directions of eigenvectors of $\mathbf{A}$.

Lemma 4. *ScaledGD in (2) with Nyström initialization (3) ensures $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t, \forall t$ for some $\mathbf{\Phi}_t \in \mathbb{R}^{r_A \times r}$.*

Lemma 4 shows that $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{X}_t = \mathbf{0}, \forall t$ also holds, namely, Nyström initialization eliminates the residual space. Building upon this, the convergence of ScaledGD can be established.

Theorem 2. *The following holds w.h.p. for ScaledGD (2) with Nyström initialization (3):*

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

If one chooses a constant learning rate e.g, $\eta = 0.1$, linear convergence can be established until reaching a neighboring area of a weakly optimal solution. The error $\|\mathbf{X}_t^\top \mathbf{A}^\dagger \mathbf{X}_t - \mathbf{I}_r\|_\mathsf{F} = \mathcal{O}(\eta r)$ is low, given that r is typically small in practice. A graphical illustration of this linear rate can be found in Fig. 1 (c). On the other hand, if the learning rate is chosen according to the prescribed accuracy ϵ, one can obtain a sublinear rate $\mathcal{O}(\frac{r}{\epsilon}\log\frac{1}{\epsilon})$ to exact weak optimality. These behaviors clearly indicate a step scheduling of learning rates (e.g., setting $\eta = 0.1, 0.01, \ldots$ every a few iterations) for both fast convergence and exact weak optimality in practice. It is also worth mentioning that the convergence under both choices of η has no dependence on κ.

Finally, we show that even in the worst case, ScaledGD guarantees that $\mathbf{X}_t$ converges to a point that is adequately close to a global solution, and the relative distance is sublinear in r.

Lemma 5. *Let $\mathbf{Q}_1$ be the first r column on $\mathbf{Q}$, and $\mathbf{\Sigma}_1$ be the top-left $r \times r$ block of $\mathbf{\Sigma}$. Denote an optimal solution to (1) as $\mathbf{X}_* = \mathbf{Q}_1 \mathbf{\Sigma}_1^{1/2}$. W.h.p. over the initialization, ScaledGD (2) with Nyström initialization (3) ensures that $\lim_{t\to\infty} \|\mathbf{X}_t - \mathbf{X}_*\|_{\mathsf{F}} \leq \mathcal{O}(r^{3/4})$.*

3. The power of initialization for asymmetric matrix factorization

3.1. Initialization and modified ScaledGD

This section demonstrates that the power of initialization is even more striking in solving asymmetric matrix factorization than symmetric ones. Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, consider the following problem

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times r}, \mathbf{Y}\in\mathbb{R}^{n\times r}} \frac{1}{2}\|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|_{\mathsf{F}}^2. \tag{4}$$

Denote $\mathrm{rank}(\mathbf{A}) = r_A$, and the compact SVD as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times r_A}$, $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$, and $\mathbf{V} \in \mathbb{R}^{n\times r_A}$. Similar to the previous section, we assume that $\sigma_1(\mathbf{A}) = 1$ and $\sigma_{r_A}(\mathbf{A}) = 1/\kappa$.

Nyström initialization. We adopt an asymmetric manner to initialize $\mathbf{X}_0$ and $\mathbf{Y}_0$ for (4), i.e.,

$$\textbf{Nyström initialization:} \quad \mathbf{X}_0 = \mathbf{A}\mathbf{\Omega}, \quad \mathbf{Y}_0 = \mathbf{0} \tag{5}$$

where $\mathbf{\Omega}$ is a Gaussian random matrix of $\mathbb{R}^{n\times r}$ with $[\mathbf{\Omega}]_{ij} \sim \mathcal{N}(0, \xi^2), \forall i, \forall j$. We can follow the same steps of Lemma 1 to show that $\mathbf{X}_0$ in (5) is rank r w.h.p. in exact- and under-parametrized settings. Moreover, there is no requirement on the magnitude of ξ, meaning that it is possible to start far from the saddle point $(\mathbf{0}, \mathbf{0})$. This asymmetry of $\mathbf{X}_0$ and $\mathbf{Y}_0$ in (5) is in contrast with small initialization which typically induces $\|\mathbf{X}_0\|_{\mathsf{F}} \approx \|\mathbf{Y}_0\|_{\mathsf{F}}$ [5, 8]. The merits will become clear shortly. Note that AltGD [7] also adopts sketch at initialization, 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. Besides the requirement on small variance of $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ and the explicit need of $\sigma_1(\mathbf{A})$, this initialization cannot eliminate the residual space. Consequently, AltGD demands early stopping in exact- and over-parametrized problems.

Modified ScaledGD. To adapt to the non-invertible $\mathbf{Y}_0^\top\mathbf{Y}_0 = \mathbf{0}$ in Nyström initialization (5), we modify the first iteration of ScaledGD. More precisely, the updates are summarized below

$$\mathbf{X}_1 = \mathbf{X}_0, \text{ and } \mathbf{X}_{t+1} = \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t(\mathbf{Y}_t^\top\mathbf{Y}_t)^{-1}, \forall t \geq 1; \tag{6a}$$

$$\mathbf{Y}_{t+1} = \mathbf{Y}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})^\top\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t)^{-1}, \forall t \geq 0. \tag{6b}$$

3.2. Nyström initialization in the exact-parametrized setting

We start with the exact-parametrized case, i.e., $r_A = r$ in (4). The benefit of Nyström initialization (5) for iteration (6) is again the alignment of $\mathbf{X}_t$ and $\mathbf{Y}_t$ to the directions of singular vectors.

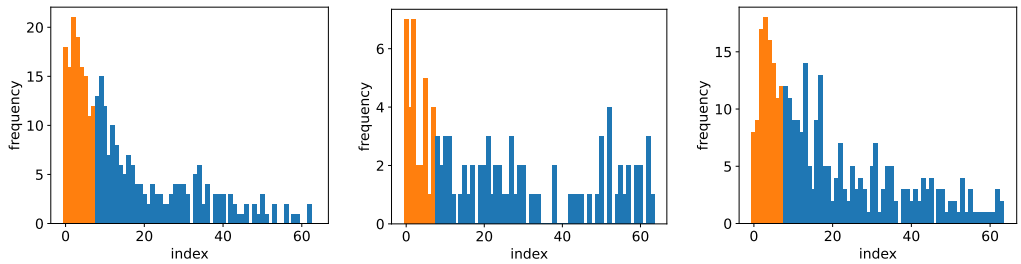
Lemma 6. *ScaledGD in (6) under Nyström initialization (5) guarantees that $\mathbf{X}_t = \mathbf{U}\mathbf{\Phi}_t$ and $\mathbf{Y}_t = \mathbf{V}\mathbf{\Psi}_t$, $\forall t \geq 0$ for some $\mathbf{\Phi}_t \in \mathbb{R}^{r\times r}$ and $\mathbf{\Psi}_t \in \mathbb{R}^{r\times r}$.*

Similar to the symmetric problems, the implication of Lemma 6 is the elimination of residual space, i.e., $(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\mathbf{X}_t = \mathbf{0}$ and $(\mathbf{I}-\mathbf{V}\mathbf{V}^\top)\mathbf{Y}_t = \mathbf{0}$. This turns out to be even more beneficial for asymmetric problems, as it induces one-step convergence of ScaledGD.

Theorem 3 (One-step convergence)**.** *With $\eta = 1$ and Nyström initialization (5), ScaledGD in (6) ensures $\mathbf{X}_1\mathbf{Y}_1^\top = \mathbf{A}$ w.h.p. over the initialization. In other words, global convergence is achieved in one step.*

Comparing to symmetric matrix factorization (cf. Theorem 1), Theorem 3 suggests that problem (4) requires less iterations to be solved owing to the asymmetry of $\mathbf{X}_0$ and $\mathbf{Y}_0$ at initialization (5). This partially agrees with results in [23], which illustrate the benefit of asymmetry in Burer-Monterio factorization for matrix sensing.

Lastly, we present a result that may be of independent interest – the asymmetric and symmetric problems are interconnected under our Nyström initialization. This link is made clear in the proof of the following corollary (to Theorem 1), which states that ScaledGD admits quadratic convergence under different choices of step sizes.



(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 others rarely have sufficiently large change.

Corollary 1 (Quadratic convergence)**.** *With Nyström initialization* (5) *and different choices of step sizes, ScaledGD in* (6) *has a similar behavior as Theorem 1, that is, a two-phase behavior w.h.p. over the initialization:*

- *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* $\|\mathbf{X}_{T_1}\mathbf{Y}_{T_1}^\top - \mathbf{A}\|_{\mathsf{F}} \le \mathcal{O}(1/\kappa^2)$*.*
- *Phase 2 (quadratic convergence). After Phase I, ScaledGD converges quadratically with* $\eta = 0.5$*. In particular,* $\|\mathbf{X}_T\mathbf{Y}_T^\top - \mathbf{A}\|_{\mathsf{F}} \le \epsilon$ *is ensured after* $T = \mathcal{O}\big(\log\log(\frac{1}{\kappa\epsilon})\big)$ *iterations.*

Extensions to over-parametrization. One-step global convergence can also be established for over-parametrized asymmetric problems under Nyström initialization. More on this can be found in Apdx. C.3, where we provide the first convergence result on ScaledGD under such a setup.

3.3. Nyström initialization in the under-parametrized setting

Lastly, we tackle the case of under-parametrization in the asymmetric problem (4), where $r_A > r$. Similar to the symmetric case in Sec.2.3, we consider a slightly weaker version of optimality.

Definition 2 (Generalized weak optimality)**.** *We say* $(\mathbf{X}, \mathbf{Y})$ *is weakly optimal if* $\mathbf{Y}^\top\mathbf{A}^\dagger\mathbf{X} - \mathbf{I}_r = \mathbf{0}$*.*

Generalized weak optimality is satisfied by any global optimum, which is proved in Lemma 13 in the appendix. With this preparation, we are ready to show that ScaledGD converges in a single step.

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

4. NoRA: Nyström low rank adapters

Our theoretical results highlight the merits of suitable initialization for matrix factorization problems. One of the key insights is that the Burer-Monterio factorization benefits from good directions of $\mathbf{X}_0$ and $\mathbf{Y}_0$ at initialization; cf. Lemmas 2 and 6. We term this as *directional alignment*. In this section, we extend the benefit of initialization to practical scenarios, showing that directional alignment is also beneficial for low-rank adapters (LoRA) in finetuning deep neural networks [11].

LoRA enhances parameter efficiency of finetuning by approximating the unknown parameter-change $\Delta\mathbf{W} \in \mathbb{R}^{m\times n}$ through Burer-Monterio 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, and $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{Y} \in \mathbb{R}^{n\times r}$ with $r \ll \min\{m, n\}$; see a more detailed recap in Apdx. A.2. 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 exists a set of well-performed adapters that lie in the column (row) span of the pretrained weight matrix [30], i.e., $\mathrm{ColSpan}(\Delta\mathbf{W}) \subseteq \mathrm{ColSpan}(\mathbf{W}_0)$ and $\mathrm{RowSpan}(\Delta\mathbf{W}) \subseteq \mathrm{RowSpan}(\mathbf{W}_0)$. In other words, $\mathbf{W}_0$ can be adopted as a suitable replacement of $\Delta\mathbf{W}$ for directional alignment.

Table 2: Training loss of NoRA and NoRA+ with stable-diffusion.

loss(↓)	LoRA	LoRA-P	NoRA	NoRA+
avg	0.092±0.012	0.093±0.012	0.084±0.017	0.084±0.015

Having $\text{ColSpan}(\mathbf{W}_0)$ alone is insufficient for directional alignment, since it does not specify which directions are more crucial. To answer this question, we examine the singular values that undergo the most significant change after LoRA finetuning on a few-shot learning task [31]. OPT-1.3B is chosen as the base model and LoRA is applied to its query and value matrices with $r = 8$; more details can be found in Apdx. E.3. For each LoRA layer, we count the indices of r singular values that exhibit the largest changes after finetuning, and summarize their frequencies across all layers in Fig. 2. It is observed that the top-r singular values tend to have larger change, explaining the success of LoRA initialization approaches that aligns $\mathbf{X}_0$ with the directions corresponding to these singular values, such as PiSSA and OLoRA [16, 17]. However, across all tested datasets, a substantial portion of non-top-r singular values also demonstrate significant variation, and the frequency is positively linked to the singular values. In other words, the directions corresponding to larger singular values tend to be more important. This is akin to the principle of Nyström initialization $\mathbf{X}_0 = \mathbf{W}_0\mathbf{\Omega}$, evidenced by its spectrum, i.e., $\mathbb{E}[\mathbf{X}_0\mathbf{X}_0^\top] \propto \mathbf{W}_0\mathbf{W}_0^\top$.

Building upon these observations, and considering the accelerated convergence with Nyström initialization in ScaledGD, we propose two novel variants of LoRA:

- **Nyström LoRA (NoRA)** applies (5) directly on top of LoRA, that is, $\mathbf{X}_0 = \mathbf{W}_0\mathbf{\Omega}$ and $\mathbf{Y}_0 = \mathbf{0}$.
- **Nyström preconditioned LoRA (NoRA+)** not only advances LoRA initialization with (5), but also leverages ScaledGD for optimization.

We note that ScaledGD has already been applied for LoRA training in [14], which we refer to as LoRA-P (P for preconditioning). We will show that both LoRA and LoRA-P benefit significantly from Nyström initialization. Due to space limitation, we summarize NoRA and NoRA+ in Algs. 1 and 2, respectively in the appendix, with additional explanations in Apdx. A.4.

Deployment efficiency. NoRA offers practical advantages over other initialization methods such as PiSSA and OLoRA. It not only bypasses the computationally expensive SVD or QR decomposition, but also avoids the need to modify to the pretrained weights. NoRA is thus an off-the-shelf solution to enhance LoRA without altering existing pipelines. We expand on this in Apdx. A.4.

5. Numerical results for NoRA

The efficiency of proposed NoRA and NoRA+ is demonstrated on large-scale finetuning tasks involving diffusion and LLMs. The experiments are conducted with PyTorch [32] on NVIDIA H100 GPUs. Details on datasets and experimental procedures can be found in Apdx. E.

5.1. Subject-driven image generation with stable-diffusion

We focus on subject-driven image generation [33] by finetuning a diffusion model with only a few user-specific images so that the modal can generate the same object in various contexts. The base model is selected as StableDiffusion v1.4 [34] (0.98B parameters). We finetune the U-Net with LoRA, whose rank is set as 4, amounting to 0.8M trainable parameters. The diffusion model is finetuned on a user-specific training set containing pictures of a dog labeled "a photo of V_{dog}," with the aim to generate proper images under the prompt "a V_{dog} eating nachos."

To demonstrate the power of initialization, we compare NoRA and NoRA+ with LoRA and LoRA-P. The averaged training loss of considered approaches are summarized in Tab. 2. It can be seen that NoRA and NoRA+ have 9.6% smaller training loss compared with LoRA and LoRA-P, demonstrating the benefits of directional alignment at initialization. The generated images are listed in Fig. 3. Some of images generated by LoRA are not natural. For instance, the third one does not have a nice expression for nachos, and the tenth is not vivid. For LoRA-P, the dog in the third image is also not

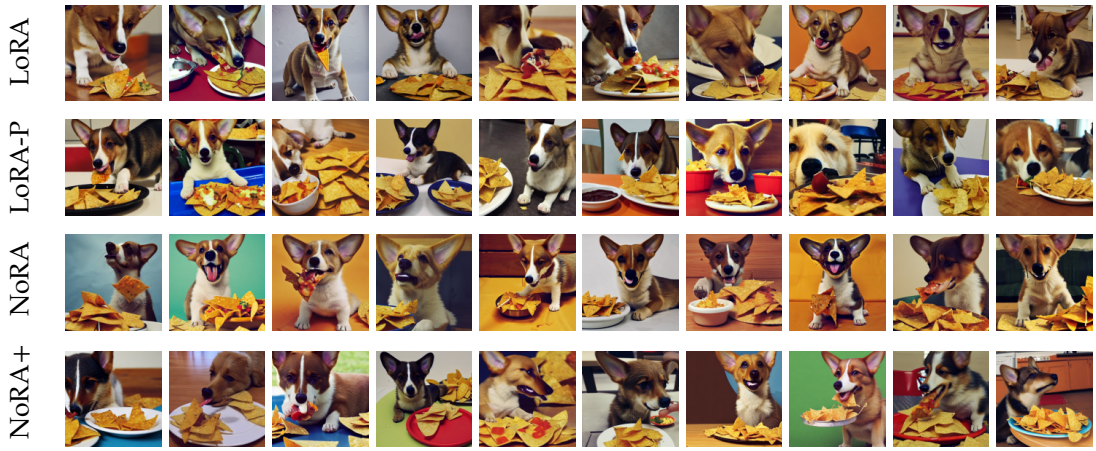


Figure 3: Generated images from NoRA and NoRA+ with stable-diffusion.

Table 3: Test accuracy of various algorithms for commonsense reasoning on LLaMA2-7B. The results marked with ‡ are taken from [38].

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	81.00	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	80.20	**78.92**

natural. NoRA and NoRA+, on the other hand, both generate high-fidelity pictures. However, there is a floating plate in the 8th image of NoRA+, but ensuring diffusion models to follow physical laws goes beyond the scope of this work. Additional results are provided in Apdx. E.5, where we finetune on images of a cat toy. The generated images from NoRA and NoRA+ have more lively facial details compare to those not using Nyström initialization.

5.2. Commonsense reasoning with LLaMA-7B and LLaMA2-7B

Our evaluation is further scaled to LLMs using LLaMA2-7B [35, 36]. We tackle commonsense reasoning tasks following the setup in [37]. Training data are merged from 8 datasets listed in Tab. 3. The test sets remain separate for individual evaluation. The rank of LoRA is chosen as 32.

The numerical results on LLaMA2-7B are presented in Tab. 3. It is observed that LoRA is unstable, henceforth the results for LoRA are taken from [38]. This instability is not observed in other tested approaches. The benefit of the Nyström initialization is particularly pronounced. These results underscore the significance of initialization for optimizing LoRA.

Additional numerical results on OPT-1.3B and Gemma-7B are provided in Apdx. E.

6. Concluding remarks

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- and over-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 with LLMs and diffusion models.

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

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

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

[19] 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.

[20] Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Proc. Int. Conf. on Artificial Intelligence and Statistics* (*AISTATS*), pages 479–488. PMLR, 2017.

[21] 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.

[22] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *Proc. Int. Conf. on Machine Learning* (*ICML*), pages 15200–15238, 2023.

[23] 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.

[24] 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.

[25] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[26] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Gradient descent with adaptive stepsize converges (nearly) linearly under fourth-order growth. *arXiv:2409.19791*, 2024.

[27] Radu-Alexandru Dragomir and Yurii Nesterov. Convex quartic problems: homogenized gradient method and preconditioning. *arXiv:2306.17683*, 2023.

[28] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proc. Int. Conf. on Machine Learning* (*ICML*), pages 964–973. PMLR, 2016.

[29] Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *Proc. Int. Conf. on Machine Learning* (*ICML*), pages 567–575. PMLR, 2013.

[30] 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.

[31] 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.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Neural Information Processing Systems* (*NeurIPS*), volume 32, 2019.

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. Conf. Computer Vision and Pattern Recognition* (*CVPR*), pages 22500–22510, 2023.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. Conf. Computer Vision and Pattern Recognition* (*CVPR*), pages 10684–10695, 2022.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[36] 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.

[37] 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.

[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] 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.

[40] Yanjie Dong, Xiaoyi Fan, Fangxin Wang, Chengming Li, Victor Leung, and Xiping Hu. Fine-tuning and deploying large language models over edges: Issues and approaches. *arXiv:2408.10691*, 2024.

[41] 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.

[42] Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. DPZero: Private fine-tuning of language models without backpropagation. In *Proc. Int. Conf. on Machine Learning* (*ICML*), 2024.

[43] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv:2402.11592*, 2024.

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

[45] 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.

[46] 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.

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

[48] 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.

[49] Bingcong Li and Georgios Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *Proc. Neural Information Processing Systems* (*NeurIPS*), volume 36, 2023.

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

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

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

[59] 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.

[60] 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.

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

[62] 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.

[63] Zachary Frangella, Joel A Tropp, and Madeleine Udell. Randomized nyström preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2):718–752, 2023.

[64] 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.

[65] 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.

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

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

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

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

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

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

[82] 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.

[83] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

[84] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *Proc. Int. Conf. on Learning Representations* (*ICLR*), 2024.

[85] 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.

[86] 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.

[87] Gemma-team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295*, 2024.

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

A. Missing details

A.1. 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.

A.2. More on related work

Convergence of over-parametrized 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 $\mathbf{Y} \in \mathbb{R}^{n\times r}$. Over-parametrization refers to the case where $\mathrm{rank}(\mathbf{A}) \le r$. The gradient flow on the extreme over-parametrized problems, where $r \ge \max\{m, n\}$, is studied in [39]. There are also papers [10, 21–23] considering the matrix sensing problem, which partially relates to our problem when there are sufficient Gaussian measures. The work of [12] considers deeper problem (i.e., having more than 3 layers) while assuming $\mathbf{A}$ is full rank. Our results on over-parametrization can be found in Apdx. B.4 and Apdx. C.3 for symmetric and asymmetric problems, respectively. The comparison of ScaledGD with other works on over-parametrized problems can be found in Tab. 1.

LoRA and parameter-efficient finetuning. LoRA [11] is a notable example of parameter-efficient finetuning (PEFT) [40] 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 [41], zeroth-order optimizers [31, 42, 43], and prefix tuning [44]. There are also various efforts to further enhance LoRA via adaptivity [45], chaining [46, 47], regularization [48, 49], low-bit training [50, 51], modifications for long-sequences [52], weight decomposition [38], and combining with sparsity [53]. Additionally, there are several approaches aiming at further reducing the number of trainable parameters in LoRA; examples include [30, 54–58]. While originally designed for finetuning LLMs, LoRA also finds its applications in other domains, such as image generation [59] and continual learning [60].

LoRA initialization. When first proposed, LoRA initialization was largely overlooked. The work of [61] justifies that whether setting $\mathbf{X}_0$ or $\mathbf{Y}_0$ to be $\mathbf{0}$ affects performance from a stability perspective. Recent works [16, 17] 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 [18] exploits stability for further improvement. However, these initialization methods are computationally expensive and lack flexibility for deployment. The proposed NoRA initialization overcomes these limitations.

Nyström sketch. Nyström sketch has well-documented success in signal processing and machine learning for coping with large-scale matrices under memory constraints. It has been applied in various settings; see e.g., [29, 62, 63]. This work only employs this approach to ensure full-rankness required in certain settings, yet the properties for recovery is not explored. We believe that other sketches are also applicable once full rankness is ensured.

A.3. LoRA for linear models as asymmetric matrix factorization

We argue that LoRA applied on linear models given a whitened dataset is equivalent to the asymmetric matrix factorization problem. The whitened dataset is widely adopted for theoretical analyses, and we refer to [12, 13, 15] for more details.

Algorithm 1 NoRA for a specific LoRA layer

Initialize: ξ – standard deviation of random matrix $\mathbf{\Omega}$
Set $\mathbf{X}_0$ and $\mathbf{Y}_0$ via Nyström initialization (5)
Standard training process

Algorithm 2 NoRA+ for a specific LoRA layer

Initialize: ξ – standard deviation of random matrix $\mathbf{\Omega}$; λ – numerical stability of matrix inversion
Set $\mathbf{X}_0$ and $\mathbf{Y}_0$ via Nyström initialization (5)
for $t = 0, \ldots, T-1$ **do**
    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

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

$$\frac{1}{2}\|(\mathbf{W}_0 + \mathbf{X}\mathbf{Y}^\top) - \mathbf{B}\|_{\mathsf{F}}^2. \tag{8}$$

It is clearly that this problem (8) is the same as (4) by setting $\mathbf{A} = \mathbf{B} - \mathbf{W}_0$.

Unfortunately, existing works provide no theoretical support on the most widely adopted initialization approach for LoRA 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 (5) 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 [5, 64, 65]. There are also recent efforts using insights from matrix factorization to further improve LoRA; see e.g., [15, 53].

A.4. More on NoRA and NoRA+

As discussed in Sec. 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 beneficial to most optimizers. While our theoretical results focus on ScaledGD, we believe that the aligned directions also improve GD. Despite the improvement may be less significant as in ScaledGD, we conjecture that the linear term in [6, Theorem 1.1] can be removed with Nyström initialization, because it can be roughly understood as the price of 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 (6). These two reasons prompt the proposed NoRA, as summarized in Alg. 1. For NoRA+ in Alg. 2, we modify the vanilla ScaledGD iterations in (6) 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 [66, 67]. 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 [14], 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 pipeline for deployment. This is because both of NoRA and LoRA do not need to modify the pretrained 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 [16, 17] are less efficient for using jointly with LoRA at deployment because both approaches modify the pretrained weights, 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.

B. Missing proofs for symmetric settings

B.1. Initialization of exact- and under-parametrized problems

B.1.1. Proof of Lemma 1

Proof. Let the compact 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

$$\mathbf{X}_0 = (\mathbf{Q}\mathbf{\Sigma})(\mathbf{Q}^\top\mathbf{\Omega}). \tag{9}$$

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 19 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 $\sigma_{\min}(\mathbf{CD}) \ge \sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})$ holds given full column rank of $\mathbf{C}$; see Lemma 17. Applying it to (9), 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 $1-\delta$. □

B.2. Missing proofs for the symmetric and exact-parametrized setting

In the exact-parametrized setting, it is convenient to define

$$\mathbf{B}_t := \mathbf{\Phi}_t\mathbf{\Phi}_t^\top \tag{10}$$

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 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.,

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

B.2.1. Proof of Lemma 2

Proof. The proof relies on $\mathbf{B}_t$ defined in (10). 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 $\mathrm{rank}(\mathbf{B}_0) = \mathrm{rank}(\mathbf{X}_0\mathbf{X}_0^\top) = r$, 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

$$\begin{aligned}
\mathbf{X}_{t+1} &= \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A})\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t)^{-1} \\
&= \mathbf{Q}\mathbf{\Phi}_t - \eta\mathbf{Q}(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{Q}^\top\mathbf{Q}\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{\Phi}_t)^{-1} \\
&\overset{(a)}{=} \mathbf{Q}\Big(\mathbf{\Phi}_t - \eta(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}\Big) \\
&\overset{(b)}{=} \mathbf{Q}\underbrace{\Big((1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}\Big)}_{:=\mathbf{\Phi}_{t+1}},
\end{aligned} \tag{11}$$

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 (11),

$$\begin{aligned}
\mathbf{B}_{t+1} = \mathbf{\Phi}_{t+1}\mathbf{\Phi}_{t+1}^\top &= (1-\eta)^2\mathbf{\Phi}_t\mathbf{\Phi}_t^\top + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}\mathbf{\Phi}_t^{-1}\mathbf{\Sigma} \\
&= (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma}.
\end{aligned} \tag{12}$$

Note that $\mathbf{B}_{t+1}$ is a PSD matrix by definition (hence the eigenvalues and singular values are the same). To see the smallest eigenvalue of $\mathbf{B}_{t+1}$ is lower bounded, we will apply Lemma 15 on (12) twice, i.e.,

$$\begin{aligned}
&\sigma_r(\mathbf{B}_{t+1}) \\
&\overset{(c)}{\geq} 2\eta(1-\eta)\sigma_r(\mathbf{\Sigma}) + \sigma_r\Big((1-\eta)^2\mathbf{B}_t + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma}\Big) \\
&\overset{(d)}{\geq} 2\eta(1-\eta)\sigma_r(\mathbf{\Sigma}) + (1-\eta)^2\sigma_r\big(\mathbf{B}_t\big) \\
&\overset{(e)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + 2\eta(1-\eta)\sigma_r(\mathbf{\Sigma})\frac{1-(1-\eta)^{2t+2}}{2\eta-\eta^2} \\
&\overset{(f)}{\geq} (1-\eta)^{2t+2}\sigma_r(\mathbf{B}_0) + (1-\eta)\sigma_r(\mathbf{\Sigma}) - (1-\eta)^{2t+3}\sigma_r(\mathbf{\Sigma}),
\end{aligned} \tag{13}$$

where (c) and (d) are because of Lemma 15; (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 (11) and (13) concludes the induction. □

B.2.2. Proof of Theorem 1

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

Lemma 7 (Phase I. Linear convergence to near optima)**.** *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 $\mathbf{\Sigma}$ from both sides of (12), we can obtain that

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

This implies that

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

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

By Lemma 2, if $\eta \le 2/3$ and there exists T_1 such that $\sigma_r(\mathbf{B}_{T_1}) \ge \sigma_r(\Sigma)/3$, then it holds that $\sigma_r(\mathbf{B}_t) \ge \sigma_r(\mathbf{\Sigma})/3, \forall t \ge T_1$. According to Lemma 1, we can choose ξ in (3) sufficiently large such that $\sigma_r(\mathbf{B}_0) \ge \sigma_r(\mathbf{\Sigma})/3$, i.e., $T_1 = 0$. Alternatively, to avoid such a requirement on ξ, we can simply choose a constant step size, e.g., $\eta = 0.5$, and run a constant number of steps, $T_1 = \mathcal{O}(1/\eta)$, to ensure $\sigma_r(\mathbf{B}_{T_1}) \ge \sigma_r(\Sigma)/3$; see Lemma 2. For simplicity of the results, our proof below goes with the first method, i.e., $T_1 = 0$.

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

where (c) is by Lemma 14. 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 $\|\mathbf{B}_{t+1} - \mathbf{\Sigma}\|_{\mathsf{F}} \le \mathcal{O}(1/\kappa^2)$ 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, $\sigma_r(\mathbf{B}_{T_2}) \ge \sigma_r(\mathbf{\Sigma})/3$ and $\|\mathbf{B}_{T_2} - \mathbf{\Sigma}\|_{\mathsf{F}} \le 2/(3\kappa^2)$ are satisfied, ScaledGD then ensures that for any $t \ge T_2$,*

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

Proof. Let $\mathbf{C}_t = \mathbf{\Sigma}^{-1}\mathbf{B}_t$. We can rewrite (12) 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

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

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

Finally, applying Lemma 16, it can be seen that a quadratic rate can be established long as $\|\mathbf{B}_{T_2} - \mathbf{\Sigma}\|_{\mathsf{F}} \le \frac{2}{3\kappa^2}$, and this condition is satisfied from Lemma 7. □

B.3. Missing proofs for the symmetric and under-parametrized setting

We start with some notation that would be helpful for this subsection. Let the compact 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 4, we will prove that $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ always holds 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 $\mathbf{B}_t = \mathbf{X}_t^\top\mathbf{A}^\dagger\mathbf{X}_t$.

B.3.1. Proof of Lemma 3

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

$$\mathbf{A} = [\mathbf{Q}_1, \mathbf{Q}_2]\begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_2\end{bmatrix}\begin{bmatrix}\mathbf{Q}_1^\top\\ \mathbf{Q}_2^\top\end{bmatrix} = \mathbf{Q}_1\mathbf{\Sigma}_1\mathbf{Q}_1^\top + \mathbf{Q}_2\mathbf{\Sigma}_2\mathbf{Q}_2^\top \tag{14}$$

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 (1) 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. □

B.3.2. Proof of Lemma 4

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

$$\begin{aligned}\mathbf{\Phi}_{t+1} &= \mathbf{\Phi}_t - \eta\big(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma}\big)\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}\\ &= (1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\underbrace{\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^{-1}}_{:=\mathbf{\Theta}_t}.\end{aligned} \tag{15}$$

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 9. □

B.3.3. Proof of Theorem 2

Proof. Using $\mathbf{\Phi}_t^\top\mathbf{\Theta}_t = \mathbf{I}_r$, definition of $\mathbf{B}_t = \mathbf{\Phi}_t^\top\mathbf{\Sigma}^{-1}\mathbf{\Phi}_t$ (at the start of Apdx. B.3), and the update of $\mathbf{\Phi}_{t+1}$ in (15), it is not difficult to verify that

$$\mathbf{B}_{t+1} = (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{I}_r + \eta^2\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t. \tag{16}$$

Subtracting $\mathbf{I}_r$ on both sides of (16), 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

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

where the last inequality is because of Lemma 10. Suppose that $\eta \leq 2/3$, from Lemma 9, one can see that there exists a time T_1 such that $\sigma_r(\mathbf{B}_t) \geq 1/3, \forall t \geq T_1$. 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

$$\begin{aligned} &\|\mathbf{B}_{t+1} - \mathbf{I}_r\|_{\mathsf{F}} \\ &\leq (1-\eta)\|\mathbf{B}_t - \mathbf{I}_r\|_{\mathsf{F}} + \eta^2\sqrt{r} + 3r\eta^2 \\ &\leq \eta\sqrt{r} + 3\eta r + (1-\eta)^{t+1-T_1}\|\mathbf{B}_{T_1} - \mathbf{I}_r\|_{\mathsf{F}} \\ &\leq \eta\sqrt{r} + 3\eta r + (1-\eta)^{t+1-T_1}\|\mathbf{B}_{T_1} - \mathbf{I}_r\|_{\mathsf{F}}. \end{aligned} \tag{17}$$

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

B.3.4. Proof of Lemma 5

Proof. We start with notation. Let

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{bmatrix}, \quad \mathbf{\Phi}_t = \begin{bmatrix} \mathbf{M}_t \\ \mathbf{N}_t \end{bmatrix}, \tag{18}$$

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., $\epsilon \to 0$ and $\mathbf{B}_t = \mathbf{I}_r$. Using (18) to rewrite $\mathbf{B}_t = \mathbf{I}_r$, we have that

$$\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1}\mathbf{M}_t + \mathbf{N}_t^\top \mathbf{\Sigma}_2^{-1}\mathbf{N}_t = \mathbf{I}_r. \tag{19}$$

The above equation implies that

$$\begin{aligned} \mathrm{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1}\mathbf{M}_t) &= \mathrm{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t) \\ &= \|\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t\|_{\mathsf{F}}^2 \overset{(a)}{\leq} r \end{aligned} \tag{20}$$

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

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

$$\begin{aligned} &\|\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t - \mathbf{I}_r\|_{\mathsf{F}}^2 \\ &= \mathrm{Tr}\Big((\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t - \mathbf{I}_r)^\top(\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t - \mathbf{I}_r)\Big) \\ &= \mathrm{Tr}\big(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t\big) + \mathrm{Tr}(\mathbf{I}_r) - 2\mathrm{Tr}(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}) \\ &\overset{(a)}{\leq} \mathrm{Tr}\big(\mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t\big) + \mathrm{Tr}(\mathbf{I}_r) + 2r^{3/2} \\ &\overset{(b)}{\leq} 2r + 2r^{3/2}, \end{aligned} \tag{21}$$

where (a) is because that i) for any $r \times r$ matrix $\mathbf{C}$, we have that $\mathrm{Tr}(\mathbf{C}) \geq r\min_i \mathbf{C}_{ii} \geq -r\|\mathbf{C}\|_{\mathsf{F}}$, ii) take $\mathbf{C} = \mathbf{M}_t^\top \mathbf{\Sigma}_1^{-1/2}$ and then apply (20); and (b) is by (20).

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

$$\frac{1}{\sigma_{r+1}(\mathbf{A})}\mathrm{Tr}\big(\mathbf{N}_t^\top \mathbf{N}_t\big) \leq \mathrm{Tr}\big(\mathbf{N}_t^\top \mathbf{\Sigma}_2^{-1}\mathbf{N}_t\big) \overset{(c)}{\leq} r \tag{22}$$

where (c) is by applying Lemma 18 on (19). 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 $\mathbf{X}_* = \mathbf{Q}[\mathbf{\Sigma}_1^{1/2}, \mathbf{0}]^\top$ and $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$. Using this fact and combining (21) and (22), we have that

$$\begin{aligned}\|\mathbf{X}_t - \mathbf{X}_*\|_\mathsf{F}^2 &= \|\mathbf{M}_t - \mathbf{\Sigma}_1^{1/2}\|_\mathsf{F}^2 + \|\mathbf{N}_t\|_\mathsf{F}^2 \\ &= \|\mathbf{\Sigma}_1^{1/2}(\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t - \mathbf{I}_r)\|_\mathsf{F}^2 + \|\mathbf{N}_t\|_\mathsf{F}^2 \\ &\leq \sigma_1(\mathbf{\Sigma}_1^{1/2})^2\|\mathbf{\Sigma}_1^{-1/2}\mathbf{M}_t - \mathbf{I}_r\|_\mathsf{F}^2 + \|\mathbf{N}_t\|_\mathsf{F}^2 \\ &= \mathcal{O}(r^{3/2}),\end{aligned} \tag{23}$$

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

B.3.5. Useful lemmas for symmetric and under-parametrized 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 15.

Lemma 9. *$\sigma_r(\mathbf{B}_t)$ 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 15 on (16) 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 14 to unroll $\sigma_r(\mathbf{B}_t)$; and (b) is because $\frac{2\eta}{2\eta-\eta^2} \geq 1$. □

Lemma 10. *Let $\mathbf{\Theta}_t$ and $\mathbf{B}_t$ defined the same as those in Apdx. B.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

$$\|\mathbf{\Theta}_t^\top\mathbf{\Sigma}\mathbf{\Theta}_t\|_\mathsf{F} = \|\mathbf{\Theta}_t^\top\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Theta}_t\|_\mathsf{F} \leq \|\mathbf{\Sigma}^{1/2}\mathbf{\Theta}_t\|_\mathsf{F}^2. \tag{24}$$

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

$$\sqrt{r} = \|\mathbf{F}_t^\top\mathbf{E}_t\|_\mathsf{F} \overset{(a)}{\geq} \sigma_r(\mathbf{F}_t)\|\mathbf{E}_t\|_\mathsf{F} \overset{(b)}{=} \sqrt{\sigma_r(\mathbf{B}_t)}\|\mathbf{E}_t\|_\mathsf{F}, \tag{25}$$

where (a) 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 $\mathbf{E}_t$). Note that (a) does not hold true for general two matrices $\mathbf{E}_t$ and $\mathbf{F}_t$. (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 (24), 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. □

B.4. Symmetric and over-parametrized setting

Nyström initialization for over-parametrization. While the initialization still follows (3), we need to adapt Lemma 1 to the over-parameterized setting, i.e., $r > r_A$.

Lemma 11 (Initialization for over-parametrization)**.** *There exists a universal constant $\tau > 0$ such that $\sigma_{r_A}(\mathbf{X}_0) \geq \xi\tau(\sqrt{r} - \sqrt{r_A - 1})\sigma_{r_A}(\mathbf{A})$ is satisfied with high probability. In other words, rank$(\mathbf{X}_0) = r_A$ w.h.p.*

Proof. Similar to the proof of Lemma 1, let the compact 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 $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$. This implies that $\mathbf{X}_0 = (\mathbf{Q}\mathbf{\Sigma})(\mathbf{Q}^\top\mathbf{\Omega})$.

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 19 on $(\mathbf{Q}^\top\mathbf{\Omega})^\top$, and using the fact $(\mathbf{Q}^\top\mathbf{\Omega})^\top$ and $(\mathbf{Q}^\top\mathbf{\Omega})$ share the same singular values, it can be seen that

$$\mathbb{P}\Big(\frac{\sigma_{r_A}(\mathbf{Q}^\top\mathbf{\Omega})}{\xi} \leq \tau(\sqrt{r} - \sqrt{r_A - 1})\Big) \leq (C_1\tau)^{r-r_A+1} + e^{-C_2 r} := \delta_2$$

where C_1 and C_2 are universal constants independent of r_A and r. This inequality shows that with probability at least $1 - \delta_2$, $\sigma_{r_A}(\mathbf{Q}^\top\mathbf{\Omega}) \geq \xi\tau(\sqrt{r} - \sqrt{r_A - 1})$.

Note that inequality $\sigma_{\min}(\mathbf{CD}) \geq \sigma_{\min}(\mathbf{C})\sigma_{\min}(\mathbf{D})$ holds given full column rank of $\mathbf{C}$; see Lemma 17. Applying it to (9), we have that

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

where (a) holds with probability at least $1 - \delta_2$. □

Next, we provide additional results of Nyström initialization on over-paramtrized setting of problem (1), where we have $r_A < r$. For a desirable convergence rate, we need to slightly modify the ScaledGD update to

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{A})\mathbf{X}_t(\mathbf{X}_t^\top\mathbf{X}_t)^\dagger. \tag{26}$$

Compared with iteration (2) for exact-parametrization, the modification is on $(\mathbf{X}_t^\top\mathbf{X}_t)^\dagger$. This pseudo-inverse is necessary because $(\mathbf{X}_t^\top\mathbf{X}_t)$ is not necessarily invertible in the over-parametrized setting. We note that unlike previous work [24] which modifies the same term to $(\mathbf{X}_t^\top\mathbf{X}_t + \lambda\mathbf{I})^{-1}$, (26) does not need the damping parameter $\lambda\mathbf{I}$ in the preconditioner. We will observe shortly in Fig. 4 that the quadratic rate is not achieved with the damping factor.

Let the compact eigendecomposition of $\mathbf{A} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$ for $\mathbf{Q} \in \mathbb{R}^{m\times r_A}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r_A\times r_A}$. We can also establish that $\mathbf{X}_t$ affords a simpler representation.

Lemma 12. *Under the Nyström initialization (3) and iteration (26), the variable $\mathbf{X}_t$ can be written as $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ for some $\mathbf{\Phi}_t \in \mathbb{R}^{r_A\times r}$. Moreover, we have that*

$$\mathbf{\Phi}_{t+1} = (1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}(\mathbf{\Phi}_t^\dagger)^\top. \tag{27}$$

Proof. We prove this by induction. Clearly, our initialization satisfies this because $\mathbf{X}_0 = \mathbf{A}\mathbf{\Omega} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{\Omega}$, i.e., $\mathbf{\Phi}_0 := \mathbf{\Sigma}\mathbf{Q}^\top\mathbf{\Omega}$. Now suppose that $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ holds for t. We then show that $\mathbf{X}_{t+1} = \mathbf{Q}\mathbf{\Phi}_{t+1}$ to finish the induction. In particular, plugging $\mathbf{X}_t = \mathbf{Q}\mathbf{\Phi}_t$ into (26), we arrive at

$$\mathbf{X}_{t+1} = \mathbf{Q}\underbrace{\left[\mathbf{\Phi}_t - \eta(\mathbf{\Phi}_t\mathbf{\Phi}_t^\top - \mathbf{\Sigma})\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^\dagger\right]}_{:=\mathbf{\Phi}_{t+1}}.$$

Clearly, the term inside the brackets is $\mathbf{\Phi}_{t+1}$. The induction is thus finished.

Now we proof the second part of this lemma. Let the SVD of $\mathbf{\Phi}_t := \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^\top$, where $\mathbf{U}_t \in \mathbb{R}^{r_A\times r_A}$, $\mathbf{\Sigma}_t \in \mathbb{R}^{r_A\times r_A}$, and $\mathbf{V}_t \in \mathbb{R}^{r\times r_A}$. We note that $\mathbf{U}_t$ is unitary for this case. With the SVD, we have that $\mathbf{\Phi}_t\mathbf{\Phi}_t^\top = \mathbf{U}_t\mathbf{\Sigma}_t^2\mathbf{U}_t^\top$, and $(\mathbf{\Phi}_t^\top\mathbf{\Phi}_t)^\dagger = \mathbf{V}_t\mathbf{\Sigma}_t^{-2}\mathbf{V}_t^\top$. Plugging these into $\mathbf{\Phi}_{t+1}$ defined earlier, we arrive at

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

This completes the proof. □

Next, let $\mathbf{B}_t = \mathbf{\Phi}_t\mathbf{\Phi}_t^\top$. With (27) we have that

$$\begin{aligned}\mathbf{B}_{t+1} &= (1-\eta)^2\mathbf{\Phi}_t\mathbf{\Phi}_t^\top + \eta(1-\eta)\mathbf{\Phi}_t\mathbf{\Phi}_t^\dagger\mathbf{\Sigma} + \eta(1-\eta)\mathbf{\Sigma}(\mathbf{\Phi}_t^\dagger)^\top\mathbf{\Phi}_t^\top + \eta^2\mathbf{\Sigma}(\mathbf{\Phi}_t^\dagger)^\top\mathbf{\Phi}_t^\dagger\mathbf{\Sigma} \\ &\overset{(a)}{=} (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}(\mathbf{\Phi}_t^\dagger)^\top\mathbf{\Phi}_t^\dagger\mathbf{\Sigma} \\ &\overset{(b)}{=} (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma},\end{aligned} \tag{28}$$

where in (a) we used the SVD of $\mathbf{\Phi}_t := \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^\top$, where $\mathbf{U}_t \in \mathbb{R}^{r_A\times r_A}$, $\mathbf{\Sigma}_t \in \mathbb{R}^{r_A\times r_A}$ and $\mathbf{V}_t \in \mathbb{R}^{r\times r_A}$, $\mathbf{\Phi}_t^\dagger = \mathbf{V}_t\mathbf{\Sigma}_t^{-1}\mathbf{U}_t^\top$, and $\mathbf{U}_t$ is unitary; and in (b) we assume that $\mathbf{B}_t$ is full rank. Note that this assumption can be easily verified given $\text{rank}(\mathbf{B}_0) = r_A$; and the iteration on $\mathbf{B}_t$ (28) is exactly the same as in exact-parametrized cases (12). The latter allows us to bound $\sigma_{r_A}(\mathbf{B}_t)$ away from 0 in the same way as Lemma 2.

In other words, the over-parametrized case under our initialization reduces to the exact-parametrized case given the same iteration on $\mathbf{B}_t$ (28) (cf. (12)). This allows as to use the same argument of Theorem 1 to derive a quadratic rate for over-parametrized case.

Theorem 5. *With high probability over the initialization, the behavior of update (26) under Nyström initialization (3) can be described as:*

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 $\|\mathbf{X}_{T_1}\mathbf{X}_{T_1}^\top - \mathbf{A}\|_\mathsf{F} \le \mathcal{O}(1/\kappa^2)$.

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

Proof. The proof is the same as Theorem 1 given the same iteration on $\mathbf{B}_t$ in (28). We omit it to avoid redundancy. □

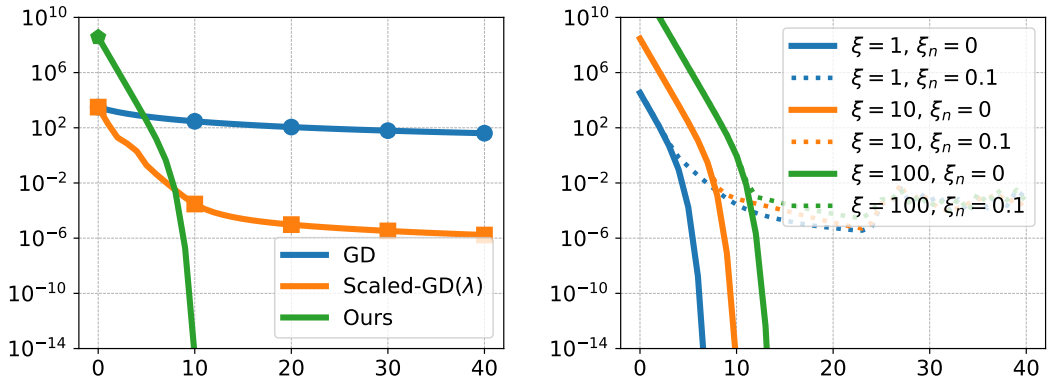
Numerical illustration. A numerical illustration for ScaledGD under Nyström initialization in over-parametrized case can be found in Fig. 4. We adopt ScaledGD-(λ) [24], the damping version of ScaledGD, as another baseline. It can be seen that only our approach achieves a quadratic rate; see Fig. 4(a). We also slightly perturb our initialization with small noise, and it can be seen that the quadratic convergence breaks down immediately. This demonstrate the critical role of initialization: i) it helps to get rid of damping using pseudo-inverse; and ii) it ensures a quadratic rate.

C. Missing proofs for asymmetric settings

C.1. Missing proofs for asymmetric and exact-parametrized setting

C.1.1. Proof of Lemma 6

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



(a) over-parametrization (b) over-parametrization

Figure 4: Convergence of ScaledGD under Nyström initialization (optimality error vs. iteration) on over-parametrized problems detailed in Apdx. E.1. (a) Comparison of GD, ScaledGD-(λ) with small initialization, and ScaledGD with our initialization. (b) Solid lines show that our initialization is not sensitive to magnitude; and dotted lines illustrate that quadratic convergence cannot be obtained even with slightly perturbed initialization, i.e., $\mathbf{X}_0 = \mathbf{A}\boldsymbol{\Omega} + \mathbf{N}$, where $[\mathbf{N}]_{ij} \sim \mathcal{N}(0, \xi_n^2)$.

will show that $\mathbf{X}_{t+1} = \mathbf{U}\boldsymbol{\Phi}_{t+1}$ and $\mathbf{Y}_{t+1} = \mathbf{V}\boldsymbol{\Psi}_{t+1}$ under iteration (6). Let us start with $\mathbf{X}_{t+1}$. Note that if $t = 0$, $\mathbf{X}_1 = \mathbf{U}\boldsymbol{\Phi}_1$ is trivial. We only focus on $t \geq 1$, where we have

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

Note that the invertible of $(\boldsymbol{\Psi}_t^\top\boldsymbol{\Psi}_t)$ will become clear in the proof of Corollary 1.

Using a similar argument, it is not hard to show that $\mathbf{Y}_t = \mathbf{V}\boldsymbol{\Psi}_t$ for all t. We do not repeat here. □

C.1.2. Proof of Theorem 3

Proof. Based on the initialization (5) and iteration (6), we can obtain that

$$\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_0 \tag{29a}$$

$$\begin{aligned}
\boldsymbol{\Psi}_1 = \mathbf{V}^\top\mathbf{Y}_1 &= \mathbf{0} - \eta\mathbf{V}^\top(\mathbf{0} - \mathbf{A})^\top\mathbf{U}\boldsymbol{\Phi}_0(\boldsymbol{\Phi}_0^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Phi}_0)^{-1} \\
&= \eta\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Phi}_0(\boldsymbol{\Phi}_0^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Phi}_0)^{-1} \\
&= \eta\boldsymbol{\Sigma}\boldsymbol{\Phi}_0(\boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0)^{-1} \\
&= \eta\boldsymbol{\Sigma}\boldsymbol{\Phi}_0^{-\top}.
\end{aligned} \tag{29b}$$

This ensures that

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

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

C.1.3. Proof of Corollary 1

Proof. The corollary is 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 3, 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 Lemma 6, we can write the iteration as

$$\mathbf{\Phi}_{t+1} = (1-\eta)\mathbf{\Phi}_t + \eta\mathbf{\Sigma}\mathbf{\Psi}_t^{-\top} \tag{30a}$$

$$\mathbf{\Psi}_{t+1} = (1-\eta)\mathbf{\Psi}_t + \eta\mathbf{\Sigma}\mathbf{\Phi}_t^{-\top}. \tag{30b}$$

This gives that

$$\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top = (1-\eta)^2\mathbf{\Phi}_t\mathbf{\Psi}_t^\top + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}(\mathbf{\Phi}_t\mathbf{\Psi}_t^\top)^{-1}\mathbf{\Sigma}. \tag{31}$$

The symmetry of $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$ directly follows from (31). For the positive definiteness of $\mathbf{\Phi}_{t+1}\mathbf{\Psi}_{t+1}^\top$, we can apply Lemma 15 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 (31) can be rewritten as

$$\mathbf{B}_{t+1} = (1-\eta)^2\mathbf{B}_t + 2\eta(1-\eta)\mathbf{\Sigma} + \eta^2\mathbf{\Sigma}\mathbf{B}_t^{-1}\mathbf{\Sigma} \tag{32}$$

which is exactly the same iteration as (12) for the symmetric exact-parametrized 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 exact-parametrized problems, i.e., Theorem 1 to get the bounds stated in this corollary. We do not repeat for conciseness. □

C.2. Missing proofs for asymmetric and under-parametrized setting

C.2.1. How good is weak optimality?

Lemma 13. *Every global optimum for (4) is also weakly optimal.*

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

$$\mathbf{A} = [\mathbf{U}_1, \mathbf{U}_2]\begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^\top \\ \mathbf{V}_2^\top\end{bmatrix} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^\top + \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^\top \tag{33}$$

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 (1) 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.

□

C.2.2. Proof of Theorem 4

Proof. The update in (6) ensures that

$$\mathbf{\Phi}_1 = \mathbf{\Phi}_0, \tag{34a}$$

$$
\begin{aligned}
\mathbf{\Psi}_1 = \mathbf{V}^\top \mathbf{Y}_1 &= \mathbf{0} - \eta \mathbf{V}^\top (\mathbf{0} - \mathbf{A})^\top \mathbf{U} \mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U} \mathbf{\Phi}_0)^{-1} \\
&= \eta \mathbf{V}^\top \mathbf{V} \mathbf{\Sigma} \mathbf{U}^\top \mathbf{U} \mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{U}^\top \mathbf{U} \mathbf{\Phi}_0)^{-1} \\
&= \eta \mathbf{\Sigma} \mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{\Phi}_0)^{-1} \\
&\overset{(a)}{:=} \eta \mathbf{\Sigma} \mathbf{\Theta}_0
\end{aligned} \tag{34b}
$$

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

From the Definition 2, 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 $\eta = 1$, generalized weak optimality can be achieved in one step for under-parametrized problems. □

C.3. Asymmetric and over-parametrized setting

Next, we establish the one step convergence with Nyström initialization in the asymmetric over-parametrized setting, where $r_A < r$. We also need to slightly modify the ScaledGD update to

$$
\mathbf{X}_1 = \mathbf{X}_0, \text{ and } \mathbf{X}_{t+1} = \mathbf{X}_t - \eta (\mathbf{X}_t \mathbf{Y}_t^\top - \mathbf{A}) \mathbf{Y}_t (\mathbf{Y}_t^\top \mathbf{Y}_t)^\dagger, \forall t \geq 1 \tag{35a}
$$

$$
\mathbf{Y}_{t+1} = \mathbf{Y}_t - \eta (\mathbf{X}_t \mathbf{Y}_t^\top - \mathbf{A})^\top \mathbf{X}_t (\mathbf{X}_t^\top \mathbf{X}_t)^\dagger, \forall t \geq 0. \tag{35b}
$$

Comparing with (6), the difference is that here we use pseudo-inverse to bypass the possible non-invertibility of $(\mathbf{X}_t^\top \mathbf{X}_t)$ and $(\mathbf{Y}_t^\top \mathbf{Y}_t)$ in the over-parametrized case. We also note that to the best of our knowledge, there is no previous result that establishes the convergence of ScaledGD (or its variants) for asymmetric over-parametrized problems.

Theorem 6. *Under Nyström initialization (5), the modified ScaledGD iterations (35) converge globally in a single step w.h.p. over the initialization, i.e., $\mathbf{X}_1 \mathbf{Y}_1^\top = \mathbf{A}$ if the learning rate is chosen as $\eta = 1$.*

Proof. Let the compact eigendecomposition of $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$ for $\mathbf{U} \in \mathbb{R}^{m \times r_A}$, $\mathbf{\Sigma} \in \mathbb{R}^{r_A \times r_A}$, and $\mathbf{V} \in \mathbb{R}^{n \times r_A}$.

The Nyström initialization ensures that $\mathbf{X}_0 = \mathbf{X}_1 = \mathbf{U} \mathbf{\Phi}_0$, where $\mathbf{\Phi}_0 \in \mathbb{R}^{r_A \times r}$ and clearly $\mathbf{\Phi}_0 = \mathbf{\Sigma} \mathbf{V}^\top \mathbf{\Omega}$. Using the expression of $\mathbf{X}_1$, iteration (35) gives that

$$
\mathbf{Y}_1 = \eta \mathbf{V} \mathbf{\Sigma} \mathbf{\Phi}_0 (\mathbf{\Phi}_0^\top \mathbf{\Phi}_0)^\dagger.
$$

Let the compact SVD of $\mathbf{\Phi}_0 := \mathbf{P} \mathbf{D} \mathbf{Q}^\top$, where $\mathbf{P} \in \mathbb{R}^{r_A \times r_A}$, $\mathbf{D} \in \mathbb{R}^{r_A \times r_A}$ and $\mathbf{Q} \in \mathbb{R}^{r \times r_A}$. Note that $\mathbf{P}$ is unitary. With the compact SVD of $\mathbf{\Phi}_0$, we have that $(\mathbf{\Phi}_0^\top \mathbf{\Phi}_0)^\dagger = \mathbf{Q} \mathbf{D}^{-2} \mathbf{Q}^\top$, which implies that

$$
\mathbf{X}_1 \mathbf{Y}_1^\top = \eta \mathbf{U} \mathbf{P} \mathbf{D} \mathbf{Q}^\top \mathbf{Q} \mathbf{D}^{-2} \mathbf{Q}^\top \mathbf{Q} \mathbf{D} \mathbf{P}^\top \mathbf{\Sigma} \mathbf{V}^\top \overset{(a)}{=} \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top = \mathbf{A}
$$

where (a) is because $\mathbf{P}$ is unitary and the choice of $\eta = 1$. □

D. Other useful lemmas

Lemma 14. *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 15. *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

$$\lambda_{\min}(\mathbf{A}+\mathbf{B}) = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top(\mathbf{A}+\mathbf{B})\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \min_{\mathbf{x}_1\neq\mathbf{0},\mathbf{x}_1=\mathbf{x}_2} \frac{\mathbf{x}_1^\top\mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^\top\mathbf{x}_1} + \frac{\mathbf{x}_2^\top\mathbf{B}\mathbf{x}_2}{\mathbf{x}_2^\top\mathbf{x}_2}. \tag{36}$$

On the other hand, we also have that

$$\lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) = \min_{\mathbf{x}_1\neq\mathbf{0},\mathbf{x}_2\neq\mathbf{0}} \frac{\mathbf{x}_1^\top\mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^\top\mathbf{x}_1} + \frac{\mathbf{x}_2^\top\mathbf{B}\mathbf{x}_2}{\mathbf{x}_2^\top\mathbf{x}_2}. \tag{37}$$

Because (36) is a constrained version of the minimization problem (37), they share the same objective, but (36) 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 16. *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 17. *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 18. *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 19 (Rudelson and Vershynin [68])**.** *Let $\mathbf{W}$ be an $d \times r$ matrix with $d \geq r$. The entries of $\mathbf{W}$ are drawn independently from $\mathcal{N}(0,1)$. Then for every $\tau > 0$, we have that*

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

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

Figure 5: The dog dataset.

Figure 6: The cat-toy dataset.

E. Missing experimental details

E.1. Details for problems with synthetic data

This subsection contains the detailed setup for the problems with synthetic data in Figs. 1 and 4. Recall that here we focus on symmetric problems under exact-, under-, and over-parametrization.

For the exact-parametrized problem in Fig. 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 the under-parametrized problem 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 $\{1.0, 0.99, 0.98, \ldots, 0.65, 0.64, 0.05, 0.025, 0.01\}$. We choose $r = 20$ to ensure the under-parametrized nature of this problem.

For the over-parametrized case in Fig. 4 (a) and (b), we choose PSD matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ in the following manner. We set $m = 1000$ and $r_A = 20$. The non-zero singular values are chosen 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 set $\mathbf{X}$ to be over-parametrized by letting $r = 60$. We choose the step size of GD as 0.01. The learning rate of ScaledGD-λ is set as 0.5, and its damping parameter λ is chosen as 0.01. The learning rate for ScaledGD with Nyström initialization is 0.5.

E.2. Datasets

The evaluation of NoRA and NoRA+ is carried out on commonly adopted datasets in the literature.

GLUE benchmark. GLUE is designed to provide general-purpose evaluation of language understanding [69]. Those adopted in our work include SST-2 (sentiment analysis, [70]), RTE[3] (inference). These datasets are released under different permissive licenses.

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

SuperGLUE benchmark. SuperGLUE [71] is another commonly adopted benchmark for language understanding, and it is more challenging compared with GLUE. The considered datasets include CB (inference, [72]), ReCoRD (question answering, [73]), WSC (coreference resolution, [74]), BoolQ (question answering, [75]), and MiltiRC (question answering, [76]). 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 [77], PIQA [78], SOCIAL-I-QA (SIQA) [79], HellaSwag [80], ARC-easy, ARC-challenge [81] and OpenbookQA [82]. These datasets are released under different permissive licenses.

Math. For mathematical problems, we consider GSM8K [83] dataset that consists of high quality linguistically diverse school math problems created by human problem writers. This dataset is under MIT license. We also adopt MetaMathQA dataset [84], which is constructed through bootstrapping mathematical questions by rewriting the question from multiple perspectives. This dataset is under MIT license.

Additional datasets. We also use SQuAD (question answering, [85]) in our experiments, which is released under license CC BY-SA 4.0.

Datasets for DreamBooth. The datasets (dog and cat-toy) used for Sec. 5.1 are obtained directly from Huggingface. The dog dataset[4] contains 5 dog images; see Fig. 5. The cat-toy[5] dataset has 4 images; see Fig. 6. Both datasets are representative examples for the purpose of DreamBooth – finetuning with only few images for personalized generalization.

E.3. Details for Fig. 2

The experiment setting and training protocols are the same as few-shot learning with OPT-1.3B in the following subsection. Here, we are interested in the change of singular values after LoRA finetuning. For each LoRA layer, we compare the singular values of $\mathbf{W}_0$ and $\mathbf{W}_0 + \mathbf{X}_T\mathbf{Y}_T^\top$, where $\mathbf{X}_T, \mathbf{Y}_T$ are LoRA weights after training, and find out the indices of r singular values that have the largest change after finetuning. We then count the indices across all LoRA layers. Fig. 2 plots indices vs. counts.

E.4. Few-shot learning with OPT-1.3B

Consider a few-shot learning task following [31]. 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 [69, 71]. Consistent with [31], we randomly sample 1,000 data points for training and another 1,000 for testing.

We embrace OPT-1.3B as our base model [86] and apply LoRA to the query and value matrices in the attention module. This aligns with common practice for models of this size. The rank of LoRA is set to 8, leading to approximately 1.5M trainable parameters, which is significantly less than the model size. We compare the proposed NoRA and NoRA+ with LoRA, prefix tuning [44], OLoRA [16], and PiSSA [17]. Note that the latter two serve as alternative methods for initializing LoRA.

For this experiment, we first search for the best batchsizes for LoRA, and the same batchsize is applied for other tested algorithms as well. Then we search additionally for the best learning rate for each algorithm. This ensures that different algorithms see the same amount of data, while still having their best performed learning rate. The hyperparameters adopted are searched over values in Tab. 4. Adam is adopted for optimization.

The performance of different algorithms is summarized in Tab. 5. 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, respectively.

[4] https://huggingface.co/datasets/diffusers/dog-example
[5] https://huggingface.co/datasets/diffusers/cat-toy-example

Table 4: 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

Table 5: Test accuracy 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	$92.9_{\pm0.9}$	$59.6_{\pm1.6}$	$73.1_{\pm2.3}$	$71.6_{\pm2.9}$	$65.2_{\pm2.6}$	$69.7_{\pm1.0}$	$64.4_{\pm3.2}$	$82.2_{\pm1.4}$	72.3
LoRA	$93.1_{\pm0.2}$	$59.1_{\pm2.0}$	$70.6_{\pm5.2}$	$72.6_{\pm3.7}$	$69.1_{\pm4.7}$	$70.8_{\pm1.0}$	$68.0_{\pm1.4}$	$81.9_{\pm1.8}$	73.2
OLoRA	$92.7_{\pm0.5}$	$60.0_{\pm2.3}$	$70.9_{\pm3.1}$	$80.3_{\pm2.7}$	$69.7_{\pm1.0}$	$71.3_{\pm1.2}$	$66.7_{\pm0.9}$	$80.0_{\pm1.4}$	74.0
PiSSA	$92.7_{\pm0.6}$	$60.6_{\pm3.7}$	$70.4_{\pm0.7}$	$78.0_{\pm7.2}$	$70.4_{\pm2.8}$	$70.9_{\pm1.2}$	$67.9_{\pm2.1}$	$82.1_{\pm0.4}$	74.1
NoRA	$93.4_{\pm0.7}$	$60.6_{\pm3.8}$	$73.2_{\pm0.6}$	$79.2_{\pm5.2}$	$72.0_{\pm1.3}$	$71.3_{\pm1.0}$	$68.5_{\pm1.2}$	$81.8_{\pm0.7}$	**75.0**
NoRA+	$93.2_{\pm0.5}$	$61.2_{\pm0.6}$	$72.9_{\pm1.3}$	$79.5_{\pm5.8}$	$72.4_{\pm3.6}$	$71.5_{\pm0.9}$	$68.4_{\pm1.2}$	$82.0_{\pm0.9}$	**75.1**

E.5. DreamBooth with stable-diffusion

Stable Diffusion V1.4 [34] is adopted as base model, where LoRA is applied to the UNet. The text-encoder is not finetuned. We adopt the default parameter-choice from Huggingface, which is summarized in Tab. 6. We adopt AdamW as the optimizer with a weight decay of 0.01.

Table 6: Hyperparameters used for DreamBooth with stable-diffusion.

Hyperparameters	Values
LoRA r	4
LoRA α	4
LoRA module	to_q, to_k, to_v, to_out
# iterations	500
batchsize	1
learning rate	1×10^{-4}
NoRA ξ	0.1

We provide additional results to further support the efficiency of NoRA by finetuning the stable-diffusion-v1.4 model using the same protocol as in Sec. 5.1. Here we adopt a dataset with 4 toy-cat images; see Fig. 6. After finetuning 500 steps using prompt "a photo of toy cat", our goal is to generate images "a toy cat wearing glasses." The generated images are shown in Fig. 7. In general, all tested algorithms do not distinguish the hands and the tail of toy cat well. However, both LoRA and LoRA-P generate images with less accurate facial details. For example, the glasses are not wearing well, or the eyes are not clear. However, the details of faces generated by NoRA and NoRA+ are quite clear.

E.6. Commonsense reasoning with LLaMA2

The base models considered is LLaMA2-7B. The experimental setup and choices of hyperparameters follow [38]. The hyperparameters are summarized in Tab. 7.

Figure 7: Generated images from NoRA and NoRA+ with stable-diffusion.

Table 7: 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

E.7. Math reasoning with Gemma-7B

Our last evaluation tackles mathematical reasoning. Gemma-7B [87] is finetuned for 2 epochs on MetaMathQA-100K dataset [84]. LoRA rank is set as 32, leading to 100M trainable parameters. The performance is assessed on GSM8K [83], and hyperparameters are summarized in Tab. 8.

Table 8: Hyperparameters used for math reasoning with Gemma-7B.

Hyper-parameters	Values
LoRA r (rank)	32
LoRA α	64
LoRA module	q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj
epoch	2
learning rate	3×10^{-4}, 4×10^{-4}, 5×10^{-4}
batchsize	128
NoRA ξ	0.02, 0.05, 0.1

The performance of various approaches is summarized in Tab. 9. We also include PiSSA [17] into the comparison. Note that PiSSA uses LoRA rank as 64 but is only finetuned for a single epoch. Despite this difference, the computational cost on backward passes is the same for PiSSA and NoRA. The results clearly show that NoRA (NoRA+) outperforms LoRA (LoRA-P), highlighting the effectiveness of our Nyström initialization.

Table 9: Performances of different algorithms for math reasoning tasks. The results marked with ‡ are taken from [17].

GSM8K	LoRA	PiSSA‡	NoRA	LoRA-P	NoRA+
Gemma-7B	76.72	77.94	78.62	77.03	78.47