Preconditioning Benefits of Spectral Orthogonalization in Muon

Jianhao Ma* Penn
Yu Huang* Penn
Yuejie Chi† Yale
Yuxin Chen*‡ Penn

January 19, 2026

Abstract

The `Muon` optimizer, a matrix-structured algorithm that leverages spectral orthogonalization of gradients, is a milestone in the pretraining of large language models. However, the underlying mechanisms of `Muon`—particularly the role of gradient orthogonalization—remain poorly understood, with very few works providing end-to-end analyses that rigorously explain its advantages in concrete applications. We take a step by studying the effectiveness of a simplified variant of `Muon` through two case studies: matrix factorization, and in-context learning of linear transformers. For both problems, we prove that simplified `Muon` converges linearly with iteration complexities independent of the relevant condition number, provably outperforming gradient descent and `Adam`. Our analysis reveals that the `Muon` dynamics decouple into a collection of independent scalar sequences in the spectral domain, each exhibiting similar convergence behavior. Our theory formalizes the preconditioning effect induced by spectral orthogonalization, offering insight into `Muon`'s effectiveness in these matrix optimization problems and potentially beyond.

1 Introduction

The emergence of `Muon`—a matrix-structured, spectrum-aware optimizer recently proposed by Jordan et al. (2024)—has marked a milestone in the pretraining of large language models (LLMs) and beyond. Standing for *MomentUm Orthogonalized by Newton-Schulz* and leveraging spectral orthogonalization of gradients, `Muon` was initially shown to set new training speed records on benchmarks like CIFAR-10 and NanoGPT, outperforming conventional optimizers (Jordan et al., 2024). Subsequent work has scaled `Muon` to multi-billion-parameter LLMs, demonstrating approximately a twofold improvement in training efficiency over the `AdamW` optimizer (Liu et al., 2025). Such empirical advances have positioned `Muon` as a compelling alternative to established optimizers such as `Adam` and `AdamW`, and have motivated theoretical investigation into the mechanisms underlying `Muon`'s practical efficiency.

1.1 The `Muon` algorithm and prior theory

Setting the stage, consider an unconstrained optimization problem:

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is a matrix variable. At each iteration $t \geq 0$, `Muon` executes the following update:

$$\boldsymbol{B}_t = \nabla f(\boldsymbol{X}_t) + \mu \boldsymbol{B}_{t-1}, \tag{2a}$$
$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta_t \,\mathsf{msign}(\boldsymbol{B}_t), \tag{2b}$$

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania.
†Department of Statistics and Data Science, Yale University.
‡Department of Electrical and Systems Engineering, University of Pennsylvania.

where $\boldsymbol{B}_t$ represents an auxiliary momentum-like iterate that aggregates the current gradient with exponentially discounted past gradients, $0 \leq \mu < 1$ controls the degree of momentum (exponential averaging), $\eta_t > 0$ stands for the learning rate at iteration t, and $\mathsf{msign}(\cdot)$ denotes the matrix sign function defined as

$$\mathsf{msign}(\boldsymbol{Z}) \coloneqq \arg\min_{\boldsymbol{O}} \left\{ \|\boldsymbol{Z} - \boldsymbol{O}\|_{\mathrm{F}} : \text{either } \boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I} \text{ or } \boldsymbol{O}^\top\boldsymbol{O} = \boldsymbol{I} \right\}. \tag{3}$$

Equivalently, if a matrix $\boldsymbol{Z}$ has compact singular value decomposition (SVD) $\boldsymbol{Z} = \boldsymbol{U}_Z\boldsymbol{\Sigma}_Z\boldsymbol{V}_Z^\top$—where $\boldsymbol{U}_Z$ (resp. $\boldsymbol{V}_Z$) denotes the left (resp. right) singular matrix—then its matrix sign is given by $\mathsf{msign}(\boldsymbol{Z}) = \boldsymbol{U}_Z\boldsymbol{V}_Z^\top$, although in practice $\mathsf{msign}(\cdot)$ is computed efficiently using Newton-Schulz iterations (Jordan et al., 2024; Higham, 2008). A notable special case of (2a) arises when momentum is disabled by setting $\mu = 0$, yielding the simplified update rule

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta_t\, \mathsf{msign}\left(\nabla f(\boldsymbol{X}_t)\right), \qquad t = 0, 1, \cdots \tag{4}$$

This important variant is commonly referred to as *simplified* ***Muon*** or the *spectral gradient method*. Turning off momentum substantially simplifies theoretical analysis (An et al., 2025; Shen et al., 2025; Davis and Drusvyatskiy, 2025; Su, 2025), while often retaining comparable empirical performance to its momentum-based counterpart for nonstochastic settings (Shen et al., 2025). In contrast to standard optimizers like `Adam` (Kingma, 2015) and `AdamW` (Loshchilov and Hutter, 2019) that apply independent per-coordinate preconditioning, a distinguishing feature of `Muon` or spectral gradient methods lies in the use of gradient orthogonalization: update directions are obtained by spectrally orthogonalizing the gradient estimates.

Motivated by `Muon`'s remarkable empirical success, the past year has witnessed a surge of theoretical efforts aimed at elucidating the mechanisms behind its effectiveness from diverse perspectives. From an optimization standpoint, Li and Hong (2025); Shen et al. (2025) established convergence guarantees of `Muon` on smooth objectives. In particular, Shen et al. (2025) showed that `Muon`'s convergence is governed by the gradient Lipschitz parameter defined w.r.t. the spectral norm, which can sometimes be substantially smaller than its Euclidean counterpart and hence offers a potential explanation for `Muon`'s accelerated convergence. Kovalev (2025) interpreted `Muon` as a trust region method with non-Euclidean trust regions and derived tighter convergence rates for certain function classes. Complementing this line of work, Chen et al. (2025) showed that: `Muon` (with decoupled weight decay) approximately enforces a spectral norm constraint on weight updates, which implicitly reduces the worst-case smoothness of the optimization landscape and enables the use of larger learning rates. Stepping beyond worst-case convergence guarantees, Davis and Drusvyatskiy (2025) compared the one-step progress of spectral updates (as in (4)) relative to Euclidean gradient updates, and showed that `Muon` yields a larger *one-step* reduction in the objective than gradient descent (`GD`) when the gradient rank exceeds the activation rank. Another recent work Su (2025) introduced an "isotropic curvature model"—proposed through heuristic arguments and validated empirically in transformer training—and derived gradient orthogonalization iterations as the optimal updates under certain assumptions.

Despite these theoretical pursuits, however, the theoretical foundation of `Muon` remains far from complete. In particular, very few existing results were able to offer end-to-end, rigorous analyses that provably demonstrate `Muon`'s advantages over classical optimizers in concrete applications.

1.2 This paper: preconditioning with `Muon`

In this work, we take a step towards theoretically justifying the effectiveness of `Muon` by investigating its preconditioning effect—a core feature built into its design via spectral orthogonalization—that is hypothesized to make the optimizer better align with the geometry of neural networks (Jordan et al., 2024; Bernstein and Newhouse, 2024b; Vasudeva et al., 2025; Lau et al., 2025). Rather than tackling the most general settings, we focus on two concrete, yet fundamental, matrix optimization problems: (a) matrix factorization, and (b) in-context learning of linear transformers. By focusing on these stylized applications, we develop end-to-end convergence theory unveiling provable advantages of `Muon` over classical optimizers like `GD` and `Adam`. Our main contributions are summarized below.

problem	algorithm		iterations	paper
matrix factorization	simplified `Muon`	exactly-parameterized	$\log \frac{1}{\varepsilon}$	this work (Theorem 1)
		over-parameterized	$\log \frac{1}{\varepsilon}$	this work (Theorem 1)
	`GD`	exactly-parameterized	$\kappa \log \frac{1}{\varepsilon}$	Chi et al. (2019)
		over-parameterized	$\kappa^3 \log \frac{1}{\varepsilon}$	Stöger and Soltanolkotabi (2021)
		lower bound	$\kappa \log \frac{1}{\varepsilon}$	folklore
	`SignGD`	lower bound	κ	this work (Theorem 2)
in-context learning	simplified `Muon`	exactly-parameterized	$\log \frac{1}{\varepsilon}$	this work (Theorem 3)
	`GD`	lower bound	$\sqrt{\kappa} \log \frac{1}{\varepsilon}$	d'Aspremont et al. (2021)
	`SignGD`	lower bound	κ	this work (Theorem 4)

Table 1: Summary of convergence theory for simplified `Muon`, `GD` and `SignGD` for both matrix factorization and in-context learning tasks. We report the numbers of iterations required to achieve ε-accuracy; only the orders are shown, with all preconstants omitted.

- *Matrix factorization.* We show in Theorem 1 that simplified `Muon` converges linearly, encompassing both exactly-parameterized and over-parameterized settings. Notably, `Muon`'s iteration complexity is provably independent of the condition number κ of the matrix to be factorized—a stark contrast to both `GD` and `SignGD` (a simplified variant of `Adam` with momentum disabled), whose iteration complexities scale at least linearly with κ (cf. Theorem 2).
- *In-context learning of linear transformers.* Akin to the matrix factorization case, we establish linear convergence of simplified `Muon` in Theorem 3, with an iteration complexity independent of the condition number of the target covariance matrix. This contrasts sharply with both `GD` and `SignGD`, for which we develop iteration complexity lower bounds (cf. Theorem 4) that scale polynomially with the condition number of interest.

See Table 1 for more detailed comparisons. For both problems, our results reveal that by normalizing the gradient spectrum at each iteration, `Muon` exhibits preconditioning benefits that yield provably faster, condition-number-free, convergence rates. These theoretical findings are complemented by a series of numerical experiments that corroborate the preconditioning benefits of `Muon`. At a more technical level, our analyses uncover that the dynamics of `Muon` decouple into a collection of independent scalar sequences in the spectral domain, each associated with one eigenvalue of the target matrix and exhibiting similar convergence behavior. While our theory is restricted to two simple matrix optimization problems and by no means exhaustive, we expect the preconditioning effect of `Muon` to manifest in broader applications.

1.3 Additional related work

We now provide additional discussion of related prior work. The convergence analyses in Li and Hong (2025); Shen et al. (2025); Chen et al. (2025) were motivated in part by Bernstein and Newhouse (2024b), which interpreted (some simplified variants of) `Adam`, `Shampoo`, and `Prodigy` as steepest descent under certain norm constraints. It is noteworthy that the idea of spectral initialization of gradients has appeared in earlier designs of optimizers (e.g., Carlson et al. (2015a,b,c); Tuddenham et al. (2022)). Another line of research studied the implicit bias of `Muon`. For example, Fan et al. (2025) showed that in multi-class linear classification, `Muon` (or its idealized variant with exact orthogonal updates) converges to solutions that maximize the margin w.r.t. the spectral norm of the weight matrix, which contrasts with the biases of `SGD` or `Adam` that favor max-margin solutions w.r.t. Euclidean or coordinate-wise norms. Moreover, spectrum-aware optimizers like `Muon` were shown to improve generalization on tasks with imbalanced or long-tailed data distributions (Vasudeva et al., 2025; Wang et al., 2025), as `Muon` (with the aid of spectral orthogonalization) tends to learn

all principal components of the data at a more uniform rate instead of over-emphasizing the dominant features. Further insights were provided by Zhang et al. (2025), who demonstrated statistical benefits of layer-wise preconditioning in simplified settings, and by Wang et al. (2025); Vasudeva et al. (2025), who showed that `Muon` yields a more isotropic singular value spectrum than `Adam`. Moreover, Tveit et al. (2025) reported that `Muon` accelerates grokking, offering further evidence of its practical advantages in long-horizon training dynamics. There have also been discussions drawing connections between `Muon` and other second-order methods—for example, Jordan et al. (2024); Shah et al. (2025) noted that `Muon`'s update can be interpreted as an approximate form of `Shampoo` (Gupta et al., 2018). Lastly, several prior work derived `Muon` and closely related methods from alternative theoretical perspectives, with some of these studies even predating the formal introduction of `Muon` (Pethick et al., 2025; Carlson et al., 2015c; Lau et al., 2025; Bernstein and Newhouse, 2024a,b; An et al., 2025).

Moving beyond `Muon`, it is worth noting that preconditioning has emerged as a powerful tool for accelerating nonconvex matrix factorization. Tong et al. (2021a) introduced `ScaledGD`, with a nonsmooth version presented in Tong et al. (2021b). They proved that in the exactly-parameterized regime with spectral initialization, `ScaledGD` achieves linear convergence at a rate independent of the condition number. Subsequent work by Zhang et al. (2021, 2023) extended these results to the over-parameterized setting, demonstrating condition-number-free convergence when suitably initialized. Xu et al. (2023) showed that `ScaledGD` remains effective under small random initialization, further broadening the scope of condition-number-free guarantees.

1.4 Notation

We also introduce a set of useful notation. For any matrix $\boldsymbol{M}$, we denote by $\sigma_i(\boldsymbol{M})$ the i-th largest singular value of $\boldsymbol{M}$, let $\sigma_{\min}(\boldsymbol{M})$ be its smallest singular value, and we let $\|\boldsymbol{M}\|$ (resp. $\|\boldsymbol{M}\|_{\mathrm{F}}$) represent its spectral norm (resp. Frobenius norm). For any $k \le d$, we let $\mathcal{O}_{d\times k}$ denote the set of orthonormal matrices in $\mathbb{R}^{d\times k}$. For any set of scalars $(a_1, \ldots, a_d)$, we denote by $\mathsf{diag}\{a_1, \ldots, a_d\}$ the diagonal matrix whose diagonal entries are $a_1, \ldots, a_d$. Finally, for any scalar $x \in \mathbb{R}$, we define the sign function as $\mathsf{sign}(x) = 1$ if $x > 0$, $\mathsf{sign}(x) = 0$ if $x = 0$, and $\mathsf{sign}(x) = -1$ if $x < 0$.

2 Main results: two case studies

In this section, we carry out both theoretical and empirical studies on two simple yet fundamental matrix optimization problems. Here and throughout, we shall focus on analyzing simplified `Muon` described in Equation (4), which discards the momentum term and thereby facilitates analysis.

2.1 Matrix factorization

The first problem considered herein is symmetric matrix factorization, which can be formulated as

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2. \tag{5}$$

Here, $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ is a rank-r positive semidefinite matrix, and $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ is a (possibly over-parameterized) factor containing k ($k \ge r$) columns. In a nutshell, we seek to factorize the target matrix $\boldsymbol{M}^\star$ as $\boldsymbol{U}\boldsymbol{U}^\top$ by solving the optimization problem (5). Throughout, we let $\boldsymbol{M}^\star = \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}$ be the eigen-decomposition of $\boldsymbol{M}^\star$, where $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_r^\star\}$ contains the nonzero eigenvalues $\lambda_1^\star \ge \cdots \ge \lambda_r^\star > 0$, and $\boldsymbol{V}^\star \in \mathbb{R}^{d\times r}$ is an orthonormal matrix whose columns correspond to the associated eigenvectors. The condition number of $\boldsymbol{M}^\star$ is defined and denoted by

$$\kappa \coloneqq \lambda_1^\star/\lambda_r^\star.$$

2.1.1 Convergence guarantees for Muon

When applied to the matrix factorization problem (5), the simplified Muon algorithm (4) yields a straightforward closed-form update rule

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \,\mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right), \qquad t = 0, 1, \cdots \tag{6}$$

For both the exactly-parameterized (i.e., $k = r$) and over-parameterized (i.e., $k > r$, or even $k > d$) settings, we establish rapid convergence of simplified Muon to the ground truth, as formalized in the theorem below.

Theorem 1. *Suppose that $\lambda_{\max}^\star \geq \lambda_1^\star \geq \cdots \geq \lambda_r^\star > 0$, and consider any $0 < \varepsilon < \lambda_{\max}^\star$.*

(a) Consider the case with $k \geq d$. Set the learning rates as $\eta_t = C_\eta\sqrt{\lambda_{\max}^\star}\rho^t$ for $1/2 \leq \rho < 1$, with C_η uniformly sampled from the interval $[1, 2]$. Set the initialization as $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$, where $0 < \alpha \leq C_\eta\sqrt{\lambda_{\max}^\star}$ and $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$. Then with probability 1, it holds that $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \leq \varepsilon$ as long as

$$T \geq \frac{1}{1-\rho}\log\left(\frac{8\lambda_{\max}^\star}{\varepsilon}\right). \tag{7}$$

(b) Consider the case with $r \leq k < d$. Set the learning rates as $\eta_t = C_{\eta,t}\sqrt{\lambda_{\max}^\star}\rho^t$ for $2/3 \leq \rho < 1$, with $C_{\eta,t}$ independently and uniformly sampled from the interval $[1, 2]$. Set the initialization $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ for some $\alpha > 0$, where $\boldsymbol{O} \in \mathcal{O}_{d\times k}$ is an orthonormal matrix sampled uniformly at random from $\mathcal{O}_{d\times k}$. Then with probability at least 0.99, we have $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \leq \varepsilon$ as soon as

$$T = \left\lceil \frac{1}{1-\rho}\log\left(\frac{16\lambda_{\max}^\star}{(1-\rho)^2\varepsilon}\right)\right\rceil, \tag{8}$$

provided that α is sufficiently small.

Remark 1. *Careful readers may note that our theory for the regime $r \leq k < d$ requires more restrictive conditions than those in the regime with $k \geq d$. We believe that these restrictions are not fundamental and can potentially be relaxed via more refined analyses, which we leave for future work.*

Remarkably, Theorem 1 uncovers that when applied to matrix factorization, the iteration complexity of Muon is entirely independent from the condition number κ of the target matrix $\boldsymbol{X}^\star$, and scales only logarithmically with the inverse accuracy level $1/\varepsilon$ (thereby establishing linear convergence for Muon). This finding suggests that the gradient orthogonalization step in Muon serves as an effective preconditioner, accelerating convergence by mitigating ill-conditioning in the gradient search directions. Even in the presence of overparameterization, Muon is guaranteed to achieve condition-number-free linear convergence.

We also briefly explain the rationale for using exponentially decaying learning rates. In contrast to GD—where the distance moved in each iteration depends on both the gradient norm and the learning rate—each Muon iteration moves a fixed distance determined solely by the learning rate η_t. Consequently, to achieve linear convergence, the length of each movement—namely, η_t—must decrease geometrically over iterations.

2.1.2 Comparisons with other optimizers

To better demonstrate the preconditioning benefits offered by Muon, we compare its convergence theory established in Theorem 1 against two prominent baselines: gradient descent, and a simplified variant of Adam with momentum turned off. Notably, the latter two optimizers are unable to achieve the desirable condition-number-free convergence rates.

Let us begin by examining GD, whose convergence properties for matrix factorization have been extensively studied. More precisely, consider the following GD update rule:

$$\text{(GD)} \qquad \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right), \qquad t = 0, 1, \cdots \tag{9}$$

The state-of-the-art convergence theory for this algorithm can be summarized as follows: by taking the learning rates $\eta_t = \Theta(1/\lambda_1^\star)$, `GD` yields $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \le \varepsilon$ in

$$\begin{cases} O\big(\kappa \log(1/\varepsilon)\big) \text{ iterations} & \text{if } k = r \text{ (exactly-parameterized)}, \\ O(\min\{\kappa^3 \log(1/\varepsilon),\, \lambda_1^\star/\varepsilon)\} \text{ iterations} & \text{if } k > r \text{ (over-parameterized)}; \end{cases} \tag{10}$$

see, e.g., Chi et al. (2019); Stöger and Soltanolkotabi (2021); Zhuo et al. (2024); Xiong et al. (2023); Xu et al. (2024) for more details. This implies that `GD` cannot attain condition-number-free convergence guarantees without compromising linear convergence.

Next, let us turn attention to a simplified variant of `Adam` given by

$$\texttt{(SignGD)} \qquad \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \,\mathsf{sign}\big((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\big), \qquad t = 0, 1, \cdots \tag{11}$$

where the sign function $\mathsf{sign}(\cdot)$ is applied entrywise. This algorithm (11), which disables momentum in `Adam`, is also referred to as `SignGD`. Note that `SignGD` is more amenable to theoretical analysis than its momentum-based counterpart, while still capturing several core features of `Adam` like entrywise preconditioning (Bernstein and Newhouse, 2024b). To demonstrate the provable advantage of `Muon` over `SignGD`, we establish the following lower bound on the iteration complexity of (11).

Theorem 2. *Let $r_0 \in (0, 1/16]$ be a universal constant. Consider the `SignGD` algorithm (11) with any non-increasing, positive learning rate sequence $\{\eta_t\}_{t\ge 0}$ satisfying $\eta_0 \le r_0$. Then, one can find a ground-truth matrix $\boldsymbol{M}^\star$ with condition number κ, along with an initialization $\boldsymbol{U}_0$ obeying $\big\|\boldsymbol{U}_0\boldsymbol{U}_0^\top - \boldsymbol{M}^\star\big\|_{\mathrm{F}} \le r_0$, such that: for any given $\varepsilon \le \frac{9r_0^2}{4096\kappa^2}$, $f(\boldsymbol{U}_T) \le \varepsilon$ cannot happen unless*

$$T \ge \frac{\kappa - 1}{4}.$$

In words, this lower bound demonstrates that a momentum-free variant of `Adam` may incur at least a linear dependency on κ in the iteration complexity. The proof of this lower bound is deferred to Section C.

2.1.3 Intuition

Thus far, we have established the advantage of simplified `Muon` over a simplified variant of `Adam` (i.e., `SignGD`). In this subsection, we seek to provide some intuitive explanations about their differences in convergence rates. To streamline the presentation, we restrict our discussion to the exactly-parameterized regime where $k = r$.

Decoupling of `Muon` dynamics into independent scalar sequences. To build intuition for the working mechanism of simplified `Muon`, we adopt for the moment a simplifying assumption:

$$\boldsymbol{U}_t = \boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top, \qquad \text{for all } t \ge 1 \tag{12}$$

for some diagonal matrix $\boldsymbol{\Sigma}_t = \mathsf{diag}\{\sigma_{1,t}, \ldots, \sigma_{r,t}\} \in \mathbb{R}^{r\times r}$ and some orthonormal matrix $\boldsymbol{R} \in \mathcal{O}_{d\times r}$. In words, (12) asserts that each `Muon` iterate $\boldsymbol{U}_t$ has its singular subspace perfectly aligned with the true subspace $\boldsymbol{V}^\star$. Although this assumption may appear overly restrictive at first glance, it will be approximately justified in our analysis in Section 3.

Under this simplifying assumption (12), the update rule (6) satisfies

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \,\mathsf{msign}\big((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\big) \\ &= \boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top - \eta_t \,\mathsf{msign}\big((\boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top\boldsymbol{R}\boldsymbol{\Sigma}_t\boldsymbol{V}^{\star\top} - \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top})\boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top\big) \\ &= \boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top - \eta_t \,\mathsf{msign}\big(\boldsymbol{V}^\star(\boldsymbol{\Sigma}_t^3 - \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{\Sigma}_t)\boldsymbol{R}^\top\big) \\ &= \boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top - \eta_t\boldsymbol{V}^\star \,\mathsf{diag\text{-}sign}(\boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star\boldsymbol{\Sigma}_t)\boldsymbol{R}^\top, \end{aligned} \tag{13}$$

where for any diagonal matrix $\boldsymbol{\Sigma} = \{\sigma_1, \ldots, \sigma_r\}$, we define $\mathsf{diag}(\boldsymbol{\Sigma}) = \{\mathsf{sign}(\sigma_1), \ldots, \mathsf{sign}(\sigma_r)\}$. If we write $\boldsymbol{U}_{t+1} = \boldsymbol{V}^\star \boldsymbol{\Sigma}_{t+1} \boldsymbol{R}^\top$ according to (12), then it readily follows from Equation (13) that

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \eta_t \, \mathsf{diag\text{-}sign}(\boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star \boldsymbol{\Sigma}_t). \tag{14}$$

Crucially, all terms in Equation (14) are diagonal, thereby allowing it to be decomposed into r independent scalar recursions:

$$\sigma_{i,t+1} = \sigma_{i,t} - \eta_t \, \mathsf{sign}\left(\sigma_{i,t}^3 - \lambda_i^\star \sigma_{i,t}\right), \qquad t = 0, 1, \cdots \tag{15}$$

for each $1 \le i \le r$, each associated with one eigenvalue of $\boldsymbol{M}^\star$. Noteworthily, the r scalar sequences in (15) evolve completely independently, with no interaction across sequences.

Owing to its simplicity, the scalar recursion in (15) admits a straightforward analysis. As we shall formally establish in Section 3.1, elementary calculations give

$$\left|\sigma_{i,t+1}^2 - \lambda_i^\star\right| = O\big(\sqrt{\lambda_{\max}^\star}\eta_t\big) = O\big(\lambda_{\max}^\star \rho^t\big), \tag{16}$$

provided that the learning rates decay exponentially as $\eta_t = C_\eta \sqrt{\lambda_{\max}^\star}\rho^t$. This linear convergence feature—with the convergence rate ρ a numerical constant within $[1/2, 1)$—mitigates the imbalance between large and small eigenvalues, thereby paving the way for condition-number-free convergence.

This intuition further hints at a connection between `Muon` and the scaled gradient descent (`ScaledGD`) method (Tong et al., 2021a). We formalize this connection and discuss its implications in Appendix A.

Why do `SignGD` and `Adam` fail? As illustrated in Theorem 2, the convergence rate of `SignGD` (a simplified variant of `Adam`) is sensitive to the condition number of $\boldsymbol{M}^\star$. This arises because `SignGD` employs a *per-coordinate preconditioner*, which disregards the richer curvature structure of the problem and hence fails to adapt as effectively as `Muon`.

To see this more formally, denote by $\boldsymbol{u}_t = \mathsf{vec}(\boldsymbol{U}_t)$ the flattened iterate, where $\mathsf{vec}(\boldsymbol{Z})$ stacks the rows of a matrix $\boldsymbol{Z}$ into a single column vector. Invoking the identities $\mathsf{vec}(\boldsymbol{AXB}) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\mathsf{vec}(\boldsymbol{X})$ and $\mathsf{msign}(\boldsymbol{Z}) = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}$ for $\boldsymbol{Z} \in \mathbb{R}^{d \times r}$, we can express the `Muon` update (6) as

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta_t \big(\boldsymbol{I} \otimes (\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t))^{-1/2}\big)\mathsf{vec}\big(\nabla f(\boldsymbol{U}_t)\big), \tag{17}$$

where $\boldsymbol{I} \otimes (\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t))^{-1/2}$ can be interpreted as a blockwise preconditioner. Crucially, this preconditioning matrix is *not* diagonal, even in the limit when $\boldsymbol{U}_t$ converges to the truth.

In contrast, `SignGD` and `Adam` employ diagonal preconditioners. For instance, the `SignGD` update (11) can be expressed as

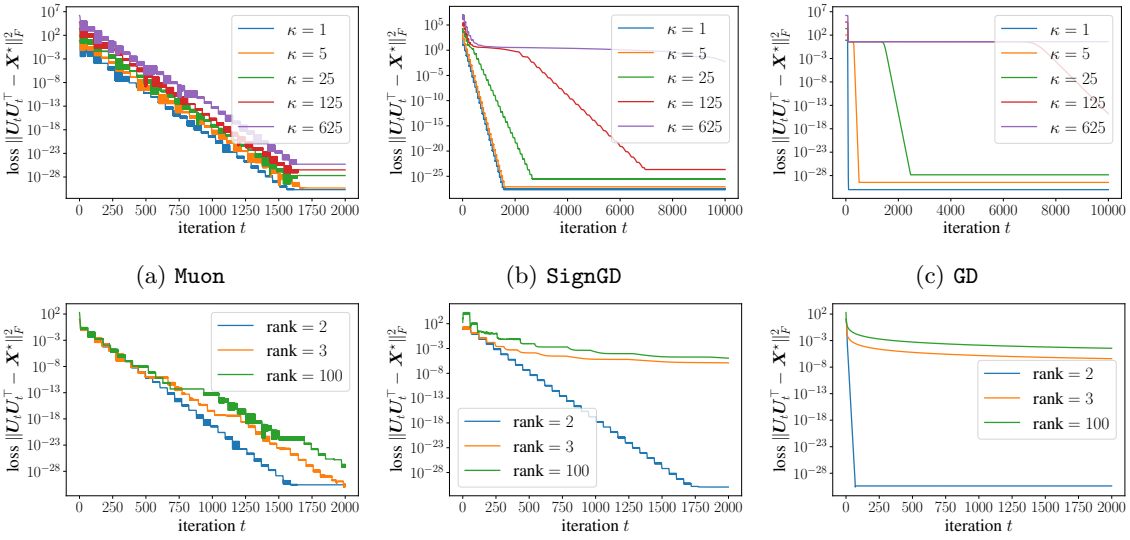
$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta_t \, \mathsf{diag}\left\{|\mathsf{vec}(\nabla f(\boldsymbol{U}_t))|^{-1}\right\} \mathsf{vec}\big(\nabla f(\boldsymbol{U}_t)\big), \tag{18}$$

where $|\boldsymbol{z}|^{-1}$ denotes the entrywise inverse of the entrywise magnitude of a vector $\boldsymbol{z}$. This diagonal preconditioner completely neglects cross-coordinate curvature. Consequently, `Adam` fails to adapt to the geometry of the matrix factorization problem, leading to slow convergence when $\boldsymbol{M}^\star$ is ill-conditioned.

2.1.4 Numerical experiments

We now carry out a series of numerical experiments to validate the theoretical separation in convergence rates between `Muon`, `GD`, and `SignGD`, with results displayed in Figure 1. In the top row (a–c) of Figure 1, we investigate the impact of the condition number $\kappa \in \{1, 5, 25, 125, 625\}$, while fixing the matrix dimension to $d = 100$, target rank $r = 2$, and search rank $k = 2$. In the bottom row (d–f) of Figure 1, we evaluate the effect of search rank $k \in \{2, 3, 100\}$, fixing the condition number $\kappa = 1$, matrix dimension $d = 100$, and target rank $r = 2$. All experiments adopt a more robust exponentially decaying learning rate schedule: the learning rate is reduced by a factor of 0.3 if the loss does not decrease for 50 consecutive iterations.

Across all settings, `Muon` exhibits fast and stable convergence, reaching machine-level precision within a few hundred to a few thousand iterations—even under large condition numbers or severe rank over-specification.



(a) `Muon` (b) `SignGD` (c) `GD`

(d) `Muon` (e) `SignGD` (f) `GD`

Figure 1: Numerical convergence behavior of `Muon`, `SignGD`, and `GD` on matrix factorization tasks under varying condition numbers and search ranks.

In contrast, both `GD` and `SignGD` experience significant slowdowns as the condition number increases or the search rank grows. These results underscore the robustness of `Muon` vis-à-vis ill-conditioning and over-parameterization. Moreover, while our theoretical guarantees for `Muon` require small initialization, we observe that in practice `Muon` converges robustly even with moderately sized initialization. In all experiments, we use an initialization scale of $\alpha = 0.1$. Rigorously elucidating why `Muon` remains stable and convergent under a broader range of initialization is an important direction for future work.

2.2 In-context learning with linear transformers

Next, we turn to the second case study, motivated by in-context learning with linear transformers. Let us first state the optimization problem before describing the motivation. Let $\{\boldsymbol{x}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ be a fixed set of N vectors. Define the empirical covariance matrix as

$$\boldsymbol{S} \coloneqq \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{19}$$

which is assumed to be invertible throughout. We aim to solve the following optimization problem:

$$\underset{\boldsymbol{Q}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad f(\boldsymbol{Q}) \coloneqq \frac{1}{2}\operatorname{tr}\big((\boldsymbol{S}\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{S}(\boldsymbol{S}\boldsymbol{Q}-\boldsymbol{I})^\top\big). \tag{20}$$

This is a simple quadratic optimization problem with $\boldsymbol{Q}^\star = \boldsymbol{S}^{-1}$ the minimizer. Letting $\kappa(\boldsymbol{S})$ denote the condition number of the matrix $\boldsymbol{S}$, we see that the quadratic form induced by (20) has an effective condition number that scales as

$$\kappa \coloneqq \kappa(\boldsymbol{S})^3. \tag{21}$$

Motivation: in-context learning of a single-layer linear transformer. In-context learning (ICL) refers to the phenomenon whereby a pretrained model can make predictions from a *prompt* on the fly (Brown

et al., 2020). More specifically, the prompt contains a sequence of N labeled examples (i.e., the context), followed by a query token, and the model must infer the query label from the context at inference time without updating its parameters. Transformers (Vaswani et al., 2017) arise as a natural model class that supports ICL. Here, we focus on a special case: in-context fixed-design linear regression, where the set of possible input vectors $\{\boldsymbol{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ is fixed with empirical covariance $\boldsymbol{S}$, and each task is indexed by a vector $\boldsymbol{w} \in \mathbb{R}^d$ with corresponding labels $y_{\boldsymbol{w},i} = \boldsymbol{w}^\top \boldsymbol{x}_i$. At a high level, the context can be summarized by the vector $\frac{1}{N}\sum_{i=1}^N y_{\boldsymbol{w},i}\boldsymbol{x}_i = \boldsymbol{S}\boldsymbol{w}$. Given a query $\boldsymbol{x}_{\mathrm{q}} \in \mathbb{R}^d$, a simple in-context predictor uses a shared meta-parameter $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ to map the query to an effective readout $\boldsymbol{Q}\boldsymbol{x}_{\mathrm{q}}$, and predicts via the bilinear form

$$\widehat{y}_{\mathrm{q}} = (\boldsymbol{S}\boldsymbol{w})^\top \boldsymbol{Q}\boldsymbol{x}_{\mathrm{q}} = \boldsymbol{w}^\top \boldsymbol{S}\boldsymbol{Q}\boldsymbol{x}_{\mathrm{q}}.$$

Averaging the squared prediction risk over tasks with $\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] = \boldsymbol{I}$, and over uniformly sampled queries $\boldsymbol{x}_{\mathrm{q}} \sim \mathsf{Unif}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, yields the expected loss that coincides with the objective function in (20). Moreover, this predictor can be realized by a single-layer *linear* transformer (attention without softmax) under a standard reparameterization (Zhang et al., 2024a; Huang et al., 2023). See Section D for more details.

Convergence guarantees for `Muon`. When applied to the optimization problem (20), the update rule of the simplified `Muon` algorithm admits a closed-form expression as follows:

$$\boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t - \eta_t \ \mathsf{msign}\big(\boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2\big), \qquad t = 0, 1, \cdots \tag{22}$$

Encouragingly, this algorithm is guaranteed to converge linearly at a rate independent of κ, as asserted by our theory below.

Theorem 3. *Let the initialization be $\boldsymbol{Q}_0 = \boldsymbol{0}$ and set the learning rate schedule as $\eta_t = \frac{C_\eta}{\sigma_{\min}(\boldsymbol{S})}\rho^t$ for some quantities $C_\eta \geq 1$ and $\rho \in [1/2, 1)$. Then, for any $\varepsilon > 0$, simplified `Muon` (22) achieves $\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\| = \|\boldsymbol{Q}_T - \boldsymbol{S}^{-1}\| \leq \varepsilon$ as long as*

$$T \geq \frac{1}{1-\rho}\log\left(\frac{C_\eta}{\sigma_{\min}(\boldsymbol{S})\varepsilon}\right). \tag{23}$$

This theorem establishes that the number of iterations needed for simplified `Muon` to yield ε-accuracy is independent of the condition number κ underlying this quadratic optimization problem. Akin to the matrix factorization counterpart, the `Muon` dynamics admit a decomposition into a set of independent scalar sequences in the spectral domain, each evolving at a comparable rate of convergence irrespective of the magnitude of the associated eigenvalue, a feature that we shall rigorize in the proof presented in Section 4.

Comparisons with other optimizers. To demonstrate the provable benefits of `Muon` compared against other optimizers, we discuss in this subsection the convergence rate of `GD` and `SignGD`.

When applied to this problem (20), `GD` follows the update rule

$$(\texttt{GD}) \qquad \boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t - \eta_t(\boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2), \qquad t = 0, 1, \cdots \tag{24}$$

Given that this problem is a strongly convex quadratic optimization problem, classical optimization theory already reveals that the number of iterations needed for `GD` to achieve ε-accuracy is lower bounded by (see, e.g., d'Aspremont et al. (2021))

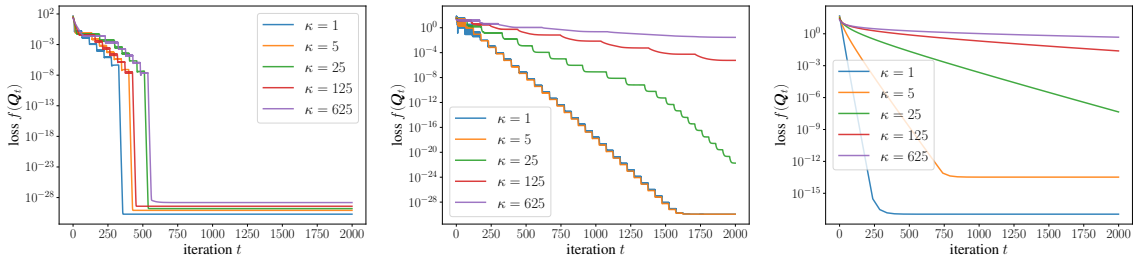
$$\Omega\left(\sqrt{\kappa}\log(1/\varepsilon)\right).$$

This lower bound for `GD` scales proportionally with $\sqrt{\kappa}$, unveiling the unavoidable dependency of its iteration complexity on the condition number.

We then switch attention to `SignGD` (recall that this is a variant of `Adam` with momentum turned off), which adopts the update rule

$$(\texttt{SignGD}) \qquad \boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t - \eta_t \ \mathsf{sign}\big(\boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2\big), \qquad t = 0, 1, \cdots \tag{25}$$

where the $\mathsf{sign}(\cdot)$ operator is applied entrywise.



(a) `Muon` (b) `SignGD` (c) `GD`

Figure 2: Numerical convergence behavior of `Muon`, `SignGD`, and `GD` on in-context learning problems with one-layer linear transformers under varying condition numbers.

Theorem 4. *Consider the `SignGD` algorithm (25) with any non-increasing, positive learning rate schedule* $\{\eta_t\}_{t\geq 0}$. *Consider any* $0 < \varepsilon \leq \sqrt{2}\eta_0/\kappa$. *Then, there exists an empirical covariance matrix* $\boldsymbol{S}$, *along with an initialization* $\boldsymbol{Q}_0$, *such that* $\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\|_{\mathrm{F}} \leq \varepsilon$ *cannot happen unless*

$$T \geq \frac{\kappa - 1}{4}.$$

In words, Theorem 4 rigorously establishes that the `SignGD` algorithm cannot achieve condition-number-free convergence for solving this problem, and is therefore substantially outperformed by `Muon`. The proof of Theorem 4 is provided in Section E.

Numerical experiments. We now evaluate and compare the numerical convergence performance of `Muon`, `SignGD`, and `GD` on in-context learning tasks with one-layer linear transformers. We vary the condition number $\kappa \in \{1, 5, 25, 125, 625\}$ while fixing the matrix dimension to $d = 100$. All experiments use an exponential decay learning rate schedule: the learning rate is reduced by a factor of 0.3 whenever the loss fails to decrease for 50 consecutive iterations. `Muon` achieves rapid convergence across all condition numbers and reaches machine precision within a few hundred steps. In contrast, `SignGD` and `GD` suffer from significantly slower rates, particularly under ill-conditioned settings, thereby validating the robustness and efficiency of `Muon` for ill-conditioned problems.

3 Analysis for matrix factorization (proof of Theorem 1)

In this section, we establish our convergence guarantees for `Muon` applied to matrix factorization (i.e., Theorem 1). Our analysis is structured into several parts. Firstly, we analyze the dynamics of `Muon` for a special scalar case. Secondly, building on this scalar recurrence analysis, we establish the desirable convergence assuming that $\boldsymbol{U}_t$ has its singular subspace perfectly aligned with $\boldsymbol{V}^\star$. With these preparations in place, Steps 3 and 4 then prove the full convergence theory for the cases with $k \geq d$ and $r \leq k < d$, respectively.

3.1 Step 1: dynamics of `Muon` in the scalar case

Before delving into the general case, let us first consider a special case that aims at solving the following scalar optimization problem:

$$\underset{u\in\mathbb{R}}{\text{minimize}} \quad (u^2 - \lambda^\star)^2, \tag{26}$$

where $\lambda^\star \geq 0$. Evidently, this problem can be viewed as a 1-dimensional special case of (5). The `Muon` algorithm (6) applied to (26) follows the scalar dynamic below:

$$u_{t+1} = u_t - \eta_t \,\mathsf{sign}\left((u_t^2 - \lambda^\star)u_t\right), \qquad t = 0, 1, \cdots \tag{27}$$

where $u_0 \in \mathbb{R}$ indicates the initialization.

In order to analyze the dynamics of (27), we first demonstrate in the following lemma that with probability 1, the iterates u_t never reach 0, as long as C_η is randomly generated.

Lemma 1. *Consider any update sequence taking the form of $u_{t+1} = u_t + \eta_t s_t$ for $t \geq 0$, where $u_0 \neq 0$ is the initialization, and $s_t \in \{1, -1\}$ for all $t \geq 0$. The learning rates are taken as $\eta_t = C_\eta \sqrt{\lambda^\star_{\max}} \rho^t$ for some $\lambda^\star_{\max} > 0$ and $\rho \in [1/2, 1)$, where the prefactor C_η is uniformly sampled from the interval $[1, 2]$ and is independent of u_0. Then, with probability 1, one has $u_t \neq 0$ for all $t \geq 0$.*

The fact that $\{u_t\}$ never hits 0 eliminates the need to analyze this undesirable stationary point. We are now positioned to develop theoretical convergence guarantees for the scalar dynamics (27).

Lemma 2 (Convergence of scalar `Muon`)**.** *Consider the scalar updates in (27), where $0 \leq \lambda^\star \leq \lambda^\star_{\max}$. Set the learning rate schedule to be $\eta_t = C_\eta \sqrt{\lambda^\star_{\max}} \rho^t$ for some quantities $1/2 \leq \rho < 1$ and $C_\eta \geq 1$. Assume that $0 < |u_0| \leq C_\eta \sqrt{\lambda^\star_{\max}} = \eta_0$. Then, with probability 1, for all $t \geq 0$, it holds that*

$$\left| |u_{t+1}| - \sqrt{\lambda^\star} \right| \leq \eta_t \leq 2\sqrt{\lambda^\star_{\max}} \rho^t, \tag{28a}$$

$$|u_{t+1}^2 - \lambda^\star| \leq 8\lambda^\star_{\max} \rho^t. \tag{28b}$$

In words, Lemma 2 reveals that `Muon` converges linearly at a rate ρ for this scalar case. Remarkably, analyzing this scalar case not only addresses this special setting, but also sheds light on the spectral dynamics underlying `Muon` for the more general case, as detailed in subsequent subsections.

Proof of Lemma 1. Regarding $t = 0$, we have $u_0 \neq 0$ by assumption. For any $t \geq 1$, we can express u_t by expanding the recurrence relation:

$$u_t = u_0 + \sum_{k=0}^{t-1} s_k \eta_k = u_0 + C_\eta \sqrt{\sigma^\star_{\max}} \left(\sum_{k=0}^{t-1} s_k \rho^k \right) =: u_0 + C_\eta S_t. \tag{29}$$

If $S_t = 0$, we have $u_t = u_0 \neq 0$. Otherwise, the condition $u_t = 0$ is equivalent to $C_\eta = -u_0/S_t$. In other words, for any given t and any fixed sequence $\{s_k\}_{k=0}^{t-1}$, there exists exactly one value of C_η that can make u_t equal 0.

Let $\mathcal{C}$ be the set containing all such critical values for all possible t and $\{s_t\}$:

$$\mathcal{C} = \bigcup_{t=1}^{\infty} \bigcup_{s \in \{-1,1\}^t} \left\{ -\frac{u_0}{\sqrt{\sigma^\star_{\max}} \sum_{k=0}^{t-1} s_k \rho^k} \,\middle|\, \sum_{k=0}^{t-1} s_k \rho^k \neq 0 \right\}, \tag{30}$$

which is clearly a countable set given that the set of time steps and the set of possible sign sequences are both countable. Therefore, when C_η is uniformly sampled from the interval $[1, 2]$, the probability of this continuous random variable taking values in a countable set is 0, i.e.,

$$\mathbb{P}(\exists t \geq 0 : u_t = 0) \leq \mathbb{P}(C_\eta \in \mathcal{C}) = 0. \tag{31}$$

Thus, it follows that, with probability 1, $u_t \neq 0$ holds for all $t \geq 0$. □

Proof of Lemma 2. First, Lemma 1 tells us that with probability 1, $u_t \neq 0$ for all $t \geq 0$. Moreover, if $u_t^2 = \lambda^\star$, then the iterate has reached the optimal solution, and will stay unchanged thereafter. Consequently, it suffices in the sequel to analyze the case where $(u_t^2 - \lambda^\star) u_t \neq 0$.

To proceed, observe that

$$(u_t^2 - \lambda^\star) u_t = (u_t - \sqrt{\lambda^\star})(u_t + \sqrt{\lambda^\star}) u_t. \tag{32}$$

- If $u_t > 0$, then $u_t(u_t + \sqrt{\lambda^\star}) > 0$, and hence

$$\mathsf{sign}\left((u_t^2 - \lambda^\star) u_t \right) = \mathsf{sign}(u_t - \sqrt{\lambda^\star}). \tag{33a}$$

- If $u_t < 0$, then $u_t(u_t - \sqrt{\lambda^\star}) > 0$, and as a result,

$$\mathsf{sign}\left((u_t^2 - \lambda^\star)u_t\right) = \mathsf{sign}(u_t + \sqrt{\lambda^\star}). \tag{33b}$$

This implies that in both of the above cases, the search direction is the sign of the difference between u_t and its nearest root of $\lambda^\star$. In light of this, we find it helpful to define

$$\Delta_t \coloneqq \big||u_t| - \sqrt{\lambda^\star}\big|. \tag{34}$$

Making use of Equations (27) and (33) allows one to easily verify that

$$\Delta_{t+1} = \big|\Delta_t - \eta_t\big| \le \max\{\Delta_t - \eta_t, \eta_t\}. \tag{35}$$

Armed with this inequality, we are ready to prove the claim (28a), which we accomplish by induction.

- *Base case ($t = 0$).* Given that $\sqrt{\lambda^\star} \le \sqrt{\lambda^\star_{\max}}$ and $|u_0| \le \eta_0$, we have

$$\Delta_0 = \big||u_0| - \sqrt{\lambda^\star}\big| \le |u_0| + \sqrt{\lambda^\star} \le \eta_0 + \sqrt{\lambda^\star_{\max}} \le 2\eta_0, \tag{36}$$

  where we have used $\eta_0 = C_\eta \sigma^\star_{\max}$ for $C_\eta \ge 1$. Combining this with (35) at $t = 0$ gives

$$\Delta_1 \le \max\{\Delta_0 - \eta_0, \eta_0\} \le \max\{\eta_0, \eta_0\} = \eta_0, \tag{37}$$

  which establishes the claim (28a) for $t = 0$.

- *Inductive step.* Assume $\Delta_{t+1} \le \eta_t$ for some $t \ge 0$. Then in view of Equation (35),

$$\Delta_{t+2} \le \max\{\Delta_{t+1} - \eta_{t+1}, \eta_{t+1}\} \le \max\{\eta_t - \eta_{t+1}, \eta_{t+1}\}. \tag{38}$$

  Equipped with our assumptions $\eta_{t+1} = \rho\eta_t$ and $1/2 \le \rho < 1$, we obtain

$$\eta_t - \eta_{t+1} = (1-\rho)\eta_t \le \rho\eta_t = \eta_{t+1}, \tag{39}$$

  which taken together with Equation (38) yields

$$\Delta_{t+2} \le \eta_{t+1}.$$

  This establishes the claim (28a) for iteration $t+2$, which in turn finishes the proof of the claim (28a) for all $t \ge 0$ by induction.

Lastly, with inequality (28a) in place, we can readily demonstrate that, for any $t \ge 0$,

$$|u_{t+1}^2 - \lambda^\star| = \big||u_{t+1}| - \sqrt{\lambda^\star}\big|\big(|u_{t+1}| + \sqrt{\lambda^\star}\big) \le \Delta_{t+1}\big(\Delta_{t+1} + 2\sqrt{\lambda^\star}\big) \le 8\lambda^\star_{\max}\rho^t \tag{40}$$

as claimed, where the last inequality holds since $\lambda^\star \le \lambda^\star_{\max}$ and $\Delta_{t+1} \le \eta_t = C_\eta\sqrt{\lambda^\star_{\max}}\rho^t \le 2\sqrt{\lambda^\star_{\max}}$. □

3.2 Step 2: dynamics of `Muon` with perfectly initialized column space

Next, we extend our analysis beyond the scalar case to another special case involving a particular—albeit often impractical—choice of initialization. As will become clear momentarily, the general case is intimately connected to this special setting.

More precisely, suppose that the initialization can be decomposed as

$$\boldsymbol{U}_0 = \boldsymbol{V}^\star \boldsymbol{\Sigma}_0 \boldsymbol{O}_{\mathsf{init}}^\top, \tag{41}$$

where $\boldsymbol{\Sigma}_0 = \mathsf{diag}\{\sigma_{1,0}, \ldots, \sigma_{r,0}\}$ is a diagonal matrix in $\mathbb{R}^{r\times r}$, and $\boldsymbol{O}_{\mathsf{init}} \in \mathbb{R}^{k\times r}$ is some arbitrary orthonormal matrix with $k \ge r$ obeying $\boldsymbol{O}_{\mathsf{init}}^\top \boldsymbol{O}_{\mathsf{init}} = \boldsymbol{I}_r$. Armed with this initialization, we can establish convergence guarantees of `Muon` by extending the scalar analysis in Lemma 2, as formalized in the lemma below.

Lemma 3. *Suppose that $\boldsymbol{U}_0$ satisfies (41). Then for all $t \geq 0$, $\boldsymbol{U}_t$ can be decomposed as*

$$\boldsymbol{U}_t = \boldsymbol{V}^\star \boldsymbol{\Sigma}_t \boldsymbol{O}_{\mathsf{init}}^\top \qquad \text{for some } \boldsymbol{\Sigma}_t = \mathsf{diag}\{\sigma_{1,t}, \ldots, \sigma_{r,t}\} \in \mathbb{R}^{r\times r}. \tag{42a}$$

In particular, for every $t \geq 0$ and $1 \leq i \leq r$, one has

$$\sigma_{i,t+1} = \sigma_{i,t} - \eta_t \,\mathsf{sign}\left((\sigma_{i,t}^2 - \lambda_i^\star)\sigma_{i,t}\right). \tag{42b}$$

Importantly, Lemma 3 reveals that: if the initialization has its left singular subspace perfectly aligned with the desired $\boldsymbol{V}^\star$, then along the entire trajectory, the "spectrum" of each `Muon` iterate decouples into r scalar sequences, each resembling the dynamics analyzed in Lemma 2. Therefore, invoking Lemma 2 yields

$$|\sigma_{i,t+1}^2 - \lambda_i^\star| \leq 8\lambda_{\max}^\star \rho^t \tag{43a}$$

for all $t \geq 0$, with the proviso that $|\sigma_{i,0}| \leq \eta_0$ for all $1 \leq i \leq r$. Taking this collectively with property (42a) leads to the following convergence bound for all $t \geq 0$:

$$\left\|\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top - \boldsymbol{M}^\star\right\| = \left\|\boldsymbol{V}^\star \boldsymbol{\Sigma}_{t+1}^2 \boldsymbol{V}^{\star\top} - \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}\right\| = \max_{1\leq i\leq r} |\sigma_{i,t+1}^2 - \lambda_i^\star| \leq 8\lambda_{\max}^\star \rho^t. \tag{43b}$$

Proof of Lemma 3. Let us prove this lemma by induction.

- *Base case with $t = 0$.* This holds trivially given our assumption (41).
- *Inductive step.* Assuming the induction hypothesis (42a) holds at time t, we can compute the gradient as

$$\nabla f(\boldsymbol{U}_t) = (\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t = \boldsymbol{V}^\star(\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)\boldsymbol{\Sigma}_t\boldsymbol{O}_{\mathsf{init}}^\top. \tag{44}$$

  Given that both $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Lambda}^\star$ are diagonal matrices, the matrix sign of $\nabla f(\boldsymbol{U}_t)$ is given by

$$\mathsf{msign}\left(\nabla f(\boldsymbol{U}_t)\right) = \boldsymbol{V}^\star \,\mathsf{diag\text{-}sign}\left((\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)\boldsymbol{\Sigma}_t\right)\boldsymbol{O}_{\mathsf{init}}^\top. \tag{45}$$

  Here, we recall that $\mathsf{diag\text{-}sign}(\boldsymbol{D}) = \mathsf{diag}\{\mathsf{sign}(D_{1,1}), \ldots, \mathsf{sign}(D_{r,r})\}$ for any diagonal matrix $\boldsymbol{D} = \mathsf{diag}\{D_{1,1}, \ldots, D_{r,r}\}$. As a consequence,

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \,\mathsf{msign}\left(\nabla f(\boldsymbol{U}_t)\right) = \boldsymbol{V}^\star\left(\boldsymbol{\Sigma}_t - \eta_t \,\mathsf{diag\text{-}sign}\left((\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)\boldsymbol{\Sigma}_t\right)\right)\boldsymbol{O}_{\mathsf{init}}^\top. \tag{46}$$

  Thus, this validates the claim (42a) for $t+1$ and demonstrates that

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \eta_t \,\mathsf{diag\text{-}sign}\left((\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)\boldsymbol{\Sigma}_t\right),$$

  as claimed in (42b).

The proof is thus complete by induction. □

3.3 Step 3: analysis for the case with $k \geq d$

Turning to the general case, we begin by analyzing the scenario with $k \geq d$. In this setting, we find it convenient to work with the decomposition $\boldsymbol{M}^\star = \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}$ with $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_d^\star\}$, where we take $r = d$ and allow some of the eigenvalues in $\{\lambda_1^\star, \ldots, \lambda_d^\star\}$ to be zero.

Recall the initialization $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$, where $\boldsymbol{O} \in \mathbb{R}^{d\times k}$ is an arbitrary orthonormal matrix obeying $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$ and $\alpha \leq \eta_0$. One can express $\boldsymbol{U}_0$ alternatively as

$$\boldsymbol{U}_0 = \alpha\boldsymbol{O} = \boldsymbol{V}^\star(\alpha\boldsymbol{I}_d)\boldsymbol{V}^{\star\top}\boldsymbol{O} =: \boldsymbol{V}^\star(\alpha\boldsymbol{I}_d)\boldsymbol{O}_{\mathsf{init}}^\top, \tag{47}$$

where $\boldsymbol{O}_{\mathsf{init}}^\top\boldsymbol{O}_{\mathsf{init}} = \boldsymbol{V}^{\star\top}\boldsymbol{O}\boldsymbol{O}^\top\boldsymbol{V}^\star = \boldsymbol{I}_d$. This indicates that the initialization $\boldsymbol{U}_0$ satisfies Condition (41). Therefore, by applying Lemma 3 and inequality (43b), we see that with probability 1, $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \leq \varepsilon$ holds as long as

$$T > \frac{1}{1-\rho}\log\left(\frac{8\lambda_{\max}^\star}{\varepsilon}\right).$$

3.4 Step 4: analysis for the case with $r \le k < d$

We now switch attention to the case with $r \le k < d$, which is substantially more challenging to analyze than the preceding setting. Here, we shall employ a random orthonormal initialization $\boldsymbol{U}_0 = \alpha \boldsymbol{O}$ obeying $\boldsymbol{O}^\top \boldsymbol{O} = \boldsymbol{I}_k$. Our proof arguments unfold in several steps, as described below.

Step 4.1: initial subspace alignment. A key property that we would like to establish is that: after the first `Muon` iteration, $\boldsymbol{U}_1$ is already well aligned with the eigenspace $\boldsymbol{V}^\star$. Note that when initialized at $\boldsymbol{U}_0 = \alpha \boldsymbol{O}$, the gradient takes the following form

$$\boldsymbol{G}_0 \coloneqq \nabla f(\boldsymbol{U}_0) = (\boldsymbol{U}_0 \boldsymbol{U}_0^\top - \boldsymbol{M}^\star)\boldsymbol{U}_0 = \underbrace{-\alpha \boldsymbol{M}^\star \boldsymbol{O}}_{=:\boldsymbol{Q}} + \alpha^3 \boldsymbol{O}, \tag{48}$$

where $\boldsymbol{Q}$ denotes the leading term for small enough α. Let us decompose $\boldsymbol{G}_0$ into two components as

$$\boldsymbol{G}_0 = \boldsymbol{G}_{0,\le r} + \boldsymbol{G}_{0,>r},$$

where $\boldsymbol{G}_{0,\le r}$ is the best rank-r approximation of $\boldsymbol{G}_0$ (i.e., it is composed of the r leading singular components of $\boldsymbol{G}_0$), and $\boldsymbol{G}_{0,>r}$ consists of the remaining $k-r$ singular components. Given that $\boldsymbol{G}_{0,\le r}$ and $\boldsymbol{G}_{0,>r}$ are orthogonal to each other, the matrix sign of $\boldsymbol{G}_0$ admits the following decomposition:

$$\mathsf{msign}(\boldsymbol{G}_0) = \mathsf{msign}(\boldsymbol{G}_{0,\le r}) + \mathsf{msign}(\boldsymbol{G}_{0,>r}). \tag{49}$$

As it turns out, $\mathsf{msign}(\boldsymbol{G}_{0,\le r})$ and $\mathsf{msign}(\boldsymbol{Q})$ can be fairly close for small enough α, as asserted by the following lemma. The proof is postponed to Section B.1.

Lemma 4. *There exists some universal constant $c_0 > 0$ such that, with probability at least 0.995,*

$$\big\| \mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q}) \big\| \le \frac{16\alpha^2\sqrt{dr}}{c_0 \lambda_r^\star} \tag{50}$$

holds as long as $4\alpha^2 \le c_0 \lambda_r^\star / \sqrt{dr}$.

In addition, given that $\boldsymbol{G}_{0,\le r}$ and $\boldsymbol{G}_{0,>r}$ are orthogonal to each other, Lemma 21 in Section F reveals the existence of a matrix $\widetilde{\boldsymbol{G}}_0 \in \mathbb{R}^{d\times k}$ such that

$$\widetilde{\boldsymbol{G}}_{0,\le r} = \boldsymbol{Q} \qquad \text{and} \tag{51a}$$

$$\big\| \mathsf{msign}(\boldsymbol{G}_0) - \mathsf{msign}(\widetilde{\boldsymbol{G}}_0) \big\| \le \sqrt{2} \big\| \mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q}) \big\| \le \frac{16\sqrt{2}\alpha^2\sqrt{dr}}{c_0\lambda_r^\star}. \tag{51b}$$

Taking this together with the first iteration $\boldsymbol{U}_1 = \alpha\boldsymbol{O} - \eta_0\, \mathsf{msign}(\boldsymbol{G}_0)$ leads to

$$\boldsymbol{U}_1 = \alpha \boldsymbol{O} - \eta_0 \big(\mathsf{msign}(\widetilde{\boldsymbol{G}}_0) + \boldsymbol{R}_0 \big) = -\eta_0\, \mathsf{msign}(\widetilde{\boldsymbol{G}}_0) + \boldsymbol{R}_1, \tag{52a}$$

where the residual terms $\boldsymbol{R}_0, \boldsymbol{R}_1$ satisfy

$$\|\boldsymbol{R}_0\| \le \frac{16\sqrt{2}\alpha^2\sqrt{dr}}{c_0\lambda_r^\star} \qquad \text{and} \qquad \|\boldsymbol{R}_1\| \le \alpha + \frac{16\sqrt{2}\eta_0\alpha^2\sqrt{dr}}{c_0\lambda_r^\star} \le 2\alpha, \tag{52b}$$

provided that $\alpha \le c_0\lambda_r^\star / (32\sqrt{2\lambda_{\max}^\star dr})$.

Step 4.2: construction of an auxiliary trajectory. To facilitate analysis, we find it helpful to construct an auxiliary trajectory $\{\widetilde{\boldsymbol{U}}_t\}_{t\geq 1}$ as follows:

$$\widetilde{\boldsymbol{U}}_1 = -\eta_0\, \mathsf{msign}(\widetilde{\boldsymbol{G}}_0), \tag{53a}$$

$$\widetilde{\boldsymbol{U}}_{t+1} = \widetilde{\boldsymbol{U}}_t - \eta_t\, \mathsf{msign}\left(\nabla f(\widetilde{\boldsymbol{U}}_t)\right), \qquad t = 1, 2, \cdots \tag{53b}$$

In words, this auxiliary trajectory is also generated by simplified `Muon` in (6), but with a slightly modified initialization that discards the residual term $\boldsymbol{R}_1$ appearing in the original iteration (52a). In particular, it follows from (52b) that

$$\left\|\widetilde{\boldsymbol{U}}_1 - \boldsymbol{U}_1\right\| = \|\boldsymbol{R}_1\| \leq 2\alpha. \tag{54}$$

Next, we demonstrate that the dynamics of this auxiliary trajectory $\{\widetilde{\boldsymbol{U}}_t\}_{t\geq 1}$ can be decomposed into a collection of independent scalar dynamics, akin to Step 2. To see this, we first claim that with high probability, $\mathsf{msign}(\boldsymbol{Q})$ can be decomposed as

$$\mathsf{msign}(\boldsymbol{Q}) = \boldsymbol{V}^\star \boldsymbol{O}'^\top \tag{55}$$

for some matrix $\boldsymbol{O}' \in \mathbb{R}^{k\times r}$ obeying $\boldsymbol{O}'^\top \boldsymbol{O}' = \boldsymbol{I}_r$.

Proof of property (55). Observe that

$$-\,\mathsf{msign}(\boldsymbol{Q}) = \mathsf{msign}(\boldsymbol{M}^\star \boldsymbol{O}) = \mathsf{msign}(\boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{B}) = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{B} (\boldsymbol{B}^\top \boldsymbol{\Lambda}^{\star 2} \boldsymbol{B})^{\dagger/2} = \boldsymbol{V}^\star\, \mathsf{msign}(\boldsymbol{\Lambda}^\star \boldsymbol{B}),$$

where we take $\boldsymbol{B} = \boldsymbol{V}^{\star\top}\boldsymbol{O}$. Lemma 10 asserts that with probability at least 0.995, $\sigma_r(\boldsymbol{\Lambda}^\star \boldsymbol{B}) > 0$, thus implying that $\big(\mathsf{msign}(\boldsymbol{\Lambda}^\star \boldsymbol{B})\big)^\top \in \mathcal{O}_{k\times r}$. This completes the proof. □

Armed with this property, we can readily repeat the analysis in Step 2 to establish convergence guarantees for $\{\widetilde{\boldsymbol{U}}_t\}$. It can be easily seen from (55) and our construction of $\widetilde{\boldsymbol{G}}_0$ that: there exist two orthonormal matrices $\boldsymbol{V} \in \mathcal{O}_{d\times k}$ and $\boldsymbol{R} \in \mathcal{O}_{k\times k}$ such that

$$\boldsymbol{V}_{:,1:r} = \boldsymbol{V}^\star, \qquad \boldsymbol{R}_{:,1:r} = \boldsymbol{O}', \qquad \text{and} \qquad \mathsf{msign}(\widetilde{\boldsymbol{G}}_0) = \boldsymbol{V}\boldsymbol{R}^\top, \tag{56}$$

where $\boldsymbol{M}_{:,1:r}$ denotes the first r columns of a matrix $\boldsymbol{M}$. In the meantime, Condition (56) allows us to express the ground truth as

$$\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top} = \boldsymbol{V}\boldsymbol{\Lambda}_{\mathsf{aug}} \boldsymbol{V}^\top, \tag{57}$$

where $\boldsymbol{\Lambda}_{\mathsf{aug}} \in \mathbb{R}^{k\times k}$ is an augmented diagonal matrix $\boldsymbol{\Lambda}_{\mathsf{aug}} = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_k^\star\}$ with $\lambda_{r+1}^\star = \cdots = \lambda_k^\star = 0$.

Recall that $\widetilde{\boldsymbol{U}}_1 = -\eta_0\, \mathsf{msign}(\widetilde{\boldsymbol{G}}_0) = -\eta_0 \boldsymbol{V}\boldsymbol{R}^\top$ (cf. (53a)). Combining this together with Equation (57), we can readily invoke Lemma 3 to show that: for each $t \geq 1$, $\widetilde{\boldsymbol{U}}_t$ can be decomposed as

$$\widetilde{\boldsymbol{U}}_t = \boldsymbol{V}\widetilde{\boldsymbol{\Sigma}}_t \boldsymbol{R}^\top \qquad \text{for some } \boldsymbol{\Sigma}_t = \mathsf{diag}\{\widetilde{\sigma}_{1,t}, \ldots, \widetilde{\sigma}_{k,t}\} \in \mathbb{R}^{k\times k}, \tag{58a}$$

where $|\widetilde{\sigma}_{i,1}| = \eta_0$ $(1 \leq i \leq k)$, and for every $t \geq 1$ and $1 \leq i \leq k$,

$$\widetilde{\sigma}_{i,t+1} = \widetilde{\sigma}_{i,t} - \eta_t\, \mathsf{sign}\left((\widetilde{\sigma}_{i,t}^2 - \lambda_i^\star)\widetilde{\sigma}_{i,t}\right). \tag{58b}$$

Repeating the same convergence analysis for (43) tells us that (see Lemma 11 for a slight extension that accounts for random learning rates): for every $t \geq 0$ and every $1 \leq i \leq k$ one has

$$|\widetilde{\sigma}_{i,t+1}^2 - \lambda_i^\star| \leq \frac{8}{(1-\rho)^2}\lambda_{\max}^\star \rho^t, \tag{59a}$$

$$\left\|\widetilde{\boldsymbol{U}}_{t+1}\widetilde{\boldsymbol{U}}_{t+1}^\top - \boldsymbol{M}^\star\right\| \leq \frac{8}{(1-\rho)^2}\lambda_{\max}^\star \rho^t. \tag{59b}$$

Consequently, one achieves

$$\left\|\widetilde{\boldsymbol{U}}_T\widetilde{\boldsymbol{U}}_T^\top - \boldsymbol{M}^\star\right\| \leq \varepsilon/2 \tag{60}$$

as long as $T > \frac{1}{1-\rho}\log\left(\frac{16\lambda_{\max}^\star}{(1-\rho)^2\varepsilon}\right)$.

Step 4.3: proximity between the original and auxiliary trajectories. With the desirable convergence property of $\{\widetilde{\boldsymbol{U}}_t\}$ in place, it remains to show that the original iterates $\{\boldsymbol{U}_t\}$ remain close to the auxiliary iterates. First, we bound the differences between $\boldsymbol{U}_t$ and $\widetilde{\boldsymbol{U}}_t$, as well as between their associated gradients, in the following lemma; the proof is deferred to Section B.2.

Lemma 5. *Assume that* $\sigma_{\min}(\nabla f(\boldsymbol{U}_t)), \sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t)) > 0$. *Then it holds that*

$$\left\|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\right\| \le \left(1 + \eta_t \frac{147\lambda_{\max}^\star}{(1-\rho)^2 \sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\right) \left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t\right\|. \tag{61}$$

In addition, we have $\max\left\{\|\boldsymbol{U}_t\|, \|\widetilde{\boldsymbol{U}}_t\|\right\} \le \frac{4\sqrt{\lambda_{\max}^\star}}{1-\rho}$.

Repeating the analysis of Lemma 1 (which we omit here for brevity), we can easily see that with probability one, $\sigma_{\min}(\nabla f(\boldsymbol{U}_t)) > 0$ and $\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t)) > 0$ hold for all $t \ge 1$. Lemma 5 then tells us that

$$\begin{aligned}
\left\|\boldsymbol{U}_T - \widetilde{\boldsymbol{U}}_T\right\| &\le \prod_{t=1}^{T-1} \left(1 + \eta_t \frac{147\lambda_{\max}^\star}{(1-\rho)^2 \sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\right) \left\|\boldsymbol{U}_1 - \widetilde{\boldsymbol{U}}_1\right\| \\
&\le \underbrace{\left\{\prod_{t=1}^{T-1} \left(1 + \frac{294\lambda_{\max}^{\star 3/2}\rho^t}{(1-\rho)^2 \sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\right)\right\}}_{=:\,\Pi_T} \left\|\boldsymbol{U}_1 - \widetilde{\boldsymbol{U}}_1\right\| \\
&\le 2\alpha\Pi_T,
\end{aligned} \tag{62}$$

where we have used $\eta_t = C_{\eta,t}\sqrt{\lambda_{\max}^\star}\rho^t \le 2\sqrt{\lambda_{\max}^\star}\rho^t$ as well as (54).

In order to invoke (62) to control $\left\|\boldsymbol{U}_T - \widetilde{\boldsymbol{U}}_T\right\|$ and Π_T, a crucial step is to lower bound $\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t)) = \min_{1\le i\le k} |(\widetilde{\sigma}_{i,t}^2 - \lambda_i^\star)\widetilde{\sigma}_{i,t}|$, as accomplished by the following lemma. See Section B.3 for the proof.

Lemma 6. *Consider any* $0 < \varepsilon \le \frac{1}{4}(\lambda_{\min}^\star)^{3/2}$. *Then for every step* $t \ge 1$, *we have*

$$\mathbb{P}\left(\sigma_{\min}\left(\nabla f(\widetilde{\boldsymbol{U}}_{t+1})\right) \le \varepsilon \mid \mathcal{F}_t\right) \le \frac{2(k-r)\sqrt[3]{\varepsilon}}{\sqrt{\lambda_{\max}^\star}\rho^t} + \frac{12r\varepsilon}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}\rho^t}, \tag{63}$$

where $\mathcal{F}_t$ *represents all events that happen up to and including time* t.

One can then exploit Lemma 6 to establish high-probability upper bounds on the quantity Π_T defined in (62). The resulting bounds are stated in the lemma below, whose proof is deferred to Section B.4.

Lemma 7. *Consider any* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$, *the following results hold.*

(i) If $k = r$, *then*

$$\Pi_T \le \exp\left(O\left(T\log\left(\frac{r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right)\right). \tag{64}$$

(ii) If $k > r$, *then*

$$\Pi_T \le \exp\left(O\left(T^2 + T\log\left(\frac{(k-r)r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right)\right). \tag{65}$$

In particular, if $T = \left\lceil \frac{1}{1-\rho}\log\left(\frac{16\lambda_{\max}^\star}{(1-\rho)^2\varepsilon}\right)\right\rceil$ *and* $\delta = \mathsf{poly}(\varepsilon/\lambda_{\max}^\star)$, *then with probability at least* $1-\delta$ *one has*

$$\Pi_T \le \left(\frac{\lambda_{\max}^\star}{\varepsilon}\right)^{\zeta_{\exp}} \text{ with } \zeta_{\exp} = \begin{cases} O\left(\frac{1}{1-\rho}\log\left(\frac{\lambda_{\max}^\star}{(1-\rho)\varepsilon}\right)\log\left(\frac{r\kappa}{1-\rho}\right)\right), & \text{if } k = r. \\ O\left(\frac{1}{(1-\rho)^2}\log^2\left(\frac{\lambda_{\max}^\star}{(1-\rho)\varepsilon}\right) + \frac{1}{1-\rho}\log\left(\frac{\lambda_{\max}^\star}{(1-\rho)\varepsilon}\right)\log\left(\frac{(k-r)r\kappa}{1-\rho}\right)\right), & \text{if } k > r. \end{cases} \tag{66}$$

Taking this lemma together with (62) yields: with probability at least $1 - 0.001(\varepsilon/\lambda^\star_{\max})$, we have

$$\left\|\boldsymbol{U}_T - \widetilde{\boldsymbol{U}}_T\right\| \le 2\alpha\left(\frac{\lambda^\star_{\max}}{\varepsilon}\right)^{\zeta_{\exp}} \le \frac{(1-\rho)\varepsilon}{16\sqrt{\lambda^\star_{\max}}}. \tag{67}$$

provided that

$$\alpha \le \frac{(1-\rho)\varepsilon}{32\sqrt{\lambda^\star_{\max}}}\left(\frac{\varepsilon}{\lambda^\star_{\max}}\right)^{\zeta_{\exp}}. \tag{68}$$

Step 4.4: putting everything together. Invoking (59b), (67) and Lemma 5, and applying the union bound, we conclude that with probability at least 0.99,

$$\begin{aligned}\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| &\le \left\|\widetilde{\boldsymbol{U}}_T\widetilde{\boldsymbol{U}}_T^\top - \boldsymbol{M}^\star\right\| + \left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \widetilde{\boldsymbol{U}}_T\widetilde{\boldsymbol{U}}_T^\top\right\| \\ &\le \left\|\widetilde{\boldsymbol{U}}_T\widetilde{\boldsymbol{U}}_T^\top - \boldsymbol{M}^\star\right\| + \left(\|\boldsymbol{U}_T\| + \left\|\widetilde{\boldsymbol{U}}_T\right\|\right)\left\|\boldsymbol{U}_T - \widetilde{\boldsymbol{U}}_T\right\| \\ &\le \frac{\varepsilon}{2} + 2\cdot\frac{4\sqrt{\lambda^\star_{\max}}}{1-\rho}\cdot\frac{(1-\rho)\varepsilon}{16\sqrt{\lambda^\star_{\max}}} = \varepsilon,\end{aligned} \tag{69}$$

provided that α is sufficiently small. This completes the proof.

4 Analysis for linear transformers (proof of Theorem 3)

Recall that the gradient of $f(\boldsymbol{Q})$ w.r.t. $\boldsymbol{Q}$ is given by

$$\nabla f(\boldsymbol{Q}) = \boldsymbol{S}(\boldsymbol{S}\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{S} = \boldsymbol{S}^2\boldsymbol{Q}\boldsymbol{S} - \boldsymbol{S}^2. \tag{70}$$

To proceed, let us denote the eigen-decomposition of $\boldsymbol{S}$ as $\boldsymbol{S} = \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}$, where $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_d^\star\}$ is a diagonal matrix containing the eigenvalues $\{\lambda_i^\star\}$ of $\boldsymbol{S}$, and $\boldsymbol{V}^\star$ consists of orthonormal columns corresponding to the eigenvectors of $\boldsymbol{S}$. As a key step of this proof, we would like to show that:

Lemma 8. *For each $t \ge 0$, the simplified* `Muon` *iterates (22) can be decomposed as*

$$\boldsymbol{Q}_t = \boldsymbol{V}^\star\boldsymbol{\Theta}_t\boldsymbol{V}^{\star\top} \tag{71a}$$

for some diagonal matrix $\boldsymbol{\Theta}_t = \mathsf{diag}\{\theta_{1,t}, \ldots, \theta_{d,t}\}$. In particular, $\{\boldsymbol{\Theta}_t\}$ evolves according to

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \eta_t\,\mathsf{diag\text{-}sign}(\boldsymbol{\Lambda}^\star\boldsymbol{\Theta}_t - \boldsymbol{I}), \qquad t = 0, 1, \cdots \tag{71b}$$

where $\mathsf{diag\text{-}sign}(\boldsymbol{M}) \coloneqq \mathsf{diag}\{\mathsf{sign}(M_{1,1}), \ldots, \mathsf{sign}(M_{d,d})\}$ for any diagonal matrix $\boldsymbol{M} = \mathsf{diag}\{M_{1,1}, \ldots, M_{d,d}\}$.

Proof of Lemma 8. The base case with $t = 0$ holds trivially, since the initialization $\boldsymbol{Q}_0 = \boldsymbol{0}$ is equivalent to taking $\boldsymbol{\Lambda}_0 = \boldsymbol{0}$. Assuming the inductive hypothesis (71a) holds at step t, we have

$$\nabla f(\boldsymbol{Q}_t) = \boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2 = \boldsymbol{V}^\star\big(\boldsymbol{\Lambda}^{\star 3}\boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2}\big)\boldsymbol{V}^{\star\top},$$

and as a result,

$$\begin{aligned}\boldsymbol{Q}_{t+1} &= \boldsymbol{Q}_t - \eta_t\,\mathsf{msign}\left(\nabla L(\boldsymbol{Q}_t)\right) \\ &= \boldsymbol{Q}_t - \eta_t\,\mathsf{msign}\left(\boldsymbol{V}^\star(\boldsymbol{\Lambda}^{\star 3}\boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2})\boldsymbol{V}^{\star\top}\right) \\ &= \boldsymbol{Q}_t - \eta_t\boldsymbol{V}^\star\,\mathsf{diag\text{-}sign}(\boldsymbol{\Lambda}^{\star 3}\boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2})\boldsymbol{V}^{\star\top} \\ &= \boldsymbol{V}^\star\big(\boldsymbol{\Theta}_t - \eta_t\,\mathsf{diag\text{-}sign}(\boldsymbol{\Lambda}^{\star 3}\boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2})\big)\boldsymbol{V}^{\star\top} \\ &= \boldsymbol{V}^\star\big(\boldsymbol{\Theta}_t - \eta_t\,\mathsf{diag\text{-}sign}(\boldsymbol{\Lambda}^\star\boldsymbol{\Theta}_t - \boldsymbol{I})\big)\boldsymbol{V}^{\star\top}.\end{aligned} \tag{72}$$

This implies the decomposition $\boldsymbol{Q}_{t+1} = \boldsymbol{V}^\star\boldsymbol{\Theta}_{t+1}\boldsymbol{V}^{\star\top}$, where the diagonal matrix $\boldsymbol{\Theta}_{t+1}$ can be computed as $\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \eta_t\,\mathsf{sign}(\boldsymbol{\Lambda}^\star\boldsymbol{\Theta}_t - \boldsymbol{I})$. The proof is thus complete by induction. □

Importantly, Lemma 8 indicates that the `Muon` dynamics can be decomposed into a collection of scalar sequences obeying

$$\theta_{i,t+1} = \theta_{i,t} - \eta_t \,\mathsf{sign}(\lambda_i^\star \theta_{i,t} - 1), \qquad t = 0, 1, \cdots \tag{73}$$

for each $1 \le i \le d$. As it turns out, the convergence rate of each scalar sequence $\{\theta_{i,t}\}_{t\ge 0}$ can be analyzed through the following lemma.

Lemma 9. *Consider a scalar sequence $\{\theta_t\}_{t\ge 0} \subset \mathbb{R}$ obeying*

$$\theta_{t+1} = \theta_t - \eta_t \,\mathsf{sign}(\lambda^\star \theta_t - 1),$$

where the scalar $\lambda^\star$ satisfies $\lambda^\star \ge \lambda^\star_{\min} > 0$. Set the learning rate schedule to be $\eta_t = \frac{C_\eta}{\lambda^\star_{\min}}\rho^t$ for some quantities $1/2 \le \rho < 1$ and $C_\eta \ge 1$. With the initialization $\theta_0 = 0$, one has

$$\left|\theta_{t+1} - \frac{1}{\lambda^\star}\right| \le \eta_t = \frac{C_\eta}{\lambda^\star_{\min}}\rho^t \qquad \text{for all } t \ge 0. \tag{74}$$

To finish up, applying Lemma 9 to each scalar sequence $\{\theta_{i,t}\}_{t\ge 0}$, we arrive at

$$\left\|\boldsymbol{Q}_{t+1} - \boldsymbol{S}^{-1}\right\| = \left\|\boldsymbol{\Theta}_{t+1} - (\boldsymbol{\Lambda}^\star)^{-1}\right\| = \max_{1\le i\le d}\left|\theta_{i,t+1} - \frac{1}{\lambda_i^\star}\right| \le \eta_t. \tag{75}$$

Thus, in order to ensure $\left\|\boldsymbol{Q}_T - \boldsymbol{S}^{-1}\right\| \le \varepsilon$, it suffices to take $T \ge \frac{1}{1-\rho}\log\left(\frac{C_\eta}{\sigma_{\min}(\boldsymbol{S})\varepsilon}\right)$.

Proof of Lemma 9. The proof is analogous to the proof of Lemma 2. Define the metric

$$\Delta_t \coloneqq \left|\theta_t - \frac{1}{\lambda^\star}\right|. \tag{76}$$

To bound Δ_t, we first observe that

$$\begin{aligned}\theta_{t+1} - \frac{1}{\lambda^\star} &= \theta_t - \frac{1}{\lambda^\star} - \eta_t \,\mathsf{sign}(\lambda^\star\theta_t - 1) \\ &= \theta_t - \frac{1}{\lambda^\star} - \eta_t \,\mathsf{sign}\left(\theta_t - \frac{1}{\lambda^\star}\right) = \mathsf{sign}\left(\theta_t - \frac{1}{\lambda^\star}\right)\left(\left|\theta_t - \frac{1}{\lambda^\star}\right| - \eta_t\right).\end{aligned} \tag{77}$$

If $\theta_t = 1/\lambda^\star$, then it is readily seen from (77) and $\mathsf{sign}(0) = 0$ that $\Delta_{t+1} = 0 \le \eta_t$. If instead $\theta_t \ne 1/\lambda^\star$, then it follows from (77) that

$$\Delta_{t+1} = \left|\Delta_t - \eta_t\right| \le \max\{\Delta_t - \eta_t, \eta_t\}. \tag{78}$$

In summary, this inequality (78) holds for both cases, which coincides with the bound (35) in the proof of Lemma 2.

When $t = 0$, it holds that

$$\Delta_1 = |\Delta_0 - \eta_0| = \left|\frac{1}{\lambda^\star} - \frac{C_\eta}{\lambda^\star_{\min}}\right| \le \frac{C_\eta}{\lambda^\star_{\min}} = \eta_0.$$

Then, repeating the same arguments as in the proof of Lemma 2, we conclude that

$$\left|\theta_{t+1} - \frac{1}{\lambda^\star}\right| = \Delta_{t+1} \le \eta_t = \frac{C_\eta}{\lambda^\star_{\min}}\rho^t$$

as claimed. □

5 Discussion

In this paper, we have rigorously characterized the preconditioning benefits of `Muon` for two matrix optimization problems: matrix factorization, and in-context learning of linear transformers. Our theory implies that `Muon`'s spectral orthogonalization acts as a form of adaptive preconditioners, effectively transforming its dynamics into independent scalar sequences in the spectral domain, each converging at a comparable rate. Both theoretical analyses and empirical studies suggest that `Muon` yields better-conditioned optimization trajectories, achieving faster convergence than `GD` and `Adam`. We anticipate that this preconditioning mechanism plays a key role in accelerating various matrix-structured optimization problems, and that it may inform the design of new spectrum-aware optimization algorithms.

As noted previously, our theoretical analysis is limited to two simple problems. This naturally opens up various avenues for future research. We conclude by highlighting two important directions.

- *Extension to other matrix-structured problems.* Given the limited scope of our analysis to two problems, a natural next step is to investigate whether the preconditioning effect of `Muon` generalizes to other matrix-structured tasks. In addition to other nonconvex matrix factorization problems described in Chi et al. (2019), one potential example is the matrix linear regression problem given by

  $$\text{minimize}_{\boldsymbol{W}\in\mathbb{R}^{m\times n}} \|\boldsymbol{W}\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2,$$

  which generalizes classical linear regression to a matrix setting. This problem not only serves as a useful testbed for theoretical analysis, but also captures the training dynamics of linear layers in neural networks. Recent papers have begun to explore this space: Davis and Drusvyatskiy (2025) derived a criterion under which `Muon` outperforms `GD` in a single step, while Das et al. (2024) investigated the preconditioning effect of `Adam` in the vector case. Extending these insights to broader matrix-valued problems could illuminate how `Muon` interacts with layer-wise structures and whether spectrum-aware optimizers yield more efficient or stable training.

- *Toward a general theory.* Another important direction is to develop a unified theoretical framework that elucidates the preconditioning and acceleration effects of `Muon` under broad, practically relevant conditions, such as gradient Lipschitz continuity. While recent research has made progress in this direction (Davis and Drusvyatskiy, 2025; Su, 2025; Shen et al., 2025), existing analyses remain limited in several key aspects: some rely on idealized models, others impose intricate per-iteration conditions whose validity has yet to be rigorously established, and many fall short of explaining the observed empirical advantage of `Muon` over classical optimizers. Overcoming these limitations will require deeper insight into both the geometry of the loss landscape—especially in transformer architectures—and the way in which `Muon`'s updates dynamically reshape the optimization trajectories. It would also be of great interest to investigate whether important structural properties arising in neural network training, such as block Hessians (Zhang et al., 2024b), can be efficiently exploited by `Muon`.

Acknowledgments

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the ONR grant N00014-25-1-2344, the NSF grants 2221009 and 2218773, the Wharton AI & Analytics Initiative's AI Research Fund, and the Amazon Research Award. Y. Chi is supported in part by NSF under grant ECCS-2537078 and AFOSR under grant FA9550-25-1-0060. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650.

An, K., Liu, Y., Pan, R., Ren, Y., Ma, S., Goldfarb, D., and Zhang, T. (2025). ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*.

Bernstein, J. and Newhouse, L. (2024a). Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*.

Bernstein, J. and Newhouse, L. (2024b). Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carlson, D., Cevher, V., and Carin, L. (2015a). Stochastic spectral descent for restricted Boltzmann machines. In *Artificial intelligence and statistics*, pages 111–119. PMLR.

Carlson, D., Hsieh, Y.-P., Collins, E., Carin, L., and Cevher, V. (2015b). Stochastic spectral descent for discrete graphical models. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):296–311.

Carlson, D. E., Collins, E., Hsieh, Y.-P., Carin, L., and Cevher, V. (2015c). Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28.

Chen, L., Li, J., and Liu, Q. (2025). Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

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

Das, R., Agarwal, N., Sanghavi, S., and Dhillon, I. S. (2024). Towards quantifying the preconditioning effect of Adam. *arXiv preprint arXiv:2402.07114*.

Davis, D. and Drusvyatskiy, D. (2025). When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*.

d'Aspremont, A., Scieur, D., and Taylor, A. (2021). Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

Fan, C., Schmidt, M., and Thrampoulidis, C. (2025). Implicit bias of spectral descent and Muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Gupta, V., Koren, T., and Singer, Y. (2018). Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Huang, Y., Cheng, Y., and Liang, Y. (2023). In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*.

Huang, Y., Wen, Z., Chi, Y., and Liang, Y. (2025). A theoretical analysis of self-supervised learning for vision transformers. In *The Thirteenth International Conference on Learning Representations*.

Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., and Bernstein, J. (2024). Muon: An optimizer for hidden layers in neural networks.

Kingma, D. P. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kovalev, D. (2025). Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*.

Lau, T. T.-K., Long, Q., and Su, W. (2025). PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*.

Li, J. and Hong, M. (2025). A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*.

Li, R.-C. (1995). New perturbation bounds for the unitary polar factor. *SIAM Journal on Matrix Analysis and Applications*, 16(1):327–332.

Liu, J., Su, J., Yao, X., Jiang, Z., Lai, G., Du, Y., Qin, Y., Xu, W., Lu, E., Yan, J., et al. (2025). Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pethick, T., Xie, W., Antonakopoulos, K., Zhu, Z., Silveti-Falls, A., and Cevher, V. (2025). Training deep learning models with norm-constrained LMOs. *arXiv preprint arXiv:2502.07529*.

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010*, pages 1576–1602. World Scientific.

Shah, I., Polloreno, A. M., Stratos, K., Monk, P., Chaluvaraju, A., Hojel, A., Ma, A., Thomas, A., Tanwer, A., and Shah, D. J. (2025). Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*.

Shen, W., Huang, R., Huang, M., Shen, C., and Zhang, J. (2025). On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*.

Stöger, D. and Soltanolkotabi, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843.

Su, W. (2025). Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*.

Tong, T., Ma, C., and Chi, Y. (2021a). Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63.

Tong, T., Ma, C., and Chi, Y. (2021b). Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409.

Tuddenham, M., Prügel-Bennett, A., and Hare, J. (2022). Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*.

Tveit, A., Remseth, B., and Skogvold, A. (2025). Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*.

Vasudeva, B., Deora, P., Zhao, Y., Sharan, V., and Thrampoulidis, C. (2025). How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, S., Zhang, F., Li, J., Du, C., Du, C., Pang, T., Yang, Z., Hong, M., and Tan, V. Y. (2025). Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Xiong, N., Ding, L., and Du, S. S. (2023). How over-parameterization slows down gradient descent in matrix sensing: The curses of symmetry and initialization. *arXiv preprint arXiv:2310.01769*.

Xu, X., Shen, Y., Chi, Y., and Ma, C. (2023). The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pages 38611–38654. PMLR.

Xu, Z., Wang, Y., Zhao, T., Ward, R., and Tao, M. (2024). Provable acceleration of Nesterov's accelerated gradient for rectangular matrix factorization and linear neural networks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 33726–33755.

Yang, T., Huang, Y., Liang, Y., and Chi, Y. (2024). In-context learning with representations: Contextual generalization of trained transformers. *Advances in Neural Information Processing Systems*, 37:85867–85898.

Zhang, G., Fattahi, S., and Zhang, R. Y. (2023). Preconditioned gradient descent for overparameterized nonconvex Burer-Monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55.

Zhang, J., Fattahi, S., and Zhang, R. Y. (2021). Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996.

Zhang, R., Frei, S., and Bartlett, P. L. (2024a). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Zhang, T. T., Moniri, B., Nagwekar, A., Rahman, F., Xue, A., Hassani, H., and Matni, N. (2025). On the concurrence of layer-wise preconditioning methods and provable feature learning. *arXiv preprint arXiv:2502.01763*.

Zhang, Y., Chen, C., Ding, T., Li, Z., Sun, R., and Luo, Z. (2024b). Why transformers need Adam: A Hessian perspective. *Advances in neural information processing systems*, 37:131786–131823.

Zhuo, J., Kwon, J., Ho, N., and Caramanis, C. (2024). On the computational and statistical complexity of over-parameterized matrix sensing. *Journal of Machine Learning Research*, 25(169):1–47.

Contents

A Connection between Muon and ScaledGD for matrix factorization

A provably efficient preconditioned optimizer for matrix factorization is ScaledGD (Tong et al., 2021a), which also achieves convergence rates independent of the condition number. As it turns out, there are some inherent connections between Muon and ScaledGD. More concretely, the update rule of simplified Muon yields

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \nabla f(\boldsymbol{U}_t)\big(\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t)\big)^{-1/2} = \boldsymbol{U}_t - \eta_t \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{\Delta}_t^2 \boldsymbol{U}_t\big)^{-1/2}, \tag{79}$$

where $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star$. In comparison, the update rule of ScaledGD is given by

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \beta_t \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{U}_t\big)^{-1} \tag{80}$$

for some learning rate $\beta_t > 0$. In other words, Muon constructs its preconditioner from the gradient, whereas ScaledGD builds its preconditioner from the iterate itself. In the idealistic case where $\boldsymbol{\Delta}_t^2 \approx c_t \boldsymbol{U}_t\boldsymbol{U}_t^\top$ for some scalar $c_t > 0$, (79) can be simplified as

$$\boldsymbol{U}_{t+1} \approx \boldsymbol{U}_t - \eta_t c_t \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{U}_t\big)^{-1}, \tag{81}$$

which coincides with the ScaledGD update (80) up to proper scaling of the learning rate.

In general, the condition $\boldsymbol{\Delta}_t^2 \approx c_t \boldsymbol{U}_t\boldsymbol{U}_t^\top$ cannot possibly hold, but it offers some useful insight in the local regime $\boldsymbol{U}_t\boldsymbol{U}_t \approx \boldsymbol{M}^\star$. Adopting once again the simplifying assumption (12), we derive

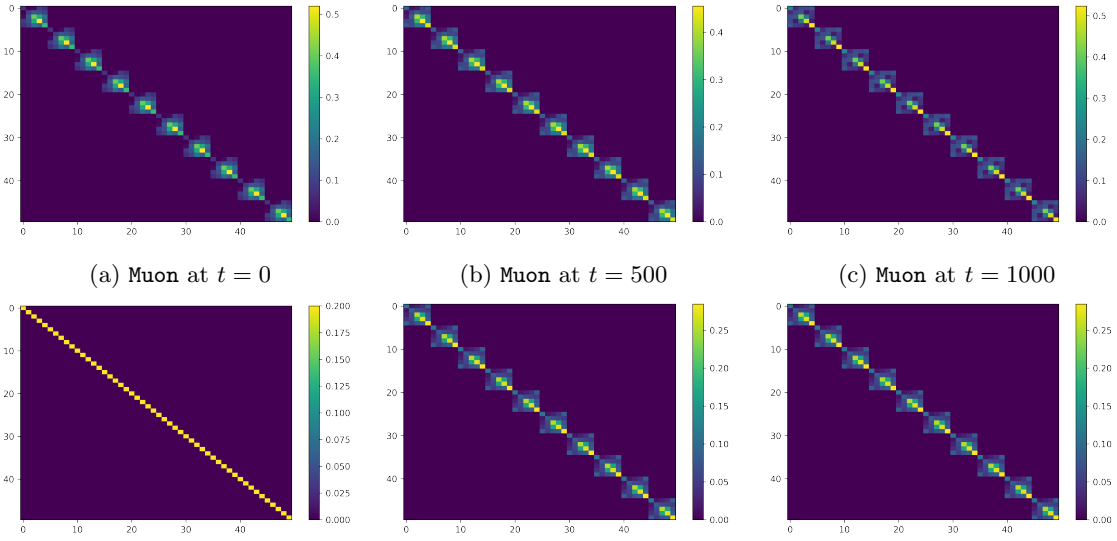
$$\boldsymbol{\Delta}_t = \boldsymbol{V}^\star\left(\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star\right)\boldsymbol{V}^{\star\top} \quad \text{and} \quad \boldsymbol{U}_t\boldsymbol{U}_t^\top = \boldsymbol{V}^\star\boldsymbol{\Sigma}_t^2\boldsymbol{V}^{\star\top} \approx \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}. \tag{82}$$

To ensure $\boldsymbol{\Delta}_t^2 \approx \boldsymbol{U}_t\boldsymbol{U}_t^\top$, one needs to show that $(\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)^2 \approx c_t\boldsymbol{\Lambda}^\star$, or equivalently,

$$(\sigma_{i,t}^2 - \lambda_i^\star)^2 \approx c_t\lambda_i^\star, \qquad 1 \le i \le r.$$

Given that $\sigma_{i,t}^2 - \lambda_i^\star = \big(\sigma_{i,t} - \sqrt{\lambda_i^\star}\big)\big(\sigma_{i,t} + \sqrt{\lambda_i^\star}\big) \approx 2\lambda_i^\star\big(\sigma_{i,t} - \sqrt{\lambda_i^\star}\big)$, this condition is equivalent to

$$4\big(\sigma_{i,t} - \sqrt{\lambda_i^\star}\big)^2 \approx c_t, \qquad 1 \le i \le r. \tag{83}$$



(a) `Muon` at $t = 0$ (b) `Muon` at $t = 500$ (c) `Muon` at $t = 1000$

(d) `ScaledGD` at $t = 0$ (e) `ScaledGD` at $t = 500$ (f) `ScaledGD` at $t = 1000$

Figure 3: Numerical comparison of the preconditioners of `Muon` and `ScaledGD` for matrix factorization at various training steps along a `Muon` trajectory.

To justify the approximate feasibility of (83), observe that at each iteration, the scalar sequence $\{\sigma_{i,t}\}$ in (15) moves by a fixed length (i.e., either η_t or $-\eta_t$) irrespective of the gradient size. In the local region where $\sigma_{i,t} \approx \lambda_i^\star$, the scalar sequence is expected to follow a zigzag trajectory oscillating around $\lambda_i^\star$. Under random initialization, one may thus anticipate $\mathbb{E}[|\sigma_{i,t} - \sqrt{\lambda_i^\star}|] \propto \eta_t$, a scale that is independent of the magnitude of $\lambda_i^\star$. This intuition suggests that in the local region, the `Muon` update may be approximated by `ScaledGD`. Note, however, that these arguments are heuristic in nature; a fully rigorous analysis of their connections is left for future work.

To further understand the connection between `Muon` and `ScaledGD`, we conduct experiments to visualize and compare their corresponding preconditioners over the course of a `Muon` trajectory, as shown in Figure 3. We consider a matrix factorization task with dimension $d = 10$ and target and search ranks $r = k = 5$, initialized with a small scale $\alpha = 10^{-10}$. We also adopt the same learning rate schedule as in previous experiments. At each step t, the `Muon` preconditioner is defined as $\boldsymbol{H}_{\texttt{Muon},t} = \boldsymbol{I} \otimes \left(\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t)\right)^{1/2}$, while the `ScaledGD` preconditioner takes the form $\boldsymbol{H}_{\texttt{ScaledGD},t} = \boldsymbol{I} \otimes \left(\boldsymbol{U}_t^\top \boldsymbol{U}_t\right)$. Throughout training, both preconditioners display a consistent block-diagonal pattern—highlighting their structural similarity and revealing the implicit connection between the two methods. Importantly, the non-diagonal structure of these preconditioners also hints at why methods using diagonal preconditioners, such as `Adam`, are not well-suited for this setting.

B Proof of auxiliary lemmas for matrix factorization

B.1 Proof of Lemma 4

The difference between $\mathsf{msign}(\boldsymbol{G}_{0,\le r})$ and $\mathsf{msign}(\boldsymbol{Q})$ can be bounded by

$$\begin{aligned}\left\|\mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q})\right\| &\overset{\text{(i)}}{\le} \frac{2}{\min\{\sigma_r(\boldsymbol{Q}), \sigma_r(\boldsymbol{G}_{0,\le r})\}} \left\|\boldsymbol{G}_{0,\le r} - \boldsymbol{Q}\right\| \\ &\overset{\text{(ii)}}{\le} \frac{2}{\sigma_r(\boldsymbol{Q}) - \alpha^3} \left\|\boldsymbol{G}_{0,\le r} - \boldsymbol{Q}\right\|\end{aligned}$$

$$\overset{\text{(iii)}}{\leq} \frac{6+\frac{2(\sigma_{r+1}(\boldsymbol{Q})+\sigma_{r+1}(\boldsymbol{G}_0))}{\min\{\sigma_r(\boldsymbol{Q}),\sigma_r(\boldsymbol{G}_0)\}-\max\{\sigma_{r+1}(\boldsymbol{Q}),\sigma_{r+1}(\boldsymbol{G}_0)\}}}{\sigma_r(\boldsymbol{Q})-\alpha^3}\|\boldsymbol{G}_0-\boldsymbol{Q}\|$$

$$\overset{\text{(iv)}}{\leq} \frac{6+\frac{2\alpha^3}{\sigma_r(\boldsymbol{Q})-2\alpha^3}}{\sigma_r(\boldsymbol{Q})-\alpha^3}\left\|\alpha^3\boldsymbol{O}\right\|$$

$$\overset{\text{(v)}}{=} \frac{6\alpha^3+\frac{2\alpha^6}{\sigma_r(\boldsymbol{Q})-2\alpha^3}}{\sigma_r(\boldsymbol{Q})-\alpha^3} \overset{\text{(vi)}}{\leq} \frac{16\alpha^3}{\sigma_r(\boldsymbol{Q})}, \tag{84}$$

provided that $\sigma_r(\boldsymbol{Q}) > 4\alpha^3$. Here, (i) follows from Lemma 17; (ii) is valid since, by Weyl's inequality,

$$\sigma_r(\boldsymbol{G}_{0,\leq r}) = \sigma_r(\boldsymbol{G}_0) \geq \sigma_r(\boldsymbol{Q}) - \|\alpha^3\boldsymbol{O}\| = \sigma_r(\boldsymbol{Q}) - \alpha^3; \tag{85}$$

(iii) applies Lemma 18; (iv) results from Equation (85), the fact $\sigma_{r+1}(\boldsymbol{Q}) = 0$, as well as the following property (by Weyl's inequality):

$$\sigma_{r+1}(\boldsymbol{G}_0) \leq \sigma_{r+1}(\boldsymbol{Q}) + \|\alpha^3\boldsymbol{O}\| = \|\alpha^3\boldsymbol{O}\| = \alpha^3;$$

(v) follows since $\boldsymbol{O}$ is orthonormal; and (vi) holds as long as $\sigma_r(\boldsymbol{Q}) > 4\alpha^3$.

To continue upper bounding (84), we develop a lower bound on $\sigma_r(\boldsymbol{Q})$ in the lemma below, whose proof is provided in Section B.5.

Lemma 10. *There exists some universal constant $c_0 > 0$ such that, with probability at least 0.995,*

$$\sigma_r(\boldsymbol{M}^\star\boldsymbol{O}) \geq \frac{c_0\lambda_r^\star}{\sqrt{dr}}. \tag{86}$$

Lemma 10 taken together with inequality (85) tells us that, with probability exceeding 0.995,

$$\sigma_r(\boldsymbol{Q}) = \alpha\sigma_r(\boldsymbol{M}^\star\boldsymbol{O}) \geq \frac{c_0\alpha\lambda_r^\star}{\sqrt{dr}}. \tag{87}$$

Therefore, if $4\alpha^2 \leq c_0\lambda_r^\star/\sqrt{dr}$, then we establish that

$$\left\|\mathsf{msign}(\boldsymbol{G}_{0,\leq r}) - \mathsf{msign}(\boldsymbol{Q})\right\| \leq \frac{16\alpha^3}{\sigma_r(\boldsymbol{Q})} \leq \frac{16\alpha^2\sqrt{dr}}{c_0\lambda_r^\star}. \tag{88}$$

B.2 Proof of Lemma 5

To begin with, the update rule (6) allows us to upper bound the size of $\boldsymbol{U}_t$ as

$$\|\boldsymbol{U}_t\| = \left\|\boldsymbol{U}_0 - \sum_{s=0}^{t-1}\eta_s\,\mathsf{msign}\left(\nabla f(\boldsymbol{U}_s)\right)\right\| \leq \|\boldsymbol{U}_0\| + \sum_{s=0}^{t-1}\eta_s \leq \alpha + \frac{2\sqrt{\lambda_{\max}^\star}}{1-\rho} \leq \frac{4\sqrt{\lambda_{\max}^\star}}{1-\rho}, \tag{89a}$$

provided that $\alpha \leq 2\sqrt{\lambda_{\max}^\star}/(1-\rho)$. The same argument applies to $\widetilde{\boldsymbol{U}}_t$, yielding

$$\left\|\widetilde{\boldsymbol{U}}_t\right\| \leq \frac{4\sqrt{\lambda_{\max}^\star}}{1-\rho}. \tag{89b}$$

Then, it follows from (6) and our construction (53) that

$$\begin{aligned}\left\|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\right\| &\leq \left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t\right\| + \eta_t\left\|\mathsf{msign}\left(\nabla f(\boldsymbol{U}_t)\right) - \mathsf{msign}\left(\nabla f(\widetilde{\boldsymbol{U}}_t)\right)\right\| \\ &\overset{\text{Lemma 17}}{\leq} \left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t\right\| + \eta_t\frac{3\left\|\nabla f(\boldsymbol{U}_t) - \nabla f(\widetilde{\boldsymbol{U}}_t)\right\|}{\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\end{aligned}$$

$$\leq \left(1+\eta_t \frac{147\lambda_{\max}^\star}{(1-\rho)^2\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\right)\|\boldsymbol{U}_t-\widetilde{\boldsymbol{U}}_t\|, \tag{90}$$

where the second line relies on our assumption that $\sigma_{\min}(\nabla f(\boldsymbol{U}_t)), \sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t)) > 0$, and the last inequality invokes the triangle inequality and (89) to obtain

$$\begin{aligned}
\|\nabla f(\boldsymbol{U}_t)-\nabla f(\widetilde{\boldsymbol{U}}_t)\| &= \|(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{M}^\star)\boldsymbol{U}_t-(\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top-\boldsymbol{M}^\star)\widetilde{\boldsymbol{U}}_t\| \\
&\leq \|\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top-\boldsymbol{M}^\star\|\|\boldsymbol{U}_t-\widetilde{\boldsymbol{U}}_t\|+\|\boldsymbol{U}_t\|\|\boldsymbol{U}_t\boldsymbol{U}_t^\top-\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top\| \\
&\leq \left(\|\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top\|+\|\boldsymbol{M}^\star\|+2\max\left\{\|\boldsymbol{U}_t\|^2,\|\widetilde{\boldsymbol{U}}_t\|^2\right\}\right)\|\boldsymbol{U}_t-\widetilde{\boldsymbol{U}}_t\| \\
&\leq \frac{49\lambda_{\max}^\star}{(1-\rho)^2}\|\boldsymbol{U}_t-\widetilde{\boldsymbol{U}}_t\|.
\end{aligned}$$

B.3 Proof of Lemma 6

Define the following quantity

$$g_{i,t} \coloneqq \left|(\widetilde{\sigma}_{i,t}^2-\lambda_i^\star)\widetilde{\sigma}_{i,t}\right|.$$

We would like to first control $g_{i,t+1}$ for a single i, followed by a union bound to cover all indices $\{1,\cdots,k\}$.

Consider any fix i, and recall from (58) that the update rule for $\widetilde{\sigma}_{i,t+1}$ is

$$\widetilde{\sigma}_{i,t+1} = \widetilde{\sigma}_{i,t}-\eta_t\,\mathsf{sign}\left((\widetilde{\sigma}_{i,t}^2-\lambda_i^\star)\widetilde{\sigma}_{i,t}\right), \tag{91}$$

where η_t is uniform sampled from $[\sqrt{\lambda_{\max}^\star}\rho^t, 2\sqrt{\lambda_{\max}^\star}\rho^t]$. Thus, conditional on past randomness, $\widetilde{\sigma}_{i,t+1}$ is uniformly distributed over $\left[\widetilde{\sigma}_{i,t}-2s_{i,t}\sqrt{\lambda_{\max}^\star}\rho^t, \widetilde{\sigma}_{i,t}-s_{i,t}\sqrt{\lambda_{\max}^\star}\rho^t\right]$, where $s_{i,t}=\mathsf{sign}\left((\widetilde{\sigma}_{i,t}^2-\lambda_i^\star)\widetilde{\sigma}_{i,t}\right)$; in other words, $\widetilde{\sigma}_{i,t+1}$ is uniformly sampled from an interval of length $\sqrt{\lambda_{\max}^\star}\rho^t$. We now divide into two cases based on whether $\lambda_i^\star=0$ or $\lambda_i^\star>0$.

Case 1: $\lambda_i^\star=0$. In this case, one has $g_{i,t+1}=|\widetilde{\sigma}_{i,t+1}|^3$, which implies that

$$\mathbb{P}\left(g_{i,t+1}\leq\varepsilon\mid\mathcal{F}_t\right)=\mathbb{P}\left(|\widetilde{\sigma}_{i,t+1}|\leq\sqrt[3]{\varepsilon}\mid\mathcal{F}_t\right)\leq\frac{2\sqrt[3]{\varepsilon}}{\sqrt{\lambda_{\max}^\star}\rho^t}. \tag{92}$$

Case 2: $\lambda_i^\star>0$. For this case, we first claim that for any $0<\varepsilon\leq\frac{1}{4}(\lambda_i^\star)^{3/2}$, it holds that

$$g_{i,t}\leq\varepsilon \quad\Longrightarrow\quad |\widetilde{\sigma}_{i,t}|\in\left[0,\frac{2\varepsilon}{\lambda_i^\star}\right]\cup\left[\sqrt{\lambda_i^\star}-\frac{2\varepsilon}{\lambda_i^\star},\sqrt{\lambda_i^\star}+\frac{2\varepsilon}{\lambda_i^\star}\right]. \tag{93}$$

Without loss of generality, assume that $\widetilde{\sigma}_{i,t}\geq 0$. To justifies this property (93), consider two sub-cases:

- If $\widetilde{\sigma}_{i,t}\geq\sqrt{\lambda_i^\star}+\frac{\varepsilon}{\lambda_i^\star}$, then we have

$$g_{i,t}=(\widetilde{\sigma}_{i,t}^2-\lambda_i^\star)\widetilde{\sigma}_{i,t}\geq\frac{2\varepsilon}{\sqrt{\lambda_i^\star}}\cdot\sqrt{\lambda_i^\star}\geq 2\varepsilon>\varepsilon. \tag{94}$$

- If $\frac{2\varepsilon}{\lambda_i^\star}<\widetilde{\sigma}_{i,t}<\sqrt{\lambda_i^\star}-\frac{2\varepsilon}{\lambda_i^\star}$, then it follows that

$$g_{i,t}=(\sqrt{\lambda_i^\star}+\widetilde{\sigma}_{i,t})(\sqrt{\lambda_i^\star}-\widetilde{\sigma}_{i,t})\widetilde{\sigma}_{i,t}\geq\sqrt{\lambda_i^\star}(\sqrt{\lambda_i^\star}-\widetilde{\sigma}_{i,t})\widetilde{\sigma}_{i,t}>\sqrt{\lambda_i^\star}\frac{2\varepsilon}{\lambda_i^\star}\left(\sqrt{\lambda_i^\star}-\frac{2\varepsilon}{\lambda_i^\star}\right)\geq\varepsilon, \tag{95}$$

  provided that $0<\varepsilon\leq\frac{1}{4}(\lambda_i^\star)^{3/2}$.

Combining the above two subcases, one can easily see that

$$\mathbb{P}\left(g_{i,t+1} \le \varepsilon \mid \mathcal{F}_t\right) \le \mathbb{P}\left(|\widetilde{\sigma}_{i,t+1}| \in \left[0, \frac{2\varepsilon}{\lambda_i^\star}\right] \cup \left[\sqrt{\lambda_i^\star} - \frac{2\varepsilon}{\lambda_i^\star}, \sqrt{\lambda_i^\star} + \frac{2\varepsilon}{\lambda_i^\star}\right] \mid \mathcal{F}_t\right) \le \frac{12\varepsilon/\lambda_i^\star}{\sqrt{\lambda_{\max}^\star}\rho^t} \le \frac{12\varepsilon}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}\rho^t}.$$

To finish up, apply the union bound over all indices $i \in \{1, \ldots, k\}$ to arrive at

$$\begin{aligned}\mathbb{P}\left(\sigma_{\min}\big(\nabla f(\widetilde{\boldsymbol{U}}_{t+1})\big) \le \varepsilon\right) &\le \sum_{i:\,\lambda_i^\star = 0} \mathbb{P}(g_{i,t+1} \le \varepsilon) + \sum_{i:\,\lambda_i^\star > 0} \mathbb{P}(g_{i,t+1} \le \varepsilon) \\ &\le \frac{2(k-r)\sqrt[3]{\varepsilon}}{\sqrt{\lambda_{\max}^\star}\rho^t} + \frac{12r\varepsilon}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}\rho^t}.\end{aligned}$$

B.4 Proof of Lemma 7

Recall that $\mathcal{F}_t$ encompasses what happens up to time t, and hence $\widetilde{\boldsymbol{U}}_t$ is fully determined by $\mathcal{F}_{t-1}$. Define

$$C_t \coloneqq \frac{294\lambda_{\max}^{\star 3/2}\rho^t}{(1-\rho)^2}, \qquad X_t = \log\left(1 + \frac{C_t}{\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))}\right), \qquad S_T = \sum_{t=1}^{T-1} X_t, \tag{96}$$

which allows us to write $\Pi_T = e^{S_T}$. In the sequel, we intend to control S_T by invoking the Chernoff-type arguments and bounding (conditional) moment generating functions (MGFs).

Step 1: a general connection between MGF and tail bounds. For any nonnegative random variable Z and any $\theta > 0$, the MGF obeys

$$\mathbb{E}\left[e^{\theta Z}\right] = 1 + \int_0^\infty \theta e^{\theta\tau}\mathbb{P}(Z \ge \tau)\mathrm{d}\tau. \tag{97}$$

This follows from integration by parts, namely, $\mathbb{E}[e^{\theta Z}] = \int_0^\infty e^{\theta z}\mathrm{d}F(z) = 1 + \int_0^\infty \theta e^{\theta\tau}\mathbb{P}(Z \ge \tau)\mathrm{d}\tau$.

Step 2: a conditional tail bound on X_t. For any $\tau \ge 0$, the definition of X_t indicates that

$$\{X_t \ge \tau\} \quad \Longleftrightarrow \quad \left\{\sigma_{\min}\big(\nabla f(\widetilde{\boldsymbol{U}}_t)\big) \le \frac{C_t}{e^\tau - 1}\right\}. \tag{98}$$

With this equivalence in mind, applying Lemma 6 to $\sigma_{\min}(\nabla f(\widetilde{\boldsymbol{U}}_t))$ reveals that, for all $\tau \ge 0$,

$$\mathbb{P}(X_t \ge \tau \mid \mathcal{F}_{t-1}) \le \min\left\{\frac{2(k-r)\sqrt[3]{C_t}}{\sqrt{\lambda_{\max}^\star}\rho^{t-1}\sqrt[3]{e^\tau - 1}} + \frac{12rC_t}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}\rho^{t-1}(e^\tau - 1)}, 1\right\}. \tag{99}$$

Step 3: an MGF bound for the case with $k = r$. When $k = r$, the first term in (99) vanishes. Define

$$A \coloneqq \frac{12rC_t}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}\rho^{t-1}} = \frac{12r}{\lambda_{\min}^\star\sqrt{\lambda_{\max}^\star}} \cdot \frac{294\lambda_{\max}^{\star 3/2}\rho^t}{(1-\rho)^2} \cdot \frac{1}{\rho^{t-1}} = \frac{3528r\kappa\rho}{(1-\rho)^2} \ge 1, \tag{100}$$

which is independent of t. Then, it follows from (99) that

$$\mathbb{P}(X_t \ge \tau \mid \mathcal{F}_{t-1}) \le \min\left\{1, \frac{A}{e^\tau - 1}\right\}. \tag{101}$$

Let $\tau_0 \coloneqq \log(1 + A)$, which satisfies $1 = \frac{A}{e^{\tau_0} - 1}$. For $\tau \ge \tau_0$, it is seen that $e^\tau - 1 \ge e^\tau/2$ due to the fact that $\tau_0 = \log(1 + A) \ge \log(2)$. Hence, it holds that, for all $\tau \ge \tau_0$,

$$\mathbb{P}(X_t \ge \tau \mid \mathcal{F}_{t-1}) \le \frac{A}{e^\tau - 1} \le 2Ae^{-\tau}. \tag{102}$$

Next, substitute this tail bound into (97) to show that: for any $\theta \in (0,1)$,

$$\begin{aligned}
\mathbb{E}\left[e^{\theta X_t} \mid \mathcal{F}_{t-1}\right] &= 1 + \int_0^{\tau_0} \theta e^{\theta\tau}\, \mathbb{P}(X_t \geq \tau \mid \mathcal{F}_{t-1}) \mathrm{d}\tau + \int_{\tau_0}^{\infty} \theta e^{\theta\tau}\, \mathbb{P}(X_t \geq \tau \mid \mathcal{F}_{t-1}) \mathrm{d}\tau \\
&\leq 1 + \int_0^{\tau_0} \theta e^{\theta\tau} \mathrm{d}\tau + \int_{\tau_0}^{\infty} \theta e^{\theta\tau} \cdot 2Ae^{-\tau} \mathrm{d}\tau \\
&= 1 + (e^{\theta\tau_0} - 1) + 2A\theta \int_{\tau_0}^{\infty} e^{-(1-\theta)\tau} \mathrm{d}\tau \\
&= e^{\theta\tau_0} + \frac{2A\theta}{1-\theta}\, e^{-(1-\theta)\tau_0}.
\end{aligned} \tag{103}$$

Recall the identity $e^{\tau_0} = 1 + A$, we have

$$e^{\theta\tau_0} = (1+A)^{\theta}, \qquad \frac{2A\theta}{1-\theta}\, e^{-(1-\theta)\tau_0} = \frac{2\theta}{1-\theta}\, A(1+A)^{-(1-\theta)} \leq \frac{2\theta}{1-\theta}\, (1+A)^{\theta}, \tag{104}$$

since $A(1+A)^{-(1-\theta)} \leq (1+A)^{\theta}$. As a result, we arrive at

$$\mathbb{E}\left[e^{\theta X_t} \mid \mathcal{F}_{t-1}\right] \leq \left(1 + \frac{2\theta}{1-\theta}\right)(1+A)^{\theta} \leq 3(1+A)^{\theta} \tag{105}$$

for any $\theta \in (0, 1/2]$.

Step 4: Chernoff bound on S_T when $k = r$. For any $\theta \in (0, 1/2]$, apply (105) recursively to obtain

$$\begin{aligned}
\mathbb{E}\left[e^{\theta S_T}\right] &= \mathbb{E}\left[\prod_{t=1}^{T-1} e^{\theta X_t}\right] = \mathbb{E}\left[\prod_{t=1}^{T-2} e^{\theta X_t} \cdot \mathbb{E}\left[e^{\theta X_{T-1}} \mid \mathcal{F}_{T-2}\right]\right] \\
&\leq 3(1+A)^{\theta} \cdot \mathbb{E}\left[\prod_{t=1}^{T-2} e^{\theta X_t}\right] \leq \cdots \\
&\leq \left(3(1+A)^{\theta}\right)^{T-1}.
\end{aligned} \tag{106}$$

Markov's inequality yields, for any $u > 0$,

$$\mathbb{P}(S_T \geq u) \leq e^{-\theta u}\, \mathbb{E}\left[e^{\theta S_T}\right] \leq \exp\big(-\theta u + (T-1)\log(3) + \theta(T-1)\log(1+A)\big).$$

Choosing

$$u = (T-1)\log(1+A) + \frac{T-1}{\theta}\log(3) + \frac{1}{\theta}\log\frac{1}{\delta} \tag{107}$$

then gives $\mathbb{P}(S_T \geq u) \leq \delta$. Taking $\theta = 1/2$, we obtain that with probability at least $1-\delta$,

$$S_T \leq (T-1)\log(1+A)2(T-1)\log(3) + 2\log\frac{1}{\delta} = O\left(T\log\left(\frac{r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right). \tag{108}$$

Exponentiating both sides yields

$$\Pi_T = \exp(S_T) \leq \exp\left(O\left(T\log\left(\frac{r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right)\right), \tag{109}$$

thereby completing the proof of Part (i) of Lemma 7.

Step 5: an MGF bound for the case with $k > r$. When $k > r$, define

$$A_{1,t} \coloneqq \frac{2(k-r)\sqrt[3]{C_t}}{\sqrt{\lambda^\star_{\max}}\rho^{t-1}} = \frac{2(k-r)}{\rho^{\frac{2t-3}{3}}}, \qquad A_2 \coloneqq \frac{12rC_t}{\lambda^\star_{\min}\sqrt{\lambda^\star_{\max}}\rho^{t-1}} = \frac{3528r\kappa\rho}{(1-\rho)^2}. \tag{110}$$

Then it is readily seen from (99) that

$$\mathbb{P}(X_t \geq \tau \mid \mathcal{F}_{t-1}) \leq \min\left\{1, \frac{A_{1,t}}{(e^\tau - 1)^{1/3}} + \frac{A_2}{e^\tau - 1}\right\}. \tag{111}$$

Let $\tau_{1,t} \coloneqq \log(1 + A_{1,t}^3)$ and $\tau_2 \coloneqq \log(1 + A_2)$, and set $\tau_0 \coloneqq \max\{\tau_{1,t}, \tau_2\}$. For $\tau \geq \tau_0$, we have

$$\mathbb{P}(X_t \geq \tau \mid \mathcal{F}_{t-1}) \leq 2^{1/3}A_{1,t}e^{-\tau/3} + 2A_2e^{-\tau}, \tag{112}$$

given that $e^\tau - 1 \geq e^\tau/2$. Invoke (97) to show that, for any $\theta \in (0, 1/3)$,

$$\begin{aligned}
\mathbb{E}[e^{\theta X_t} \mid \mathcal{F}_{t-1}] &\leq 1 + \int_0^{\tau_0} \theta e^{\theta\tau}\mathrm{d}\tau + \int_{\tau_0}^{\infty} \theta e^{\theta\tau}\big(2^{1/3}A_{1,t}e^{-\tau/3} + 2A_2e^{-\tau}\big)\mathrm{d}\tau \\
&= e^{\theta\tau_0} + 2^{1/3}A_{1,t}\theta\int_{\tau_0}^{\infty} e^{-(1/3-\theta)\tau}\mathrm{d}\tau + 2A_2\theta\int_{\tau_0}^{\infty} e^{-(1-\theta)\tau}\mathrm{d}\tau \\
&= e^{\theta\tau_0} + \frac{2^{1/3}A_{1,t}\theta}{1/3-\theta}\, e^{-(1/3-\theta)\tau_0} + \frac{2A_2\theta}{1-\theta}\, e^{-(1-\theta)\tau_0}.
\end{aligned} \tag{113}$$

Now, recognizing that $\tau_{1,t} = \log(1 + A_{1,t}^3)$ and $\tau_2 = \log(1 + A_2)$, we can further derive

$$e^{\theta\tau_0} \leq e^{\theta\tau_{1,t} + \theta\tau_2} \leq (1 + A_{1,t}^3)^\theta(1 + A_2)^\theta, \tag{114}$$

and also (since $\tau_0 \geq \tau_{1,t}$ and $\tau_0 \geq \tau_2$)

$$A_{1,t}e^{-(1/3-\theta)\tau_0} \leq A_{1,t}(1 + A_{1,t}^3)^{-(1/3-\theta)} \leq (1 + A_{1,t}^3)^\theta, \qquad A_2e^{-(1-\theta)\tau_0} \leq (1 + A_2)^\theta. \tag{115}$$

Substitution into (113) reveals that, for any given $\theta \in (0, 1/6]$,

$$\mathbb{E}[e^{\theta X_t} \mid \mathcal{F}_{t-1}] \leq C_1(1 + A_{1,t}^3)^\theta\,(1 + A_2)^\theta, \tag{116}$$

where C_1 is a constant given by $C_1 = 1 + \frac{2^{1/3}\theta}{1/3-\theta} + \frac{2\theta}{1-\theta}$.

Step 6: Chernoff bound on S_T when $k > r$. Iterating conditional expectations as before and invoking (116), we arrive at

$$\begin{aligned}
\mathbb{E}[e^{\theta S_T}] &\leq \prod_{t=1}^{T-1}\left(C_1(1 + A_2)^\theta(1 + A_{1,t}^3)^\theta\right) \\
&= \exp\left((T-1)\log(C_1) + \theta(T-1)\log(1 + A_2) + \theta\sum_{t=1}^{T-1}\log(1 + A_{1,t}^3)\right).
\end{aligned} \tag{117}$$

Akin to Step 4, Markov's inequality then yields

$$\mathbb{P}(S_T \geq u) \leq \exp\left(-\theta u + (T-1)\log(C_1) + \theta(T-1)\log(1 + A_2) + \theta\sum_{t=1}^{T-1}\log(1 + A_{1,t}^3)\right).$$

Clearly, choosing

$$u = (T-1)\log(1 + A_2) + \sum_{t=1}^{T-1}\log(1 + A_{1,t}^3) + \frac{T-1}{\theta}\log(C_1) + \frac{1}{\theta}\log\frac{1}{\delta}. \tag{118}$$

yields $\mathbb{P}(S_T \geq u) \leq \delta$. Taking $\theta = 1/6$ above and recognizing the facts that

$$\sum_{t=1}^{T-1} \log(1 + A_{1,t}^3) = \sum_{t=1}^{T-1} \log\left(1 + \frac{8(k-r)^3}{\rho^{2t-3}}\right) = O\left(T^2 \log\frac{1}{\rho} + T\log(k-r)\right),$$
$$\log(1 + A_2) = O\bigg(\log\Big(\frac{r\kappa}{1-\rho}\Big)\bigg),$$

we can use $\rho \geq 2/3$ to demonstrate that

$$S_T \leq O\left(T^2 + T\log\left(\frac{(k-r)r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right),$$

with probability at least $1-\delta$, and as a consequence,

$$\Pi_T = \exp(S_T) \leq \exp\left(O\left(T^2 + T\log\left(\frac{(k-r)r\kappa}{1-\rho}\right) + \log\frac{1}{\delta}\right)\right). \tag{119}$$

This establishes Part (ii) of Lemma 7.

B.5 Proof of Lemma 10

Recalling that $\boldsymbol{M}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}$, we can derive

$$\sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) = \sigma_r(\boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top} \boldsymbol{O}) \geq \lambda_r^\star \cdot \sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{O}) \geq \lambda_r^\star \cdot \sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{O}_{:,1:r}), \tag{120}$$

where $\boldsymbol{O}_{:,1:r} \in \mathbb{R}^{d\times r}$ is composed of the first r columns of $\boldsymbol{O}$.

To proceed, observe that $\boldsymbol{O}_{:,1:r}$ has the same distribution as $\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1/2}$, where $\boldsymbol{G} \in \mathbb{R}^{d\times r}$ is a random matrix with i.i.d. standard Gaussian entries. Hence, it suffices to develop a high-probability lower bound for $\sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1/2})$. Towards this end, we first make the observation that

$$\sigma_r\big(\boldsymbol{V}^{\star\top}\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1/2}\big) \geq \sigma_r\big(\boldsymbol{V}^{\star\top}\boldsymbol{G}\big)\sigma_r\big((\boldsymbol{G}^\top\boldsymbol{G})^{-1/2}\big) = \frac{\sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{G})}{\sigma_1(\boldsymbol{G})}. \tag{121}$$

It is clearly seen that $\boldsymbol{V}^{\star\top}\boldsymbol{G}$ is also a random matrix with i.i.d. standard Gaussian entries. In view of Lemmas 19 and 20, there exists some universal constant $c_0 > 0$ such that

$$\frac{\sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{G})}{\sigma_1(\boldsymbol{G})} \geq c_0\frac{1/\sqrt{r}}{\sqrt{d}} = \frac{c_0}{\sqrt{dr}} \tag{122}$$

holds with probability at least 0.995. Taking the above arguments together, we arrive at

$$\sigma_r(\boldsymbol{M}^\star\boldsymbol{O}) \geq \frac{c_0\lambda_r^\star}{\sqrt{dr}} \tag{123}$$

with probability at least 0.995.

B.6 Scalar dynamics with time-varying prefactors in learning rates

This subsection presents a slight extension of Lemma 2 to accommodate slightly broader learning rates.

Lemma 11. *Consider the scalar updates in (27), where $0 \leq \lambda^\star \leq \lambda_{\max}^\star$. Set the learning rate schedule to be*

$$\eta_t = C_{\eta,t}\sqrt{\lambda_{\max}^\star}\rho^t \qquad \text{for some } 1 \leq C_{\eta,t} \leq 2 \text{ and } 2/3 \leq \rho < 1.$$

Assume that $0 < |u_0| \leq \eta_0$. Then, with probability 1, for all $t \geq 0$, it holds that

$$\big||u_t| - \sqrt{\lambda^\star}\big| \leq \frac{2}{1-\rho}\sqrt{\lambda_{\max}^\star}\rho^t, \tag{124a}$$
$$|u_t^2 - \lambda^\star| \leq \left(\frac{4}{(1-\rho)^2} + \frac{4}{1-\rho}\right)\lambda_{\max}^\star\rho^t. \tag{124b}$$

Proof of Lemma 11. Similarly to the proof of Lemma 2, define $\Delta_t := \big||u_t| - \sqrt{\lambda^\star}\big|$, which satisfies (see (35))

$$\Delta_{t+1} = \big|\Delta_t - \eta_t\big|. \tag{125}$$

Next, define the tail sum $S_t := \sum_{s=t}^{\infty} \eta_s$. We claim for the moment that

$$\Delta_t \le S_t \qquad \text{for all } t \ge 0. \tag{126}$$

Once Equation (126) is established, the first claim (124a) follows immediately since

$$S_t = \sum_{s=t}^{\infty} C_{\eta,s}\sqrt{\lambda^\star_{\max}}\rho^s \le 2\sqrt{\lambda^\star_{\max}}\sum_{s=t}^{\infty}\rho^s = \frac{2}{1-\rho}\sqrt{\lambda^\star_{\max}}\rho^t, \tag{127}$$

where we have used $C_{\eta,s} \le 2$ for all $s \ge 0$.

It remains to prove the claim (126), which we accomplish by induction.

- *Base case ($t = 0$).* Recalling that $\sqrt{\lambda^\star} \le \sqrt{\lambda^\star_{\max}}$ and $|u_0| \le \eta_0$, we have

$$\Delta_0 = \big||u_0| - \sqrt{\lambda^\star}\big| \le |u_0| + \sqrt{\lambda^\star} \le \eta_0 + \sqrt{\lambda^\star_{\max}} \le 2\sqrt{\lambda^\star_{\max}} + \sqrt{\lambda^\star_{\max}} = 3\sqrt{\lambda^\star_{\max}}, \tag{128}$$

  which follows since $\eta_0 = C_{\eta,0}\sqrt{\lambda^\star_{\max}} \le 2\sqrt{\lambda^\star_{\max}}$. Moreover, since $C_{\eta,s} \ge 1$ for all s, we obtain

$$S_0 = \sum_{s=0}^{\infty}\eta_s = \sum_{s=0}^{\infty} C_{\eta,s}\sqrt{\lambda^\star_{\max}}\rho^s \ge \sqrt{\lambda^\star_{\max}}\sum_{s=0}^{\infty}\rho^s = \frac{1}{1-\rho}\sqrt{\lambda^\star_{\max}} \ge 3\sqrt{\lambda^\star_{\max}}, \tag{129}$$

  with the proviso that $\rho \ge 2/3$. Therefore, $S_0 \ge 3\sqrt{\lambda^\star_{\max}} \ge \Delta_0$, thus validating the base case.

- *Inductive step.* Now assume that $\Delta_t \le S_t$ for some $t \ge 0$. To bound Δ_{t+1}, we divide into two cases.

  - *Case 1:* $\Delta_t \ge \eta_t$. In this case, Equation (125) yields $\Delta_{t+1} = \Delta_t - \eta_t \le S_t - \eta_t = S_{t+1}$, which holds since $S_{t+1} = S_t - \eta_t$.
  - *Case 2:* $\Delta_t \le \eta_t$. In this case, Equation (125) yields $\Delta_{t+1} = \eta_t - \Delta_t \le \eta_t$. Thus it suffices to show that $\eta_t \le S_{t+1}$. Given that $C_{\eta,t} \le 2$ and $C_{\eta,s} \ge 1$ for all s, we derive

$$\eta_t = C_{\eta,t}\sqrt{\lambda^\star_{\max}}\rho^t \le 2\sqrt{\lambda^\star_{\max}}\rho^t,$$
$$S_{t+1} = \sum_{s=t+1}^{\infty}\eta_s \ge \sqrt{\lambda^\star_{\max}}\sum_{s=t+1}^{\infty}\rho^s = \frac{\rho}{1-\rho}\sqrt{\lambda^\star_{\max}}\rho^t \ge 2\sqrt{\lambda^\star_{\max}}\rho^t,$$

    provided that $\rho \ge 2/3$. This establishes that $\Delta_{t+1} \le S_{t+1}$.

  Combining these cases justifies Equation (126) at time $t+1$, which in turn establishes the claim (126).

Equipped with Equation (124a), we can now readily prove Equation (124b). For any $t \ge 0$,

$$|u_t^2 - \lambda^\star| = \big||u_t| - \sqrt{\lambda^\star}\big|\big(|u_t| + \sqrt{\lambda^\star}\big) = \Delta_t\big(\Delta_t + 2\sqrt{\lambda^\star}\big) \le \Delta_t\big(\Delta_t + 2\sqrt{\lambda^\star_{\max}}\big), \tag{130}$$

where we used $\lambda^\star \le \lambda^\star_{\max}$. Applying Equation (124a) leads to

$$|u_t^2 - \lambda^\star| \le \frac{2}{1-\rho}\sqrt{\lambda^\star_{\max}}\rho^t\left(\frac{2}{1-\rho}\sqrt{\lambda^\star_{\max}}\rho^t + 2\sqrt{\lambda^\star_{\max}}\right) \le \left(\frac{4}{(1-\rho)^2} + \frac{4}{1-\rho}\right)\lambda^\star_{\max}\rho^t$$

as claimed. □

C Lower bound for SignGD in matrix factorization (Proof of Theorem 2)

In this proof, we first establish a convergence lower bound for a two-dimensional quadratic optimization problem, and then show that a 2×2 matrix factorization instance can be reduced to this problem, thereby inheriting the same lower bound.

Step 1: a convergence lower bound for a quadratic optimization problem. Specifically, consider the following 2-dimensional quadratic minimization problem:

$$\underset{\boldsymbol{z}\in\mathbb{R}^2}{\text{minimize}} \quad f(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z}, \tag{131}$$

where the matrix $\boldsymbol{H}$ is symmetric positive semidefinite given by

$$\boldsymbol{H} = \frac{1}{2}\begin{pmatrix} \kappa+1 & \kappa-1 \\ \kappa-1 & \kappa+1 \end{pmatrix} \tag{132}$$

with two eigenvalues $\kappa \geq 1$ and 1. Clearly, the condition number of this matrix (or the Hessian of $f(\cdot)$) is κ, and the optimal objective value of the problem (131) is 0, attained at $\boldsymbol{z} = \boldsymbol{0}$. When applied to this problem, the SignGD algorithm proceeds as

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta_t \,\mathsf{sign}\left(\nabla f(\boldsymbol{z}_t)\right) = \boldsymbol{z}_t - \eta_t \,\mathsf{sign}(\boldsymbol{H}\boldsymbol{z}_t), \qquad t = 0, 1, \cdots \tag{133}$$

where $\eta_t > 0$ is the learning rate at iteration t, and the $\mathsf{sign}(\cdot)$ operator is applied entrywise.

We now present a convergence lower bound for SignGD on this structured quadratic objective. The proof is deferred to Section C.1.

Lemma 12. *Consider solving the problem (131) using SignGD (cf. (133)). Let $\{\eta_t\}_{t\geq 0}$ be any non-increasing sequence of learning rates, and consider any accuracy level obeying $0 < \varepsilon \leq \eta_0/\kappa$. Then, one can find an initialization $\boldsymbol{z}_0 \in [-2\eta_0, 2\eta_0]^2$ such that $\|\boldsymbol{z}_t\|_2 \leq \varepsilon$ can only happen after $t \geq \frac{\kappa-1}{4}$.*

Step 2: reduction of matrix factorization to quadratic optimization. Next, we demonstrate that a 2×2 instance of matrix factorization can be reduced to the quadratic optimization problem studied in Step 1.

To be precise, consider the following matrix factorization problem:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{2\times 2}}{\text{minimize}} \quad F(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{H}\right\|_{\mathrm{F}}^2, \qquad \text{with } \boldsymbol{H} = \frac{1}{2}\begin{pmatrix} \kappa+1 & \kappa-1 \\ \kappa-1 & \kappa+1 \end{pmatrix},\ \kappa \geq 1. \tag{134}$$

Set $\boldsymbol{U}^\star = \boldsymbol{H}^{1/2}$ to be the symmetric square root of $\boldsymbol{H}$. The SignGD algorithm proceeds as

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \,\mathsf{sign}\left(\nabla F(\boldsymbol{U}_t)\right) = \boldsymbol{U}_t - \eta_t \,\mathsf{sign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{H})\boldsymbol{U}_t\right), \quad t = 0, 1, \cdots \tag{135}$$

where $\mathsf{sign}(\cdot)$ is applied entrywise. The following lemma—whose proof is provided in Section C.2—develops a lower bound on the iteration complexity of SignGD.

Lemma 13. *Consider any learning rate sequence $\{\eta_t\}$ that is non-increasing in t. Then, there exists a universal constant $r_0 \in (0, 1/16)$ such that: for any target accuracy $\varepsilon > 0$ satisfying $\varepsilon \leq \frac{9r_0^2}{4096\kappa^2}$ and any initial $\eta_0 \leq r_0$, one can find an initialization $\boldsymbol{U}_0$ obeying $\|\boldsymbol{U}_0 - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq r_0$ such that the SignGD trajectory (135) cannot yield $F(\boldsymbol{U}_T) \leq \varepsilon$ unless*

$$T \geq \frac{\kappa-1}{4}. \tag{136}$$

This concludes the proof of Theorem 2.

C.1 Proof of Lemma 12

The proof is carried out in the following steps.

Step 1: a rotated basis aligned with the sign geometry. For each $t \geq 0$, define

$$\widetilde{\boldsymbol{z}}_t \coloneqq \boldsymbol{R}^\top \boldsymbol{z}_t \quad \text{with } \boldsymbol{R}^\top = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}; \qquad \text{and} \quad \widetilde{\boldsymbol{H}} \coloneqq \begin{pmatrix} \kappa & \\ & 1 \end{pmatrix}. \tag{137}$$

In words, $\widetilde{\boldsymbol{z}}_t = [\widetilde{z}_{1,t}, \widetilde{z}_{2,t}]^\top$ is obtained by rotating the original iterate $\boldsymbol{z}_t$. These allow one to express both the objective value and its gradient at iteration t as

$$f(\boldsymbol{z}_t) = \frac{1}{2}\widetilde{\boldsymbol{z}}_t^\top \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{z}}_t = \frac{1}{2}\big(\kappa\widetilde{z}_{1,t}^2 + \widetilde{z}_{2,t}^2\big), \qquad \nabla f(\boldsymbol{z}_t) = \boldsymbol{H}\boldsymbol{z}_t = \boldsymbol{R}\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{z}}_t = \frac{1}{\sqrt{2}}\begin{pmatrix} \kappa\widetilde{z}_{1,t} - \widetilde{z}_{2,t} \\ \kappa\widetilde{z}_{1,t} + \widetilde{z}_{2,t} \end{pmatrix}. \tag{138}$$

It is easily seen that $\boldsymbol{s}_t \coloneqq \mathsf{sign}(\nabla f(\boldsymbol{z}_t)) \in \{\pm 1\}^2$. Thus, the `SignGD` update in the rotated basis becomes

$$\widetilde{\boldsymbol{z}}_{t+1} = \boldsymbol{R}^\top\big(\boldsymbol{z}_t - \eta_t \boldsymbol{s}_t\big) = \widetilde{\boldsymbol{z}}_t - \eta_t \boldsymbol{R}^\top \boldsymbol{s}_t. \tag{139}$$

Given that there are only 4 possibilities in $\{\pm 1\}^2$, there are also only 4 possible update directions:

$$\boldsymbol{R}^\top \boldsymbol{s}_t = \begin{cases} (\sqrt{2}, 0) & \text{if } \boldsymbol{s}_t = (1,1), \\ (-\sqrt{2}, 0) & \text{if } \boldsymbol{s}_t = (-1,-1), \\ (0, -\sqrt{2}) & \text{if } \boldsymbol{s}_t = (1,-1), \\ (0, \sqrt{2}) & \text{if } \boldsymbol{s}_t = (-1,1). \end{cases} \tag{140}$$

This implies that in the eigenbasis (i.e., the above rotated coordinate system), each step of `SignGD` updates exactly one coordinate—either $\widetilde{z}_{1,t}$ or $\widetilde{z}_{2,t}$, but never both.

Step 2: a condition that governs the sign patterns of the updates. As it turns out, there exists a condition—based on the ratio of $|\widetilde{z}_{1,t}|$ and $|\widetilde{z}_{2,t}|$—that determines when `SignGD` updates each coordinate.

Lemma 14. *Consider any iteration t.*

- *If $|\kappa\widetilde{z}_{1,t}| > |\widetilde{z}_{2,t}|$, then*
$$\widetilde{z}_{1,t+1} = \widetilde{z}_{1,t} - \sqrt{2}\eta_t\, \mathsf{sign}(\widetilde{z}_{1,t}), \qquad \widetilde{z}_{2,t+1} = \widetilde{z}_{2,t}. \tag{141}$$
- *If $|\kappa\widetilde{z}_{1,t}| < |\widetilde{z}_{2,t}|$, then*
$$\widetilde{z}_{1,t+1} = \widetilde{z}_{1,t}, \qquad \widetilde{z}_{2,t+1} = \widetilde{z}_{2,t} - \sqrt{2}\eta_t\, \mathsf{sign}(\widetilde{z}_{2,t}). \tag{142}$$

Proof of Lemma 14. According to (138), the signs of the two coordinates of $\nabla f(\boldsymbol{z}_t)$ differ when

$$(\kappa\widetilde{z}_{1,t} - \widetilde{z}_{2,t})(\kappa\widetilde{z}_{1,t} + \widetilde{z}_{2,t}) < 0 \quad \Longleftrightarrow \quad (\kappa\widetilde{z}_{1,t})^2 < \widetilde{z}_{2,t}^2 \quad \Longleftrightarrow \quad |\kappa\widetilde{z}_{1,t}| < |\widetilde{z}_{2,t}|. \tag{143}$$

If $|\kappa\widetilde{z}_{1,t}| > |\widetilde{z}_{2,t}|$, then both components of $\nabla f(\boldsymbol{z}_t)$ have signs equal to $\mathsf{sign}(\widetilde{z}_{1,t})$, and hence the update vector is $(\pm\sqrt{2}, 0)$ in the rotated basis. If instead $|\kappa\widetilde{z}_{1,t}| < |\widetilde{z}_{2,t}|$, then the signs of the two components of $\nabla f(\boldsymbol{z}_t)$ are equal to $-\mathsf{sign}(\widetilde{z}_{2,t})$ and $\mathsf{sign}(\widetilde{z}_{2,t})$, respectively, and hence the update vector in the rotated basis is $(0, \pm\sqrt{2})$. □

Step 3: a learning rate barrier. We now develop a general lower bound for the following sequence that updates one coordinate at a time. Specifically, consider a sequence $\boldsymbol{x}_t = [x_{1,t}, x_{2,t}]^\top$, $t \geq 0$, that follows the update rule below:

- If $|x_{1,t}| < |x_{2,t}|/\kappa$, then $x_{2,t+1} = x_{2,t} - \eta_t$ and $x_{1,t+1} = x_{1,t}$;
- If $|x_{1,t}| > |x_{2,t}|/\kappa$, then $x_{1,t+1} = x_{1,t} - \eta_t$ and $x_{2,t+1} = x_{2,t}$.

- If $|x_{1,t}| = |x_{2,t}|/\kappa$, then $x_{1,t+1}$ and $x_{2,t+1}$ can be chosen arbitrarily.

Lemma 15. *Consider the above sequence $\{\boldsymbol{x}_t\}_{0\le t\le T}$ for any finite T. Let $\{\eta_t\}_{t\ge 0}$ be a non-increasing sequence of learning rates. For any target accuracy $0 < \varepsilon \le \eta_0/\kappa$, there exists an initialization $\boldsymbol{x}_0 \in [0,\eta_0]^2$ such that $x_{2,t} < \kappa\varepsilon$ can only happen when $\eta_t < 4\varepsilon$.*

Proof of Lemma 15. Let us initialize at $\boldsymbol{x}_0 = [x_{1,0}, \kappa\varepsilon]^\top$, where $x_{1,0}$ is defined recursively as follows.

- Let $T_0 = \min\{T, \max\{t : \eta_t \ge 4\varepsilon\}\}$; choose $x_{1,T_0} \in [2\varepsilon, \eta_{T_0} - 2\varepsilon]$.
- Define the previous iterates backward:

$$x_{1,t} \coloneqq \eta_t - x_{1,t+1}, \qquad \text{for } t = T_0 - 1, T_0 - 2, \ldots, 0. \tag{144}$$

Now we show by induction that $x_{1,t} \in [2\varepsilon, \eta_t - 2\varepsilon]$ for all $0 \le t \le T_0$. The base case with $t = T_0$ holds trivially by construction. Assume the induction hypothesis holds at $t+1$, i.e., $x_{1,t+1} \in [2\varepsilon, \eta_{t+1} - 2\varepsilon]$. Then it follows from the assumption $\eta_t \ge \eta_{t+1}$ that

$$\begin{aligned} x_{1,t} &= \eta_t - x_{1,t+1} \ge \eta_t - \eta_{t+1} + 2\varepsilon \ge 2\varepsilon > 0, \\ x_{1,t} &= \eta_t - x_{1,t+1} \le \eta_t - 2\varepsilon, \end{aligned} \tag{145}$$

thus justifying the induction hypothesis at t. Hence, we establish by induction that $x_{1,t} \in [2\varepsilon, \eta_t - 2\varepsilon]$ holds for all $0 \le t \le T_0$. As immediate consequences, for all $0 \le t \le T_0$ one has: (i) $x_{1,t} > \varepsilon$; (ii) the update rule described above for $\{\boldsymbol{x}_t\}$ always applies only to the first coordinate $x_{1,t}$, with $x_{2,t}$ frozen at $\kappa\varepsilon$ (given that $|x_{1,t}|/|x_{2,t}| > \varepsilon/(\kappa\varepsilon) = 1/\kappa$). This concludes the proof. □

Step 4: putting all this together. Let us initialize `SignGD` to $\widetilde{\boldsymbol{z}}_0 = \sqrt{2}\boldsymbol{x}_0$, with $\boldsymbol{x}_0$ constructed in the proof of Lemma 15. Clearly, one has $\widetilde{\boldsymbol{z}}_0 \in [0, \sqrt{2}\eta_0]^2$, which together with $\boldsymbol{z}_0 = \boldsymbol{R}\widetilde{\boldsymbol{z}}_0$ gives $\boldsymbol{z}_0 \in [0, 2\eta_0]^2$. Moreover, it is seen from Lemma 14 that $\big\{\big(\frac{1}{\sqrt{2}}\widetilde{z}_{1,t}, \frac{1}{\sqrt{2}}\widetilde{z}_{2,t}\big)\big\}$ follows the same dynamics as $\{\boldsymbol{x}_t\}$ in Lemma 15—and hence $\frac{1}{\sqrt{2}}\widetilde{z}_{2,t} = \kappa\varepsilon$—before η_t drops below 4ε. To reduce $\widetilde{z}_{2,t}$ from $\sqrt{2}\kappa\varepsilon$ to below ε using learning rates at most 4ε, with each iteration changing the coordinate by at most $\sqrt{2}\eta_t$, the number of iterations needs to at least exceed

$$\frac{\sqrt{2}\kappa\varepsilon - \varepsilon}{4\sqrt{2}\varepsilon} \ge \frac{\kappa - 1}{4}, \tag{146}$$

thus completing the proof.

C.2 Proof of Lemma 13

The proof comprises several steps as described below. Throughout this proof, we shall focus on initializations residing within the following subspace:

$$\mathcal{S} := \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} : (a,b) \in \mathbb{R}^2 \right\}. \tag{147}$$

For any $\boldsymbol{U} = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \in \mathcal{S}$, we shall refer to (a,b) as its induced parameters.

Step 1: invariance of the set $\mathcal{S}$ under `SignGD` updates. We first show that, when initialized in $\mathcal{S}$, the entire trajectory of `SignGD` stays within $\mathcal{S}$.

Lemma 16 (Invariance of $\mathcal{S}$)**.** *If $\boldsymbol{U} \in \mathcal{S}$, then $\nabla F(\boldsymbol{U}) \in \mathcal{S}$ and $\mathsf{sign}(\nabla F(\boldsymbol{U})) \in \mathcal{S}$. Consequently, $\boldsymbol{U}_0 \in \mathcal{S}$ implies $\boldsymbol{U}_t \in \mathcal{S}$ for all t.*

Proof of Lemma 16. Note that any $\boldsymbol{U} \in \mathcal{S}$ can be written as

$$\boldsymbol{U} = \begin{pmatrix} a & b \\ b & a \end{pmatrix} = a\boldsymbol{I} + b\boldsymbol{J}, \qquad \text{with } \boldsymbol{I} = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \text{ and } \boldsymbol{J} = \begin{pmatrix} & 1 \\ 1 & \end{pmatrix}. \tag{148}$$

As can be easily verified, products of such matrices from $\mathcal{S}$ remain in $\mathcal{S}$. As a result, $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^2 \in \mathcal{S}$, so $(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{H}) \in \mathcal{S}$, and multiplying by $\boldsymbol{U} \in \mathcal{S}$ yields $\nabla F(\boldsymbol{U}) = (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{H})\boldsymbol{U} \in \mathcal{S}$. If a matrix has equal diagonals and equal off-diagonals, then applying $\mathsf{sign}(\cdot)$ entrywise preserves these equalities. □

Consequently, it suffices to focus on analyzing the dynamics within $\mathcal{S}$.

Step 2: equivalent updates of induced parameters. Set

$$\boldsymbol{R} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \text{and hence} \quad \boldsymbol{R}^\top \boldsymbol{H} \boldsymbol{R} = \begin{pmatrix} \kappa & \\ & 1 \end{pmatrix}. \tag{149}$$

For any $\boldsymbol{U} \in \mathcal{S}$ with induced parameters (a, b), one can easily verify that

$$\boldsymbol{R}^\top \boldsymbol{U} \boldsymbol{R} = \mathsf{diag}\{\lambda_1, \lambda_2\}, \qquad \text{with } \lambda_1 = a + b,\ \lambda_2 = a - b. \tag{150}$$

Define

$$\lambda_1^\star = \sqrt{\kappa},\ \lambda_2^\star = 1, \qquad \delta_1 := \lambda_1 - \sqrt{\kappa},\ \delta_2 := \lambda_2 - 1, \tag{151}$$

where $\lambda_1^\star$ and $\lambda_2^\star$ correspond to the two eigenvalues of $\boldsymbol{U}^\star = \boldsymbol{H}^{1/2}$. These allow us to convert the gradient into *exact* diagonal form as

$$\boldsymbol{R}^\top \nabla F(\boldsymbol{U}) \boldsymbol{R} = \mathsf{diag}\{g_1(\lambda_1), g_2(\lambda_2)\}, \qquad \text{with } g_1(\lambda) \coloneqq (\lambda^2 - \kappa)\lambda,\ g_2(\lambda) \coloneqq (\lambda^2 - 1)\lambda. \tag{152}$$

Equivalently, the gradient in the original basis can be expressed as

$$\nabla F(\boldsymbol{U}) = \begin{pmatrix} G_{\mathrm{d}} & G_{\mathrm{o}} \\ G_{\mathrm{o}} & G_{\mathrm{d}} \end{pmatrix}, \qquad \text{with } G_{\mathrm{d}} = \frac{g_1(\lambda_1) + g_2(\lambda_2)}{2},\ G_{\mathrm{o}} = \frac{g_1(\lambda_1) - g_2(\lambda_2)}{2}. \tag{153}$$

Given that the update is entrywise, the induced parameter update on (a, b) can be written as

$$a_{t+1} = a_t - \eta_t\, \mathsf{sign}(G_{\mathrm{d},t}), \qquad b_{t+1} = b_t - \eta_t\, \mathsf{sign}(G_{\mathrm{o},t}). \tag{154}$$

Step 3: local gradient signs. Next, expand $g_1(\cdot)$ (resp. $g_2(\cdot)$) around $\lambda_1^\star = \sqrt{\kappa}$ (resp. $\lambda_2^\star = 1$) as

$$g_1(\sqrt{\kappa} + \delta_1) = \big((\sqrt{\kappa} + \delta_1)^2 - \kappa\big)(\sqrt{\kappa} + \delta_1) = (2\sqrt{\kappa}\delta_1 + \delta_1^2)(\sqrt{\kappa} + \delta_1) = 2\kappa\delta_1 + 3\sqrt{\kappa}\delta_1^2 + \delta_1^3, \tag{155a}$$

$$g_2(1 + \delta_2) = \big((1 + \delta_2)^2 - 1\big)(1 + \delta_2) = (2\delta_2 + \delta_2^2)(1 + \delta_2) = 2\delta_2 + 3\delta_2^2 + \delta_2^3. \tag{155b}$$

Fix a universal radius $r_0 \in (0, 1/16)$ and consider the local region with

$$|\delta_1| \le \sqrt{\kappa} r_0, \qquad |\delta_2| \le r_0. \tag{156}$$

In this region, the higher-order terms are dominated by the linear terms: indeed, using Equation (155a) and $|\delta_1| \le \sqrt{\kappa} r_0$, we can derive

$$|3\sqrt{\kappa}\delta_1^2 + \delta_1^3| \le \big(3\sqrt{\kappa}\delta_1 + |\delta_1|^2\big)\,|\delta_1| \le \big(3\kappa r_0 + \kappa r_0^2\big)\,|\delta_1| \le \frac{1}{2}\kappa|\delta_1| \tag{157}$$

for $r_0 \le 1/16$, which allows us to express

$$g_1(\sqrt{\kappa} + \delta_1) = 2\kappa\delta_1 + \Delta_1 \qquad \text{for some } |\Delta_1| \le \frac{1}{2}\kappa|\delta_1|. \tag{158a}$$

Similarly, it follows from Equation (155b) and $|\delta_2| \le r_0$ that

$$g_2(1+\delta_2) = 2\delta_2 + \Delta_2 \qquad \text{for some } |\Delta_2| \le \frac{1}{2}|\delta_2|. \tag{158b}$$

Recall the expressions of G_{d} and G_{o} in (153), which combined with (158) yields

$$G_{\mathrm{d}} = \kappa\delta_1 + \delta_2 + \frac{\Delta_1 + \Delta_2}{2}, \tag{159a}$$

$$G_{\mathrm{o}} = \kappa\delta_1 - \delta_2 + \frac{\Delta_1 - \Delta_2}{2}. \tag{159b}$$

Moreover, it follows from Equations (158a) and (158b) that

$$\left|\frac{\Delta_1 \pm \Delta_2}{2}\right| \le \frac{|\Delta_1| + |\Delta_2|}{2} \le \frac{1}{4}(\kappa|\delta_1| + |\delta_2|). \tag{160}$$

As a result, one has

$$\mathsf{sign}(G_{\mathrm{d}}) = \mathsf{sign}(\kappa\delta_1 + \delta_2), \qquad \mathsf{sign}(G_{\mathrm{o}}) = \mathsf{sign}(\kappa\delta_1 - \delta_2), \tag{161}$$

provided that

$$\min\{|\kappa\delta_1 + \delta_2|, |\kappa\delta_1 - \delta_2|\} > \frac{1}{4}(\kappa|\delta_1| + |\delta_2|). \tag{162}$$

Step 4: `SignGD` exhibiting matching dynamics as in Lemma 12. Define

$$\boldsymbol{s}_t \coloneqq \begin{pmatrix} \mathsf{sign}(G_{\mathrm{d},t}) \\ \mathsf{sign}(G_{\mathrm{o},t}) \end{pmatrix} \in \{\pm 1\}^2. \tag{163}$$

The iterative updates of the (a, b) parameters described in (154) can be written compactly as

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta_t \boldsymbol{s}_t \qquad \text{with } \boldsymbol{u}_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix}. \tag{164}$$

Such update rules can be translated into updates over the eigenvalues. More specifically, set the eigenvalues of $\boldsymbol{U}_t$ to be $\sqrt{\kappa} + \delta_{1,t}$ and $1 + \delta_{2,t}$, which combined with the fact that $\boldsymbol{U}_t \in \mathcal{S}$ gives

$$\boldsymbol{R}^\top \boldsymbol{U}_t \boldsymbol{R} = \mathsf{diag}\{\sqrt{\kappa} + \delta_{1,t}, 1 + \delta_{2,t}\}. \tag{165}$$

A little algebra then allows us to translate Equation (164) into

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \widetilde{\eta}_t \boldsymbol{R}^\top \boldsymbol{s}_t \qquad \text{with } \widetilde{\eta}_t \coloneqq \sqrt{2}\eta_t, \tag{166}$$

where $\boldsymbol{\delta}_t = [\delta_{1,t}, \delta_{2,t}]^\top$, and $\{\widetilde{\eta}_t\}$ is clearly also a non-increasing learning rate sequence.

The above update rule (166) bears similarity with the one (139) analyzed in Lemma 12. By initializing $\boldsymbol{\delta}_0$ to be $\widetilde{\boldsymbol{z}}_0$ as in the proof of Lemma 12—except that η_t is replaced with $\widetilde{\eta}_t$ and ε replaced with ε_q (to be specified shortly) in the construction of this initialization—we see from the proof of Lemma 12 that

$$\kappa|\delta_{1,0}| \ge 2|\delta_{2,0}|,$$

which satisfies the condition described in (162). Thus, combining it with Equation (161) leads to

$$\boldsymbol{s}_0 = \begin{pmatrix} \mathsf{sign}(\kappa\delta_{1,0} + \delta_{2,0}) \\ \mathsf{sign}(\kappa\delta_{1,0} - \delta_{2,0}) \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{\delta}_1 = \boldsymbol{\delta}_0 - \widetilde{\eta}_0 \boldsymbol{R}^\top \begin{pmatrix} \mathsf{sign}(\kappa\delta_{1,0} + \delta_{2,0}) \\ \mathsf{sign}(\kappa\delta_{1,0} - \delta_{2,0}) \end{pmatrix}, \tag{167}$$

which is precisely the update rule of $\widetilde{\boldsymbol{z}}_1$ in the proof of Lemma 12. Continuing these arguments and taking advantage of the properties derived in the proof of Lemma 12, we can readily see that: for any $t \le T_0$ with $T_0 \coloneqq \min\big\{\max\{t : \widetilde{\eta}_t \ge 4\varepsilon_q\}, \lceil \frac{\kappa-1}{4} \rceil\big\}$, one has

$$\kappa|\delta_{1,t}| \ge 2|\delta_{2,t}|,$$

which obeys the condition described in Equation (162) and in turns results in

$$\boldsymbol{s}_t = \begin{pmatrix} \mathsf{sign}(\kappa\delta_{1,t} + \delta_{2,t}) \\ \mathsf{sign}(\kappa\delta_{1,t} - \delta_{2,t}) \end{pmatrix} \qquad \Longrightarrow \qquad \boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \widetilde{\eta}_t \boldsymbol{R}^\top \begin{pmatrix} \mathsf{sign}(\kappa\delta_{1,t} + \delta_{2,t}) \\ \mathsf{sign}(\kappa\delta_{1,t} - \delta_{2,t}) \end{pmatrix}. \tag{168}$$

Consequently, by construction (again see the proof of Lemma 12) one has

$$\delta_{2,t} \ge \kappa\varepsilon_q \qquad \text{for every } t \le T_0. \tag{169}$$

Step 5: connecting $F(\boldsymbol{U})$ with $\boldsymbol{\delta}_t$-updates. On $\mathcal{S}$, the objective admits an exact eigen-form:

$$F(\boldsymbol{U}) = \frac{1}{4}\big((\lambda_1^2 - \kappa)^2 + (\lambda_2^2 - 1)^2\big) = \frac{1}{4}\big((2\sqrt{\kappa}\delta_1 + \delta_1^2)^2 + (2\delta_2 + \delta_2^2)^2\big). \tag{170}$$

where as before we take the eigenvalues of $\boldsymbol{U}$ to be $\sqrt{\kappa} + \delta_1$ and $1 + \delta_2$. In the local region described in Equation (156) with $r_0 \le 1/16$, we have $|\delta_2| \le r_0 \le 1/16$, hence

$$|2\delta_2 + \delta_2^2| \ge 2|\delta_2| - \delta_2^2 \ge \frac{3}{2}|\delta_2|. \tag{171}$$

Substitution into Equation (170) yields the local lower bound:

$$F(\boldsymbol{U}) \ge \frac{1}{4}(2\delta_2 + \delta_2^2)^2 \ge \frac{9}{16}\delta_2^2. \tag{172}$$

Therefore, any iterate $\boldsymbol{U}_T$ obeying $F(\boldsymbol{U}_T) \le \varepsilon$ necessarily satisfies

$$|\delta_{2,T}| \le \frac{4}{3}\sqrt{\varepsilon}. \tag{173}$$

As a consequence, setting the target level ε_q in Step 4 as $\varepsilon_q \coloneqq \frac{4}{3}\sqrt{\varepsilon}$, we see from Lemma 12 that

$$T \ge \frac{\kappa - 1}{4},$$

provided that $(\delta_{1,t}, \delta_{2,t})$ satisfies Condition (156). To finish up, it suffices to note that Condition (156) is guaranteed as long as

$$\eta_0 \le r_0, \qquad \kappa\varepsilon_q = \frac{4}{3}\kappa\sqrt{\varepsilon} \le r_0 \le \frac{1}{16}.$$

This follows from the fact that, for all $t \ge T$, we have $\delta_{1,t} \in [2\varepsilon_q, \eta_t - 2\varepsilon_q]$ and $\delta_{2,t} \equiv \kappa\varepsilon_q$ according to the proof of Lemma 12.

D Derivation of the training objective in Section 2.2

In this section, we provide a more detailed explanation of how the objective (20) arises from the framework of in-context learning (ICL). A common way to formalize ICL is to place a distribution over tasks (Garg et al., 2022), viewing each task as a function h drawn from a function class $\mathcal{H}$. A prompt consists of N input–label pairs followed by a query:

$$P = (\boldsymbol{x}_1, h(\boldsymbol{x}_1), \ldots, \boldsymbol{x}_N, h(\boldsymbol{x}_N), \boldsymbol{x}_{\mathrm{q}}),$$

where inputs $\boldsymbol{x}_i$ and query $\boldsymbol{x}_{\mathrm{q}}$ are sampled independently from certain data distribution $\mathcal{D}_{\mathcal{X}}$, and the task function h is drawn from $h \sim \mathcal{D}_{\mathcal{H}}$.

A model is said to have *in-context learned* the function class $\mathcal{H}$ if, when presented with a *fresh* task h' drawn from $\mathcal{H}$ and a corresponding fresh prompt, it can reliably predict the output $h'(\boldsymbol{x}_{\mathrm{q}})$ without updating its parameters. To understand how models acquire this ability through training, Garg et al. (2022) proposed a meta-learning protocol: at each training step, a task h and a sequence of data points are sampled to form a prompt, and the model parameters are updated to minimize the prediction error on the query. They empirically demonstrated that transformers trained in this manner can in-context learn, e.g., linear function classes. Motivated by these findings, a growing body of theoretical work has adopted this framework to study the optimization dynamics (Ahn et al., 2023; Zhang et al., 2024a; Huang et al., 2023).

Our instantiation: linear tasks with a fixed support set. Let us focus on linear regression tasks, where $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ for a task parameterized by vector $\boldsymbol{w} \in \mathbb{R}^d$. We adopt the fixed-design setting (Yang et al., 2024): the first N input tokens $\{\boldsymbol{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ in the prompt are fixed, with empirical covariance $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^\top$. We draw $\boldsymbol{w} \sim \mathcal{D}$ with $\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] = \boldsymbol{I}$, and generate noiseless labels $y_{\boldsymbol{w},i} = \boldsymbol{w}^\top \boldsymbol{x}_i$. The query is sampled uniformly from the support set, i.e., $\boldsymbol{x}_{\mathrm{q}} \sim \mathrm{Unif}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Following standard ICL practice (Garg et al., 2022; Zhang et al., 2024a; Ahn et al., 2023), we embed the prompt as

$$\boldsymbol{E}_{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N & \boldsymbol{x}_{\mathrm{q}} \\ y_{\boldsymbol{w},1} & y_{\boldsymbol{w},2} & \cdots & y_{\boldsymbol{w},N} & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(N+1)}. \tag{174}$$

The goal of ICL training is to optimize a model in order to reliably predict $\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{q}}$ from $\boldsymbol{E}_{\boldsymbol{w}}$.

Single-layer linear transformer. A standard single-layer transformer with input $\boldsymbol{E}_{\boldsymbol{w}}$ computes its output using softmax attention (Vaswani et al., 2017):

$$F_{\mathsf{softmax}}(\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V; \boldsymbol{E}_{\boldsymbol{w}}) := \boldsymbol{W}_V \boldsymbol{E}_{\boldsymbol{w}} \cdot \mathsf{softmax}\left(\frac{(\boldsymbol{W}_K \boldsymbol{E}_{\boldsymbol{w}})^\top (\boldsymbol{W}_Q \boldsymbol{E}_{\boldsymbol{w}})}{\gamma}\right),$$

where $\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V \in \mathbb{R}^{(d+1)\times(d+1)}$ represent the key, query, and value weight matrices, $\gamma > 0$ is a normalization factor, and the softmax operator $\mathsf{softmax}(\cdot)$ is applied column-wise. In this work, we consider a simplified model that is more amenable to theoretical analysis and commonly adopted in existing theoretical literature for ICL (Zhang et al., 2024a; Ahn et al., 2023; Huang et al., 2023). Specifically, we remove the softmax nonlinearity and merge $\boldsymbol{W}_Q, \boldsymbol{W}_V$ into a single $\boldsymbol{W}_{KQ}$, and take $\gamma = N$, resulting in

$$F_{\mathsf{linear}}(\boldsymbol{W}_V, \boldsymbol{W}_{KQ}; \boldsymbol{E}_{\boldsymbol{w}}) = \boldsymbol{W}_V \boldsymbol{E}_{\boldsymbol{w}} \Big(\frac{\boldsymbol{E}_{\boldsymbol{w}}^\top \boldsymbol{W}_{KQ} \boldsymbol{E}_{\boldsymbol{w}}}{N}\Big). \tag{175}$$

Furthermore, we take $\boldsymbol{W}_V$ and $\boldsymbol{W}_{KQ}$ to be the following specific forms as adopted in (Huang et al., 2023; Yang et al., 2024; Huang et al., 2025):

$$\boldsymbol{W}_V = \begin{pmatrix} \boldsymbol{0}_{d\times d} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 1 \end{pmatrix}, \quad \boldsymbol{W}_{KQ} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 0 \end{pmatrix}.$$

Therefore, the model can be parameterized by $\boldsymbol{Q}$, and the prediction for $\boldsymbol{x}_{\mathrm{q}}$ is read off from the bottom-right entry:

$$\widehat{y}_{\mathrm{q}} := \widehat{y}_{\mathrm{q}}(\boldsymbol{Q}; \boldsymbol{E}_{\boldsymbol{w}}) = [F_{\mathsf{linear}}(\boldsymbol{Q}; \boldsymbol{E}_{\boldsymbol{w}})]_{(d+1),(N+1)}. \tag{176}$$

By direct calculation, this admits a simplified closed-form expression:

$$\widehat{y}_{\mathrm{q}} = \begin{pmatrix} \boldsymbol{0}_d^\top & 1 \end{pmatrix} \left(\frac{\boldsymbol{E}_{\boldsymbol{w}} \boldsymbol{E}_{\boldsymbol{w}}^\top}{N}\right) \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{0}_d^\top \end{pmatrix} \boldsymbol{x}_{\mathrm{q}} = \frac{1}{N}\sum_{i=1}^N (\boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i^\top \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}} = \boldsymbol{w}^\top \boldsymbol{S} \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}}.$$

in-context learning objective. The training goal is to optimize $\boldsymbol{Q}$ to minimize the expected squared prediction risk, where the randomness comes from $\boldsymbol{w}$ and $\boldsymbol{x}_{\mathrm{q}}$ across prompts. Therefore,

$$f(\boldsymbol{Q}) \coloneqq \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{\mathrm{q}}}\Big[\big(\widehat{y}_{\mathrm{q}} - \boldsymbol{w}^\top \boldsymbol{x}\big)^2\Big] = \frac{1}{2N}\sum_{i=1}^{N} \|\boldsymbol{S}\boldsymbol{Q}\boldsymbol{x}_i - \boldsymbol{x}_i\|_2^2 = \frac{1}{2}\operatorname{tr}\big((\boldsymbol{S}\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{S}(\boldsymbol{S}\boldsymbol{Q}-\boldsymbol{I})^\top\big). \tag{177}$$

Minimizing this objective is exactly equivalent to solving the quadratic optimization problem (20).

E Lower bounds for SignGD in ICL (Proof of Theorem 4)

Consider any $\kappa \geq 2$. In what follows, we will construct an instance (i.e., a covariance matrix $\boldsymbol{S}$ obeying $\kappa(\boldsymbol{S})^3 = \kappa$), on which SignGD needs $\Omega(\kappa)$ iterations to achieve the target accuracy.

Step 1: construction of a 2-dimensional instance. Let $d = 2$ and define the rotation matrix

$$\boldsymbol{R} \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}. \tag{178}$$

Set the covariance matrix to be

$$\boldsymbol{S} \coloneqq \boldsymbol{R}\begin{pmatrix} \kappa^{1/3} & 0 \\ 0 & 1 \end{pmatrix}\boldsymbol{R}^\top. \tag{179}$$

It then follows that $\kappa(\boldsymbol{S}) = \kappa^{1/3}$, hence $\kappa(\boldsymbol{S})^3 = \kappa$.

Step 2: invariance of a 2-dimensional slice. Define the set

$$\mathcal{S} \coloneqq \left\{\boldsymbol{Q}(a,b) \coloneqq \begin{pmatrix} a & b \\ b & a \end{pmatrix} :\ (a,b) \in \mathbb{R}^2\right\}. \tag{180}$$

We now claim that: if $\boldsymbol{Q}_t \in \mathcal{S}$, then $\boldsymbol{Q}_{t+1}$ remains within $\mathcal{S}$.

Proof. To justify this claim, we first note that for any $\boldsymbol{Q} \in \mathcal{S}$, $\boldsymbol{Q}$ commutes with $\boldsymbol{R}\mathsf{diag}(\cdot)\boldsymbol{R}^\top$, hence $\boldsymbol{Q}$ commutes with $\boldsymbol{S}$ (cf. (179)), and therefore $\boldsymbol{S}\boldsymbol{Q}\boldsymbol{S} \in \mathcal{S}$. The gradient of the objective $f(\cdot)$ is

$$\nabla f(\boldsymbol{Q}) = \boldsymbol{S}^2\boldsymbol{Q}\boldsymbol{S} - \boldsymbol{S}^2, \tag{181}$$

which also falls within $\mathcal{S}$ whenever $\boldsymbol{Q} \in \mathcal{S}$. Additionally, the entrywise sign map preserves the structure $\left(\begin{smallmatrix} a & b \\ b & a \end{smallmatrix}\right)$, and as a result, $\mathsf{sign}(\nabla f(\boldsymbol{Q})) \in \mathcal{S}$. These taken together prove that $\boldsymbol{Q}_{t+1} \in \mathcal{S}$. □

Thus, it suffices to analyze the induced dynamics within $\mathcal{S}$. In what follows, we shall write $\boldsymbol{Q}_t = \boldsymbol{Q}(a_t, b_t)$, with (a_t, b_t) the induced parameters.

Step 3: an equivalent form of the objective. Any $\boldsymbol{Q}(a,b) \in \mathcal{S}$ is diagonalizable in the basis $\boldsymbol{R}$:

$$\boldsymbol{Q}(a,b) = \boldsymbol{R}\begin{pmatrix} q_1 & 0 \\ 0 & q_2 \end{pmatrix}\boldsymbol{R}^\top, \quad \text{where } \ q_1 = a+b \text{ and } q_2 = a-b. \tag{182}$$

Recall the diagonal form of $\boldsymbol{S}$ in (179). Letting $\sigma_1 = \kappa^{1/3}$ and $\sigma_2 = 1$, we can write

$$f\big(\boldsymbol{Q}(a,b)\big) = \frac{\sigma_1}{2}(\sigma_1 q_1 - 1)^2 + \frac{\sigma_2}{2}(\sigma_2 q_2 - 1)^2 = \frac{\kappa}{2}\big(q_1 - \kappa^{-1/3}\big)^2 + \frac{1}{2}\big(q_2 - 1\big)^2. \tag{183}$$

Similarly, if we express the solution $\boldsymbol{Q}^\star = \boldsymbol{S}^{-1}$ as

$$\boldsymbol{Q}^\star = \boldsymbol{Q}^\star(a^\star, b^\star) = \boldsymbol{R}\begin{pmatrix} q_1^\star & 0 \\ 0 & q_2^\star \end{pmatrix}\boldsymbol{R}^\top, \quad \text{where } \ q_1^\star = a^\star + b^\star \text{ and } q_2^\star = a^\star - b^\star, \tag{184}$$

then it can be easily verified that

$$q_1^\star = \kappa^{-1/3}, \quad q_2^\star = 1 \quad \Longrightarrow \quad a^\star = \frac{q_1^\star + q_2^\star}{2} = \frac{\kappa^{-1/3}+1}{2}, \quad b^\star = \frac{q_1^\star - q_2^\star}{2} = \frac{\kappa^{-1/3}-1}{2}. \tag{185}$$

Now, let us define the error coordinates

$$\boldsymbol{z} \coloneqq \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \coloneqq \begin{pmatrix} a - a^\star \\ b - b^\star \end{pmatrix}, \tag{186}$$

allowing us to write

$$q_1 - q_1^\star = (a - a^\star) + (b - b^\star) = z_1 + z_2 \quad \text{and} \quad q_2 - q_2^\star = (a - a^\star) - (b - b^\star) = z_1 - z_2.$$

It then follows from Equation (183) that

$$f\big(\boldsymbol{Q}(a,b)\big) = \frac{\kappa}{2}(z_1+z_2)^2 + \frac{1}{2}(z_1 - z_2)^2 = \frac{1}{2}\boldsymbol{z}^\top \begin{pmatrix} \kappa+1 & \kappa-1 \\ \kappa-1 & \kappa+1 \end{pmatrix} \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z} =: g(\boldsymbol{z}), \tag{187}$$

where

$$\boldsymbol{H} \coloneqq \frac{1}{2}\begin{pmatrix} \kappa+1 & \kappa-1 \\ \kappa-1 & \kappa+1 \end{pmatrix}. \tag{188}$$

In particular, $g(\cdot)$ is a quadratic function with minimizer $\boldsymbol{z} = \boldsymbol{0}$ and gradient $\nabla g(\boldsymbol{z}) = 2\boldsymbol{H}\boldsymbol{z}$.

Step 4: `SignGD` exhibiting matching dynamics as in Lemma 12. Given the invariance of $\mathcal{S}$ and the fact that (a_t, b_t) are the diagonal and off-diagonal entries of $\boldsymbol{Q}_t$, the `Muon` update induces

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta_t \,\mathsf{sign}(\boldsymbol{H}\boldsymbol{z}_t), \qquad t = 0, 1, 2, \ldots, \tag{189}$$

where $\boldsymbol{z}_t$ is defined by Equation (186) w.r.t. the t-th iterate, and $\boldsymbol{H}$ is given in Equation (188). This matches precisely the `SignGD` recursion studied in Lemma 12.

Therefore, for any non-increasing $\{\eta_t\}_{t\geq 0}$ and any $0 < \varepsilon \leq \sqrt{2}\eta_0/\kappa$, Lemma 12 guarantees that one can choose an initialization $\boldsymbol{z}_0 \in [0, 2\eta_0]^2$ such that $\|\boldsymbol{z}_t\|_2 \leq \varepsilon/\sqrt{2}$ can only occur after

$$t \geq \frac{\kappa - 1}{4}. \tag{190}$$

Step 5: translating it back to $\boldsymbol{Q}_t$. Recalling that $\boldsymbol{Q}_t = \boldsymbol{Q}(a_t, b_t)$ and $\boldsymbol{Q}^\star = \boldsymbol{Q}(a^\star, b^\star)$, we have

$$\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_{\mathrm{F}}^2 = 2(a_t - a^\star)^2 + 2(b_t - b^\star)^2 = 2\|\boldsymbol{z}_t\|_2^2, \quad \Longrightarrow \quad \|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_{\mathrm{F}} = \sqrt{2}\,\|\boldsymbol{z}_t\|_2. \tag{191}$$

Hence, with the above-mentioned initialization, achieving $\|\boldsymbol{Q}_t - \boldsymbol{Q}^\star\|_{\mathrm{F}} \leq \varepsilon$ requires at least $(\kappa-1)/4$ iterations. This establishes the `SignGD` lower bound claimed in Theorem 4.

F Technical lemmas

In this section, we gather a couple of technical lemmas that are useful in our analysis. We begin with two lemmas concerned with perturbation bounds for matrix signs and rank-r approximations.

Lemma 17 (Adapted from Theorem 2 in Li (1995))**.** *For arbitrary two matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$ $(m > n)$ of full column rank, we have*

$$\|\mathsf{msign}(\boldsymbol{X}) - \mathsf{msign}(\boldsymbol{Y})\| \leq \frac{3}{\sigma_n(\boldsymbol{X})}\,\|\boldsymbol{X} - \boldsymbol{Y}\|. \tag{192}$$

Lemma 18 (Adapted from Equation (4.4) in Wedin (1972)). *For any two matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$, denote the best rank-$r$ approximations by $\boldsymbol{X}_r, \boldsymbol{Y}_r$, respectively. We define the eigengap $\delta = \min\{\sigma_r(\boldsymbol{X}), \sigma_r(\boldsymbol{Y})\} - \max\{\sigma_{r+1}(\boldsymbol{X}), \sigma_{r+1}(\boldsymbol{Y})\}$. Then, we have*

$$\|\boldsymbol{X}_r - \boldsymbol{Y}_r\| \le \|\boldsymbol{X} - \boldsymbol{Y}\| \left(3 + \frac{\sigma_{r+1}(\boldsymbol{X}) + \sigma_{r+1}(\boldsymbol{Y})}{\delta}\right). \tag{193}$$

Next, we gather two lemmas regarding the singular values of Gaussian random matrices.

Lemma 19 (Adapted from Theorem 6.1 in Wainwright (2019)). *Suppose that $\boldsymbol{G} \in \mathbb{R}^{d_1\times d_2}$ is a standard Gaussian matrix, where $d_1 \ge d_2$. Then, it holds that*

$$\mathbb{P}\big(\|\boldsymbol{G}\| \ge 3\sqrt{d_1}\big) \le \exp\left(-d_1/2\right). \tag{194}$$

Lemma 20 (Adapted from Equation (3.2) in Rudelson and Vershynin (2010)). *Suppose that the entries of the $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ are i.i.d. standard Gaussian random variables. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\big(\sigma_{\min}(\boldsymbol{G}) \le \varepsilon d^{-1/2}\big) \le \varepsilon. \tag{195}$$

Finally, we show that for any two orthonormal matrices in $\mathcal{O}_{d\times r}$ with $r < d$, it is plausible to augment each into a square orthonormal matrix, without increasing their spectral-norm difference by much. See Section F.1 for the proof of this result.

Lemma 21. *Let $\boldsymbol{O}_1, \boldsymbol{O}_2 \in \mathcal{O}_{d\times r}$, where $r < d$. Then there exist $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathcal{O}_{d\times(d-r)}$ such that*

$$\boldsymbol{A}_1 := [\boldsymbol{O}_1, \boldsymbol{R}_1] \in \mathcal{O}_{d\times d}, \qquad \boldsymbol{A}_2 := [\boldsymbol{O}_2, \boldsymbol{R}_2] \in \mathcal{O}_{d\times d},$$
$$\text{and} \qquad \|\boldsymbol{A}_1 - \boldsymbol{A}_2\| \le \sqrt{2}\,\|\boldsymbol{O}_1 - \boldsymbol{O}_2\|.$$

F.1 Proof of Lemma 21

Denote by $\mathcal{S}_i := \mathsf{span}(\boldsymbol{O}_i)$ the r-dimensional subspace spanned by the columns of $\boldsymbol{O}_i$. Let $\theta_1, \ldots, \theta_r \in [0, \pi/2]$ represent the principal angles between $\mathcal{S}_1$ and $\mathcal{S}_2$ (see, e.g., Golub and Van Loan (2013, Chapter 6.4.3) and Chen et al. (2021, Section 2.2)). Define $\theta_{\max} := \max_{1\le i\le r} \theta_i$.

Step 1: computing distance between two subspaces. Consider any pair of orthonormal bases $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{d\times r}$ with $\mathsf{span}(\boldsymbol{Q}_i) = \mathcal{S}_i$. Classical matrix perturbation theory (e.g., Edelman et al. (1998, Section 4.3)) asserts that

$$\inf_{\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathcal{O}_{d\times r}:\, \mathsf{span}(\boldsymbol{Q}_1)=\mathcal{S}_1, \mathsf{span}(\boldsymbol{Q}_2)=\mathcal{S}_2} \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\| = 2\sin\Big(\frac{\theta_{\max}}{2}\Big), \tag{196}$$

thus implying that

$$2\sin\Big(\frac{\theta_{\max}}{2}\Big) \le \|\boldsymbol{O}_1 - \boldsymbol{O}_2\|. \tag{197}$$

Additionally, let $\mathcal{S}_i^{\perp}$ denote the $(d-r)$-dimensional orthogonal complement of $\mathcal{S}_i$. The maximum principal angle between $\mathcal{S}_1^{\perp}$ and $\mathcal{S}_2^{\perp}$ is again $\theta_{\max}$. This implies that

$$\inf_{\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathcal{O}_{d\times(d-r)}:\ \mathsf{span}(\boldsymbol{B}_1)=\mathcal{S}_1^{\perp}, \mathsf{span}(\boldsymbol{B}_2)=\mathcal{S}_2^{\perp}} \|\boldsymbol{B}_1 - \boldsymbol{B}_2\| = 2\sin\Big(\frac{\theta_{\max}}{2}\Big). \tag{198}$$

Step 2: choosing orthogonal complements with controlled distance. By Equation (198), one can find orthonormal bases $\boldsymbol{R}_1 \in \mathbb{R}^{d\times(d-r)}$ (resp. $\boldsymbol{R}_2 \in \mathbb{R}^{d\times(d-r)}$) of $\mathcal{S}_1^{\perp}$ (resp. $\mathcal{S}_2^{\perp}$) such that

$$\|\boldsymbol{R}_1 - \boldsymbol{R}_2\| = 2\sin\Big(\frac{\theta_{\max}}{2}\Big). \tag{199}$$

Combining Equations (197) and (199) yields

$$\|\boldsymbol{R}_1 - \boldsymbol{R}_2\| \le \|\boldsymbol{O}_1 - \boldsymbol{O}_2\|. \tag{200}$$

By construction, $\boldsymbol{A}_i := [\boldsymbol{O}_i, \boldsymbol{R}_i]$ forms a square orthogonal matrix.

Step 3: bounding the distance between $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. Observe that

$$\boldsymbol{A}_1 - \boldsymbol{A}_2 = [\boldsymbol{O}_1 - \boldsymbol{O}_2, \boldsymbol{R}_1 - \boldsymbol{R}_2].$$

For any $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \in \mathbb{R}^d$ with $\boldsymbol{x}_1 \in \mathbb{R}^r$ and $\boldsymbol{x}_2 \in \mathbb{R}^{d-r}$, it holds that

$$(\boldsymbol{A}_1 - \boldsymbol{A}_2)\boldsymbol{x} = (\boldsymbol{O}_1 - \boldsymbol{O}_2)\boldsymbol{x}_1 + (\boldsymbol{R}_1 - \boldsymbol{R}_2)\boldsymbol{x}_2.$$

This allows one to establish that

$$\begin{aligned}
\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|^2 &= \sup_{\|\boldsymbol{x}\|_2=1} \|(\boldsymbol{A}_1 - \boldsymbol{A}_2)\boldsymbol{x}\|^2 \\
&\leq \sup_{\|\boldsymbol{x}\|_2=1} \Big(\|\boldsymbol{O}_1 - \boldsymbol{O}_2\| \|\boldsymbol{x}_1\|_2 + \|\boldsymbol{R}_1 - \boldsymbol{R}_2\| \|\boldsymbol{x}_2\|_2 \Big)^2 \\
&\leq \sup_{\|\boldsymbol{x}\|_2=1} \Big(\|\boldsymbol{O}_1 - \boldsymbol{O}_2\| \|\boldsymbol{x}_1\|_2 + \|\boldsymbol{O}_1 - \boldsymbol{O}_2\| \|\boldsymbol{x}_2\|_2 \Big)^2 \quad \text{(by Equation (200))} \\
&= \|\boldsymbol{O}_1 - \boldsymbol{O}_2\|^2 \sup_{\|\boldsymbol{x}\|_2=1} \big(\|\boldsymbol{x}_1\|_2 + \|\boldsymbol{x}_2\|_2 \big)^2 \\
&\leq 2\|\boldsymbol{O}_1 - \boldsymbol{O}_2\|^2,
\end{aligned}$$

where the last line holds since, by Cauchy-Schwarz, $|\boldsymbol{x}_1\|_2 + \|\boldsymbol{x}_2\|_2 \leq \sqrt{2}\sqrt{\|\boldsymbol{x}_1\|_2^2 + \|\boldsymbol{x}_2\|_2^2} = \sqrt{2}\|\boldsymbol{x}\|_2^2$. This completes the proof.