MUON WITH NESTEROV MOMENTUM: HEAVY-TAILED NOISE AND (RANDOMIZED) INEXACT POLAR DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most first-order optimizers treat matrix-valued parameters as vectors, ignoring the intrinsic geometry of hidden-layer weights in neural networks. Muon addresses this mismatch by updating along the polar factor of a momentum matrix, but its theoretical understanding has lagged behind practice. In particular, practical implementations incorporate Nesterov momentum, compute the polar factor only approximately, and operate with stochastic gradients that may be heavy-tailed. We close this gap by developing a convergence theory for Muon with Nesterov momentum and inexact polar decomposition in non-convex matrix optimization under heavy-tailed noise. Our analysis builds on a unified framework for inexact polar decomposition that captures practical iterative approximations such as Newton-Schulz and quantifies how their errors propagate through the optimization dynamics. Under uniform bounds on polar errors, we establish an optimal iteration and sample complexity of $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$ for finding an ε-stationary point, where $\alpha \in (1, 2]$ denotes the heavy-tail index. For the inexact-polar setting with $\sigma_1 = 0$, we also provide guarantees that do not require prior knowledge of α. We analyze a randomized low-rank polar decomposition that is substantially more efficient than full-space methods while remaining compatible with our theory. Numerical experiments further demonstrate the effectiveness of the proposed inexact and randomized variants.

1 INTRODUCTION

Most widely used first-order optimizers [23; 33] still treat matrix-valued parameters as vectors, even though the weights of hidden linear layers are inherently structured matrices [42]. Recent norm-based perspectives suggest that this geometric choice is consequential, and that different tensor structures may benefit from different update norms [4]. Muon, introduced by [21], is a compelling realization of this idea: it replaces the raw momentum matrix with its orthogonalized polar factor before applying the update. For solving the optimization problem $\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} f(\boldsymbol{X})$, the update rule of Muon with Nesterov momentum is given by:

$$\boldsymbol{G}_k = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{G}_k^i, \quad \boldsymbol{C}_k = \beta \boldsymbol{C}_{k-1} + \boldsymbol{G}_k, \quad \boldsymbol{M}_k = \beta \boldsymbol{C}_k + \boldsymbol{G}_k, \quad \boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta \mathcal{T}(\boldsymbol{M}_k), \tag{1}$$

where $\beta \in (0, 1)$ and $\eta > 0$ are the momentum parameter and step size, respectively, B is the batch size, $\boldsymbol{G}_k^i$ is an estimate of the gradient $\nabla f(\boldsymbol{X}_k)$, and $\mathcal{T}(\boldsymbol{M}_k)$ is an approximation to the polar factor of $\boldsymbol{M}_k$. When computed exactly, $\mathcal{T}(\boldsymbol{M}_k) = \boldsymbol{U}_k \boldsymbol{V}_k^\top$ for the singular value decomposition $\boldsymbol{M}_k = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top$, so Muon replaces the raw momentum direction with its orthogonalized counterpart.

This orthogonalization step is central to the method, but it also constitutes the primary computational bottleneck. For large models, computing the exact polar factor of $\boldsymbol{M}_k$ is often prohibitively expensive. Consequently, practical implementations rely on a small number of iterations of fast matrix polynomial methods, such as Newton-Schulz [15; 21] and PolarExpress [2], to obtain $\mathcal{T}(\boldsymbol{M}_k)$, an efficient approximation of the polar factor. This discrepancy creates a noticeable gap between theory and practice. Existing analyses of Muon typically focus on either exact polar decomposition or Polyak-momentum variants [43; 5; 42; 44]. In contrast, the default PyTorch implementation employs Nesterov momentum together with an inexact polar decomposition [39; 36]. This practical variant is already being deployed at scale: Liu et al. [30] trained Moonlight, 3B and 16B parameter language models, using Muon and reported roughly $2\times$ computational efficiency gain over AdamW. More

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recently, Muon with Nesterov momentum and inexact polar decomposition has been used to train one of the largest open-source language models, DeepSeek-V4 [8].

A second mismatch between existing theory and modern practice lies in the assumptions on the stochastic gradients $\boldsymbol{G}_k^i$. Most optimization analyses assume bounded variance [28], while empirical evidence suggests that stochastic gradients in modern deep networks can exhibit heavy-tailed noise [45; 3]. Consequently, a realistic theory of Muon should simultaneously account for three ingredients: Nesterov momentum, inexact polar decomposition, and heavy-tailed stochastic gradients.

Our goal in this paper is to provide precisely such a theory for non-convex matrix optimization. We develop a unified framework for analyzing inexact polar decomposition (Assumption 2.1), and establish convergence guarantees (Theorem 3.5) for Muon with Nesterov momentum under heavy-tailed noise (Assumption 3.2). In Table 1, we provide a detailed comparison of our analysis with existing prior works on Muon.

Table 1: Comparisons of the theory of Muon across prior works. In the fourth column, $\sigma_1 \geq 0$ and $\alpha \in (1, 2]$ denote the heavy tail parameters in Assumption 3.2. The classical bounded-variance regime corresponds to $\sigma_1 = 0$ and $\alpha = 2$. The last column measures $\min_{0\leq k\leq K} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]$. Our rate uses the uniform polar bounds of Theorem 3.5, fixed batch size, and uniformly bounded initial error.

Paper	Nesterov Momentum	Inexact Polar Decomposition	Heavy Tailed Noise?	Convergence Rate
Riabinin et al. [41]	✗	✗	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
Kovalev [25]	✗	✗	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
Li and Hong [29]	✗	✗	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
Shen et al. [43]	✗	✗	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
Sato et al. [42]	✓	✗	$\sigma_1 = 0, \alpha = 2$	—
Chang et al. [5]	✓	✗	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(\log K/K^{1/4}\right)$
Shulgin et al. [44]	✗	✓	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
Kim and Oh [22]	✗	✓	$\sigma_1 = 0, \alpha = 2$	$\mathcal{O}\left(1/K^{1/4}\right)$
He et al. [14]	✗	✓	$\sigma_1 = 0, \alpha \in (1, 2]$	$\mathcal{O}\left(\log K/K^{(\alpha-1)/(3\alpha-2)}\right)$
Yu et al. [51]	✓	✗	$\sigma_1 \geq 0, \alpha \in (1, 2]$	$\mathcal{O}\left(1/K^{(\alpha-1)/(3\alpha-2)}\right)$
Ours (Theorem 3.5)	✓	✓	$\sigma_1 \geq 0, \alpha \in (1, 2]$	$\mathcal{O}\left(1/K^{(\alpha-1)/(3\alpha-2)}\right)$

Finally, we analyze a **randomized** low-rank polar decomposition procedure (Algorithm 1) that is significantly more efficient than full-space polar decomposition methods while remaining fully compatible with our theoretical framework. This algorithm exploits the widely observed phenomenon that the parameters of modern neural networks are often approximately low rank [13; 17; 14; 53]. A related low-rank polar decomposition method was recently proposed [14]; however, their analysis relies on a stopping criterion involving the computationally expensive quantity $\|\boldsymbol{M}_k - \mathcal{T}(\boldsymbol{M}_k)\|_*$.[1] In contrast, our analysis avoids this requirement altogether, thereby removing a key computational bottleneck and making the resulting method considerably more practical for large-scale settings.

Main Contributions. We summarize the main contributions of the paper below.

- *Heavy-tailed noise analysis for non-convex matrix optimization.* We provide a convergence analysis under relaxed assumptions on the stochastic gradients (Assumption 3.2), allowing for heavy-tailed noise beyond standard bounded-variance settings. This enables us to capture the behavior of Muon in modern deep learning applications. To our knowledge, this is the first analysis that simultaneously treats the specific Nesterov update in (1), generalized heavy-tailed noise with $\sigma_1 > 0$, and an inexact polar oracle.
- *Optimal convergence guarantee.* Under uniform polar-error bounds and uniformly bounded initial error, we establish optimal iteration and sample complexity for Muon with inexact polar decomposition and fixed batch size. Specifically, Theorem 3.5 implies an ε-stationarity complexity of $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$. To our knowledge, this is the first work that achieves optimal sample complexity for Muon with Nesterov momentum, even under bounded-variance assumption.
- *Unified inexact polar decomposition framework.* We develop a unified framework for inexact polar decomposition via Assumption 2.1, under which the Newton-Schulz iteration arises as a special case (see Proposition 2.2). This framework enables us to precisely quantify how approximation errors propagate through the optimization dynamics (see Theorem 3.5).

[1] $\|\cdot\|_*$ denotes the nuclear norm of a matrix.

- *Adaptivity to unknown tail index* $\alpha \in (1, 2]$. For the inexact-polar setting with $\sigma_1 = 0$ in Assumption 3.2, we establish convergence in Corollary 3.6 without requiring prior knowledge of the tail parameter α, yielding an adaptive method that is practically applicable when the distributional properties of stochastic gradients are unknown.
- *Randomized polar decomposition.* Inspired by [12], we analyze a randomized variant of polar decomposition (Algorithm 1) that enables efficient low-rank implementations. With Frobenius normalization and either theoretical polynomial, Lemma B.1 verifies the uniform polar bounds required by Theorem 3.5.
- *Numerical experiments.* We complement our theory on two standard benchmarks in Section 4: nanoGPT pretraining [37; 20] and CNN training on CIFAR-10 [26]. Randomized Muon performs comparably to the full-space Muon while reducing per-step optimizer FLOPs by roughly $8.5\times$ on nanoGPT ($2136 \to 251$ GFLOPs) and $1.8\times$ on CIFAR-10 ($14.80 \to 8.08$ GFLOPs). Additionally, replacing the dense Gaussian sketch with a coordinate (Kaczmarz) sketch yields a further reduction in both settings. We also measure a maximum realized alignment loss of 0.59 on the finite nanoGPT runs (Appendix E.6).

Related Works. Recent work has increasingly focused on geometry-aware and structure-exploiting optimization methods. For instance, Muon [21] orthogonalizes matrix-valued momentum using an approximate polar factor for hidden-layer updates. This perspective is naturally extended by the norm-constrained LMO framework underlying Scion [38]. Building on these insights, Gluon [41] unifies Muon and Scion, effectively bridging the gap between theoretical guarantees and layer-wise practical implementation. Alongside these methods, second-order approaches such as Shampoo [11] exploit tensor-structured preconditioning, which SOAP [49] further refines by incorporating Adam-style updates within Shampoo's eigenbasis to improve stability and efficiency. Taking a fundamentally different approach, Lion [6] employs a sign-based momentum discovered via symbolic search, deliberately prioritizing memory efficiency over complex matrix geometry.

2 (Randomized) Inexact Polar Decomposition

Given the SVD of the momentum matrix $\boldsymbol{M}_k = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top$, the exact polar factor is the solution to $\max_{\|\boldsymbol{O}\|_{\mathrm{op}} \le 1} \langle \boldsymbol{M}_k, \boldsymbol{O} \rangle$, and the maximum is attained at $\boldsymbol{O} = \boldsymbol{U}_k \boldsymbol{V}_k^\top$. However, computing this factor exactly is often prohibitively expensive in high-dimensional settings. Practical implementations, therefore, rely on iterative approximation schemes such as Newton-Schulz [15] and PolarExpress [2], which avoid an explicit SVD and instead repeatedly apply a polynomial map to a normalized matrix. Throughout this section, for an odd scalar polynomial p and a rectangular matrix $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the notation $p(\boldsymbol{Z})$ denotes the corresponding singular-value map $\boldsymbol{U}p(\boldsymbol{\Sigma})\boldsymbol{V}^\top$. Starting from $\boldsymbol{Z}_0 = \boldsymbol{M}_k/\delta$ with $\delta \ge \|\boldsymbol{M}_k\|_{\mathrm{op}}$ (practitioners often use $\delta = \|\boldsymbol{M}_k\|_F$), these methods generate iterates $\boldsymbol{Z}_{t+1} = \varphi(\boldsymbol{Z}_t), t = 0, 1, \ldots, q-1$, where φ is chosen to drive the singular values toward 1. Equivalently, if $p_0(x) = x$ and $p_{t+1}(x) = \varphi(p_t(x))$, then $\boldsymbol{Z}_q = p_q(\boldsymbol{M}_k/\delta)$ serves as a fast, hardware-friendly approximation of $\boldsymbol{U}_k \boldsymbol{V}_k^\top$. Two standard choices of polynomials are:

- Newton-Schulz with a degree-3 polynomial: $\varphi(x) = \frac{1}{2}(3x - x^3)$,
- Newton-Schulz with a degree-5 polynomial: $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$.

For a fixed polynomial φ, Muon implements $\mathcal{T}(\boldsymbol{M}_k) = p_q\left(\boldsymbol{M}_k/\delta\right)$ with a small number of iterations, typically q in $5 - 10$.

An inexact polar map should remain nearly feasible in operator norm and retain most of the optimal alignment with $\boldsymbol{M}_k$. Let $\mathcal{F}_k$ denote the history before drawing the current mini-batch. Here $\mathcal{F}_k$ contains the initialization and all earlier mini-batches and polar draws; in particular, $\boldsymbol{X}_k$ and $\boldsymbol{C}_{k-1}$ are $\mathcal{F}_k$-measurable. Define $\mathcal{G}_k := \sigma(\mathcal{F}_k, \boldsymbol{G}_k^1, \ldots, \boldsymbol{G}_k^B)$, the history after forming the current gradients and momentum but before the polar draw. Thus $\mathcal{F}_k \subseteq \mathcal{G}_k \subseteq \mathcal{F}_{k+1}$; Appendix A details this sampling order. We model the polar oracle as $\mathcal{T}(\boldsymbol{M}_k) = \mathcal{T}_k(\boldsymbol{M}_k; \zeta_k)$, where $\mathcal{T}_k$ is a fixed measurable map and the fresh seed ζ_k is independent of $\mathcal{G}_k$.

Assumption 2.1. We assume there exist deterministic bounds $\gamma_k \in [0, 1)$ and $\nu_k \ge 0$ such that

$$\mathbb{E}\left[\langle \boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \boldsymbol{M}_k\right] \ge (1 - \gamma_k)\|\boldsymbol{M}_k\|_* \qquad \text{and} \qquad \mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}}^2 \mid \boldsymbol{M}_k\right] \le (1 + \nu_k)^2.$$

Under this oracle model, the conditional law of $\mathcal{T}(\boldsymbol{M}_k)$ given $\mathcal{G}_k$ depends only on $\boldsymbol{M}_k$, so both bounds also hold when conditioning on $\mathcal{G}_k$. Assumption 2.1 recovers the exact polar decomposition $\mathcal{T}(\boldsymbol{M}_k) = \boldsymbol{U}_k\boldsymbol{V}_k^\top$ with $\gamma_k = \nu_k = 0$. Indeed, $\|\mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}} = \|\boldsymbol{U}_k\boldsymbol{V}_k^\top\|_{\mathrm{op}} = 1$, and

$$\langle \boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle = \langle \boldsymbol{M}_k, \boldsymbol{U}_k\boldsymbol{V}_k^\top\rangle = \mathrm{tr}\big(\boldsymbol{V}_k\boldsymbol{\Sigma}_k\boldsymbol{U}_k^\top\boldsymbol{U}_k\boldsymbol{V}_k^\top\big) = \mathrm{tr}(\boldsymbol{\Sigma}_k) = \|\boldsymbol{M}_k\|_*.$$

Moreover, the approximation condition $\|\mathcal{P}(\boldsymbol{M}_k) - \mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}} \le \varepsilon_k$ with $\varepsilon_k < 1$ (where $\mathcal{P}(\boldsymbol{M}_k) := \boldsymbol{U}_k\boldsymbol{V}_k^\top$) from [44] is a special case of Assumption 2.1. Indeed, we note that

$$\begin{aligned}\langle \boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle &= \langle \boldsymbol{M}_k, \mathcal{P}(\boldsymbol{M}_k)\rangle - \langle \boldsymbol{M}_k, \mathcal{P}(\boldsymbol{M}_k) - \mathcal{T}(\boldsymbol{M}_k)\rangle \\ &\ge \|\boldsymbol{M}_k\|_* - \|\boldsymbol{M}_k\|_*\|\mathcal{P}(\boldsymbol{M}_k) - \mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}} \ge (1-\varepsilon_k)\|\boldsymbol{M}_k\|_*,\end{aligned}$$

while the triangle inequality gives $\|\mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}} \le \|\mathcal{P}(\boldsymbol{M}_k)\|_{\mathrm{op}} + \|\mathcal{T}(\boldsymbol{M}_k) - \mathcal{P}(\boldsymbol{M}_k)\|_{\mathrm{op}} \le 1 + \varepsilon_k$. Beyond these assumptions, we show that the few-step Newton-Schulz updates [15] discussed above also fall within this framework.

Proposition 2.2. Suppose $\varphi(x) = \frac{1}{2}(3x - x^3)$ or $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$. Let $p_0(x) = x$ and $p_{t+1}(x) = \varphi(p_t(x))$ for $t = 0, 1, \ldots, q-1$. Let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be a compact SVD of a rank-r matrix with singular values $\sigma_1, \ldots, \sigma_r > 0$, and let $\delta \ge \|\boldsymbol{M}\|_{\mathrm{op}}$. Define

$$\mathcal{T}(\boldsymbol{M}) := p_q(\boldsymbol{M}/\delta) = \boldsymbol{U}\,\mathrm{diag}\left(p_q(\sigma_1/\delta), \cdots, p_q(\sigma_r/\delta)\right)\boldsymbol{V}^\top.$$

Then, $\mathcal{T}(\boldsymbol{M})$ satisfies Assumption 2.1 with $\nu = 0$ and $\gamma = 1 - \frac{\sum_{i=1}^r \sigma_i p_q(\sigma_i/\delta)}{\sum_{i=1}^r \sigma_i}$.

The proposition above shows that Assumption 2.1 captures the deterministic inexact polar decompositions used by Muon. A particularly relevant work is [22], which finds error bounds on finite-step Newton-Schulz, similar to Proposition 2.2, and provides an analysis of Muon with Polyak momentum under bounded variance noise. However, when the matrices are large, even a small number of Newton-Schulz steps requires several dense matrix multiplications in the full ambient space. To reduce this cost, we propose a randomized lifted variant that first compresses $\boldsymbol{M}$ into a low-dimensional subspace and then performs the inexact polar computation in that reduced space.

Algorithm 1 Lifted Randomized Polar Decomposition

Require: Matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, sketch size $1 \le \ell \le \min(m,n)$, integers $h, q \ge 0$, and scaling $\delta \ge \|\boldsymbol{M}\|_{\mathrm{op}}$ for $\boldsymbol{M} \ne \boldsymbol{0}$.

If $\boldsymbol{M} = \boldsymbol{0}$, return $\boldsymbol{0}$; for the uniform guarantee, set $\delta = \|\boldsymbol{M}\|_F$.

Draw a fresh Gaussian sketch $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times\ell}$, independently of the input and prior history;	**Project Down**
Form $\boldsymbol{Y} := (\boldsymbol{M}\boldsymbol{M}^\top)^h\boldsymbol{M}\boldsymbol{\Omega}$; let $\boldsymbol{Q} := \mathrm{orth}(\boldsymbol{Y})$ span its actual range;	
Set $\boldsymbol{B} := \boldsymbol{Q}^\top\boldsymbol{M} \in \mathbb{R}^{\mathrm{rank}(\boldsymbol{Y})\times n}$;	
Set $\boldsymbol{Z}_0 := \boldsymbol{B}/\delta$;	**Iterative Approximation**
for $t = 0, 1, \ldots, q-1$ **do**	
$\quad\boldsymbol{Z}_{t+1} := \varphi(\boldsymbol{Z}_t)$;	
end for	
Return $\mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega}) := \boldsymbol{Q}\boldsymbol{Z}_q$.	**Project Up**

Algorithm 1 draws a Gaussian sketch $\boldsymbol{\Omega} \in \mathbb{R}^{n\times\ell}$ (typically $\ell = s + p$) and forms $\boldsymbol{Y} := (\boldsymbol{M}\boldsymbol{M}^\top)^h\boldsymbol{M}\boldsymbol{\Omega}$ to capture the dominant column space. Then, an orthonormal basis $\boldsymbol{Q} := \mathrm{orth}(\boldsymbol{Y})$ is computed via a QR decomposition and used to form $\boldsymbol{B} := \boldsymbol{Q}^\top\boldsymbol{M}$, with $\mathrm{rank}(\boldsymbol{Y}) \le \ell$ rows. We then compute an inexact polar decomposition of $\boldsymbol{B}$, and lift the result back via multiplication with $\boldsymbol{Q}$. The randomness of the method is induced solely by the Gaussian sketch $\boldsymbol{\Omega}$, which is drawn afresh independently of $\mathcal{G}_k$ at every optimizer iteration k. Its computational efficiency arises from performing the inexact polar decomposition in the reduced ℓ-dimensional subspace. Our approach builds on the randomized sketching framework of [12]. As in this work, Algorithm 1 first constructs a low-dimensional sketch of the range of the input matrix. The key difference is that we avoid computing an SVD of the compressed matrix; instead, we obtain an approximate polar factor through inexpensive iterative Newton-Schulz updates.

The compression induced by the sketching matrix yields a significant computational advantage. Compared with applying Newton-Schulz directly in the full space, which for $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ incurs a cost of $\mathcal{O}\left(q(4d_1d_0^2 + 2d_0^3)\right)$ with $d_0 := \min(m,n)$ and $d_1 := \max(m,n)$, our method first compresses $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ into $\boldsymbol{B} \in \mathbb{R}^{\ell\times n}$, where $\ell = s+p \ll d_0$, and then performs the same iteration only in this ℓ-dimensional subspace. The resulting complexity becomes $\mathcal{O}\left((4h+6)mn\ell + q(4n\ell^2 + 2\ell^3)\right)$, so the cubic dependence on d_0 is replaced by a cubic dependence on the much smaller sketch size ℓ. In the square case $m = n = d$, this reduces the scaling from $\mathcal{O}(qd^3)$ to $\mathcal{O}\left((4h+6)d^2\ell + q(4d\ell^2 + 2\ell^3)\right)$. For instance, with $(d, \ell, q, h) = (4096, 256, 5, 1)$, the factorization cost is reduced by roughly $40\times$.

The next proposition shows that the randomized lifted scheme (Algorithm 1) remains spectrally bounded and preserves a controlled amount of alignment with the momentum matrix $\boldsymbol{M}$.

Proposition 2.3. Let $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ have rank r with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Let $1 \leq s < r$, let $p \geq 2$ and $\ell = s + p \leq \min(m,n)$, and let h, q be nonnegative integers. Define $\varrho_s := {}^{\sigma_{s+1}}/_{\sigma_s} \in [0,1]$. For any $\delta \geq \|\boldsymbol{M}\|_{\mathrm{op}}$, the output $\mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})$ from Algorithm 1, with $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times\ell}$ and $\varphi(x) = \frac{1}{2}(3x - x^3)$ or $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$, satisfies

$$\mathbb{E}_{\boldsymbol{\Omega}}\left[\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})\rangle\right] \geq \frac{1}{\delta}\left[\sum_{j=1}^{s}\sigma_j^2 - \frac{s}{p-1}\varrho_s^{4h}\sum_{j>s}\sigma_j^2\right]_+.$$

Moreover, almost surely, $\left\|\mathcal{T}_{h,q}\left(\boldsymbol{M};\boldsymbol{\Omega}\right)\right\|_{\mathrm{op}} \leq 1$.

Proposition 2.3 gives a spectrum-dependent alignment bound; Appendix B.4 discusses when its bracket is positive. For the convergence rates, we also need alignment bounded away from zero uniformly over iterations and horizons. With $\delta = \|\boldsymbol{M}\|_F$ for $\boldsymbol{M} \neq \boldsymbol{0}$ and $\mathcal{T}(\boldsymbol{0}) = \boldsymbol{0}$, Lemma B.1 proves that either theoretical polynomial admits uniform bounds $\bar{\nu} = 0$ and $\bar{\gamma} = 1 - d_0^{-1/2}$ in full space, or $\bar{\gamma} = 1 - d_0^{-3/2}$ for Algorithm 1, where $d_0 = \min(m,n)$. These conservative bounds require no spectral gap and verify the uniform polar conditions in Theorem 3.5. Other polynomial variants require separate verification. Our experiments also use a faster Kaczmarz-style sparse sketch, described in Appendix E.1.

3 Main Theoretical Results

In this section, we establish convergence guarantees for Muon with Nesterov momentum under heavy-tailed stochastic gradient noise. Our analysis proceeds in two steps. We first study an idealized version of Muon with exact polar decomposition, which isolates the effects of the momentum dynamics and heavy-tailed noise. We then extend the argument to inexact polar decompositions, capturing practical implementations based on approximate polar iterations. Throughout, the initial iterate $\boldsymbol{X}_0$ is deterministic; let $f_\star$ denote a finite lower bound on f, let $d_0 := \min\{m,n\}$ for $\boldsymbol{X} \in \mathbb{R}^{m\times n}$. We use the histories $\mathcal{F}_k$ and $\mathcal{G}_k$ defined in Section 2; in particular, $\boldsymbol{X}_k$ is $\mathcal{F}_k$-measurable. Unless otherwise stated, $\mathcal{O}(\cdot)$ hides constants depending only on α and d_0. We assume that f is L-smooth as stated in the following assumption.

Assumption 3.1. We assume that f has an L-Lipschitz gradient in the Frobenius norm, that is, $\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\|_F \leq L\|\boldsymbol{X} - \boldsymbol{Y}\|_F$ for all $\boldsymbol{X}, \boldsymbol{Y}$. In particular,

$$f(\boldsymbol{Y}) \leq f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X}\rangle + \frac{L}{2}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2.$$

Assumption 3.1 is standard in the analysis of first-order methods [32; 50; 35] and is also commonly adopted in recent analyses of Muon [42; 5; 43].

Existing analyses of Muon typically focus on a bounded-variance condition, $\mathbb{E}[\|\boldsymbol{G}_k^i - \nabla f(\boldsymbol{X}_k)\|_F^2 \mid \mathcal{F}_k] \leq \sigma_0^2$ (see [42; 5]). To better accommodate the training regimes encountered in modern deep learning, we instead adopt the following generalized moment assumption.

Assumption 3.2 (*Generalized Heavy-Tailed Noise*)**.** Conditional on $\mathcal{F}_k$, we assume that $\boldsymbol{G}_k^1, \ldots, \boldsymbol{G}_k^B$ are independent, and satisfy $\mathbb{E}\left[\boldsymbol{G}_k^i \mid \mathcal{F}_k\right] = \nabla f(\boldsymbol{X}_k)$ and

$$\mathbb{E}\left[\|\boldsymbol{G}_k^i - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \mathcal{F}_k\right] \leq \sigma_0^\alpha + \sigma_1^\alpha\|\nabla f(\boldsymbol{X}_k)\|_F^\alpha$$

with $\sigma_0, \sigma_1 \geq 0$ and $\alpha \in (1,2]$ for all $i \in [B]$.

Assumption 3.2 covers several familiar regimes. When $\alpha = 2$ and $\sigma_1 = 0$, it reduces to the standard bounded-variance condition. When $\alpha \in (1, 2)$, it accommodates heavy-tailed stochastic gradients, while the term involving σ_1 further allows the noise to grow with the gradient magnitude. Related assumptions have been studied for vector-valued problems in [31]; the special case $\sigma_1 = 0$ recovers the α-bounded moment assumption considered in prior works [18; 52; 7]. To the best of our knowledge, [14] is the only prior work that analyzes Muon with $\alpha \in (1, 2]$, but it focuses on Polyak momentum and $\sigma_1 = 0$. Moreover, [31] provides a simple counterexample showing that Assumption 3.2 with $\sigma_1 = 0$ may fail to hold even for 1-dimensional regression problems, thereby highlighting the need for $\sigma_1 > 0$ in practical settings. The generalization to $\alpha \in (1, 2]$ is particularly relevant in light of empirical evidence for heavy-tailed gradient noise in both image classification and large language model training [45; 3; 52; 1]. Finally, Assumption 3.2 yields a bound on the noise level of $\boldsymbol{G}_k$ (Lemma A.11) that decreases monotonically with the batch size B. This shows that mini-batching reduces the effective noise level, and will be used in both the exact and inexact analyses below.

3.1 Warm-up: Analysis for Exact Polar Decomposition

We begin with the idealized setting in which the polar decomposition is computed exactly, namely $\mathcal{T}(\boldsymbol{M}_k) = \boldsymbol{U}_k \boldsymbol{V}_k^\top$ whenever $\boldsymbol{M}_k = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top$ is the SVD of $\boldsymbol{M}_k$. This setting isolates the effect of Nesterov momentum without the additional complication of approximation error in the polar step.

It is convenient to rescale the momentum matrices by $(1-\beta)$, i.e., $\widetilde{\boldsymbol{C}}_k := (1-\beta)\boldsymbol{C}_k$ and $\widetilde{\boldsymbol{M}}_k := (1-\beta)\boldsymbol{M}_k$. This rescaling does not change the polar factor, since multiplying a matrix by a positive scalar only rescales its singular values. Thus, $\mathcal{T}(\widetilde{\boldsymbol{M}}_k) = \mathcal{T}(\boldsymbol{M}_k) = \boldsymbol{U}_k \boldsymbol{V}_k^\top$. However, $\widetilde{\boldsymbol{M}}_k$ satisfies a nice simple recursion.

Lemma 3.3. For $\widetilde{\boldsymbol{C}}_k := (1-\beta)\boldsymbol{C}_k$ and $\widetilde{\boldsymbol{M}}_k := (1-\beta)\boldsymbol{M}_k$, we have, for all $k \geq 1$,

$$\widetilde{\boldsymbol{C}}_k = \beta \widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)\boldsymbol{G}_k \qquad \text{and} \qquad \widetilde{\boldsymbol{M}}_k = \beta \widetilde{\boldsymbol{M}}_{k-1} + (1-\beta)\big((1+\beta)\boldsymbol{G}_k - \beta \boldsymbol{G}_{k-1}\big).$$

Our analysis begins with a descent inequality for the exact-polar update (Lemma C.1). Following standard Muon analyses [43; 5], this bound depends on the momentum error term $\boldsymbol{S}_k = \widetilde{\boldsymbol{M}}_k - \nabla f(\boldsymbol{X}_k)$. The main technical challenge is therefore to control $\boldsymbol{S}_k$, which is not itself a martingale difference because of Nesterov momentum. We address this by unrolling the recursion for $\boldsymbol{S}_k$ into three components: the initial bias $\boldsymbol{S}_0$, a drift term $\boldsymbol{D}_k = \nabla f(\boldsymbol{X}_{k-1}) - \nabla f(\boldsymbol{X}_k)$, and a geometrically weighted sum of stochastic gradient noises $\boldsymbol{\xi}_k = \boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)$ (Lemma C.2). Smoothness bounds the drift, while Lemma A.11, the martingale inequality in Lemma A.8, and the conditional weighted estimate in Lemma A.9 control the noise terms. The latter estimate handles the gradient-dependent noise using only first moments of the gradient norm. Finally, for Muon with Nesterov momentum and exact polar decomposition, we establish the following convergence guarantee under heavy-tailed noise.

Theorem 3.4. Suppose Assumptions 3.1 and 3.2 hold. For a horizon $K \geq 2$, Muon (1) with exact polar decompositions, step size $\eta = 1/K^{(2\alpha-1)/(3\alpha-2)}$, momentum parameter $\beta = 1 - 1/K^{\alpha/(3\alpha-2)}$, and batch size $B > \{2\sqrt{\pi}(1+\sqrt{d_0})\mathcal{C}_\alpha^{1/\alpha}\sigma_1\}^{\alpha/\alpha-1}$ satisfies

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Big(\frac{f(\boldsymbol{X}_0)-f_\star}{\rho_{\text{ex}} K^{\alpha-1/3\alpha-2}} + \frac{L}{\rho_{\text{ex}} K^{\alpha-1/3\alpha-2}} + \frac{\sigma_0}{\rho_{\text{ex}} B^{\alpha-1/\alpha} K^{\alpha-1/3\alpha-2}} + \frac{L}{\rho_{\text{ex}} K^{2\alpha-1/3\alpha-2}} + \frac{\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho_{\text{ex}} K^{2\alpha-2/3\alpha-2}} + \frac{\sigma_1 \|\nabla f(\boldsymbol{X}_0)\|_F}{\rho_{\text{ex}} B^{\alpha-1/\alpha} K}\Big), \quad (2)$$

where $\rho_{\text{ex}} = 1 - (1+\sqrt{d_0})2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1/B^{\alpha-1/\alpha} > 0$ and $0 < \mathcal{C}_\alpha \leq 2$ is a scalar depending only on α.

The choices of η and β do not require the smoothness constant L. For all asymptotic rates below, fix the problem parameters and $\boldsymbol{X}_0$, take B independent of K and the target accuracy, and assume $\mathbb{E}[\|\boldsymbol{S}_0\|_F]$ is bounded uniformly in K. With B satisfying the batch size condition, the dominant term on the right side of (2) is $\mathcal{O}(1/K^{(\alpha-1)/(3\alpha-2)})$. This aligns with the concurrent work of [51] that only studies exact polar decomposition. Moreover, to obtain an ε-accurate point, the algorithm requires $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$ iterations, which match the optimal complexity for non-convex minimization under heavy-tailed noise [52; 51]. In particular, when $\alpha = 2$, this bound recovers the $\mathcal{O}(\varepsilon^{-4})$

complexity, which is known to be optimal under the standard bounded-variance setting [27; 10]. Compared with $\mathcal{O}\left(\log K/K^{1/4}\right)$ rate obtained for the Nesterov variant under bounded variance in [5, Theorem 3.1(1)], our result covers the heavy-tailed noise regime ($\alpha \in (1,2), \sigma_1 > 0$) and yields a faster asymptotic rate in the special case $\alpha = 2, \sigma_1 = 0$.

The lower bound on B is required only because we allow more general regime $\sigma_1 > 0$ in Assumption 3.2. When $\sigma_1 = 0$, no such restriction is needed, and even $B = 1$ is admissible. Therefore, a large batch size is not a limitation of our approach. This dependence of B on σ_1 aligns with the findings of [31] for the $\sigma_1 > 0$ regime. Notably, since the batch size threshold is independent of K, the sample complexity has the same order as the iteration complexity for fixed B, i.e., $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$.

3.2 Analysis for Inexact Polar Decomposition

We now turn to the practically relevant setting in which the polar decomposition is computed only approximately, captured by Assumption 2.1. The proof follows a similar overall structure to the exact case, but the descent argument must now account for the loss of alignment between the momentum matrix $\widetilde{\boldsymbol{M}}_k$ and the approximate polar factor $\mathcal{T}(\boldsymbol{M}_k)$. The key observation is that Assumption 2.1 implies $\mathbb{E}[\langle \widetilde{\boldsymbol{M}}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \mathcal{G}_k] \geq (1-\gamma_k)\|\widetilde{\boldsymbol{M}}_k\|_*$, because $\widetilde{\boldsymbol{M}}_k = (1-\beta)\boldsymbol{M}_k$. Thus, after taking outer expectations, the inexact polar step still preserves a fraction of the alignment achieved in the exact case. Together with the control of the momentum drift as before, this yields the following result.

Theorem 3.5. Suppose Assumptions 2.1, 3.1, and 3.2 hold. Assume deterministic constants $\bar{\gamma} \in [0,1)$ and $0 \leq \bar{\nu} < \infty$, independent of the iteration k and horizon K, satisfy $\gamma_k \leq \bar{\gamma}$ and $\nu_k \leq \bar{\nu}$ in every run considered. Then, for any $\eta > 0$ and $\beta \in (0,1)$, Muon (1) with inexact polar decomposition and fixed batch size $B > \left\{2\sqrt{\pi}\left(1+\sqrt{d_0}(1+\bar{\nu})\right)\mathcal{C}_\alpha^{1/\alpha}\sigma_1/(1-\bar{\gamma})\right\}^{\alpha/\alpha-1}$ satisfies

$$\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(\frac{f(\boldsymbol{X}_0)-f_\star}{\rho\eta K} + \frac{\beta^2 L(1+\bar{\nu})^2\eta}{\rho(1-\beta)} + \frac{(1+\bar{\nu})\sigma_0(1-\beta)^{\alpha-1/\alpha}}{\rho B^{\alpha-1/\alpha}} + \frac{L\eta(1+\bar{\nu})^2}{\rho} + \frac{(1+\bar{\nu})\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho(1-\beta)K} + \frac{(1+\bar{\nu})\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho B^{\alpha-1/\alpha}K}\Bigg),$$

where $\rho := 1-\bar{\gamma}-\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1/B^{\alpha-1/\alpha} > 0$ and $\mathcal{C}_\alpha \in (0,2]$ depends only on α.

The uniform constants $\bar{\gamma}, \bar{\nu}$ and fixed B ensure that $\rho > 0$ is independent of K. Larger polar-error bounds increase the required batch size and worsen the constants. Per-iteration bounds $\gamma_k < 1$ alone do not ensure a rate if their maxima approach one as K grows. Under these uniform bounds and the initial-error condition stated above, choosing $\eta = K^{-(2\alpha-1)/(3\alpha-2)}$ and $\beta = 1 - K^{-\alpha/(3\alpha-2)}$ gives $\mathcal{O}(K^{-(\alpha-1)/(3\alpha-2)})$ convergence (dominant terms are highlighted in red). Thus the iteration and sample complexity is $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$, since B is fixed independently of accuracy. Lemma B.1 supplies the required uniform polar bounds $\bar{\gamma}, \bar{\nu}$ for the full-space and Gaussian constructions with Frobenius normalization.

Convergence Guarantees without the Knowledge of α. The parameter choices in the previous theorems depend on the heavy-tail parameter α in Assumption 3.2, which may be difficult to estimate in practice. We therefore also provide a universal schedule that guarantees convergence without requiring knowledge of α in the inexact-polar setting with $\sigma_1 = 0$.

Corollary 3.6. Suppose the assumptions and uniform polar bounds of Theorem 3.5 hold with $\sigma_1 = 0$. Set $\rho_0 := 1 - \bar{\gamma} > 0$, independent of K. For a horizon $K \geq 2$, Muon (1) with inexact polar decomposition, step size $\eta = \frac{1}{K^{3/4}}$, momentum parameter $\beta = 1 - \frac{1}{\sqrt{K}}$, and any batch size $B \geq 1$ satisfies

$$\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(\frac{f(\boldsymbol{X}_0)-f_\star}{\rho_0 K^{\frac{1}{4}}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{3}{4}}} + \frac{(1+\bar{\nu})\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho_0\sqrt{K}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{1}{4}}} + \frac{(1+\bar{\nu})\sigma_0}{\rho_0 B^{\frac{\alpha-1}{\alpha}}K^{\frac{\alpha-1}{2\alpha}}}\Bigg).$$

Corollary 3.6 removes the need to know α when choosing the parameters η and β. In the regime $\sigma_1 = 0$, the batch size can also be chosen arbitrarily, so no algorithmic parameter depends on α. Under the uniform polar bounds and the initial-error condition stated above, the rate is $\mathcal{O}(K^{-(\alpha-1)/2\alpha})$,

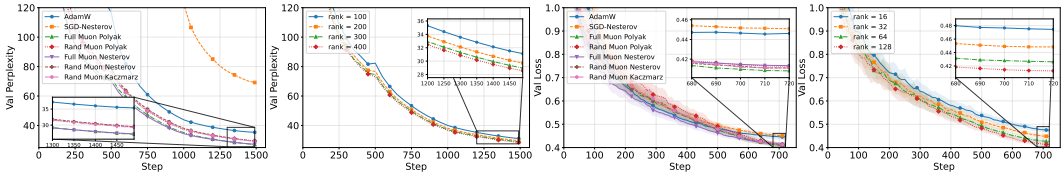
which leads to an iteration complexity of $\mathcal{O}(\varepsilon^{-2\alpha/(\alpha-1)})$. However, when $\alpha = 2$, this recovers the optimal $\mathcal{O}(\varepsilon^{-4})$ complexity.

4 Numerical Experiments

In this section, we present numerical experiments evaluating randomized Muon with Nesterov momentum against six baselines: AdamW, SGD with Nesterov momentum, Muon with Polyak momentum, Muon with Nesterov momentum, randomized Muon with Polyak momentum, and randomized Muon with Nesterov momentum, but Kaczmarz sketch. We consider two standard benchmarks: pretraining a nanoGPT model on the FineWeb dataset [37] and training a CNN on CIFAR-10 [26]. For each benchmark, we include additional ablation studies in Appendix E.4 and E.5 to examine the sensitivity of randomized Muon with Nesterov momentum. Appendix E.6 reports a maximum realized alignment loss of 0.59 on the finite nanoGPT runs; these measurements do not certify uniform bounds over all horizons. Full experimental setup details for both benchmarks are deferred to Appendix E.

4.1 Pretraining of nanoGPT on FineWeb10B

We pretrain a 135M-parameter nanoGPT [20] on FineWeb10B [37] using the `modded-nanogpt` speedrun codebase of [20], following the setup adopted in recent analyses of inexact and low-rank Muon [44; 14]. We report validation perplexity on the FineWeb10B validation set [40; 48; 47] to amplify small differences in validation loss. Each configuration is averaged over 5 random seeds. Table 2 reports the experimental results, and Figure 1 plots the convergence trajectories for the optimizer comparison and the effect of varying the randomized rank. Full implementation details are deferred to Appendix E.2.



(a) nanoGPT: optimizer (b) nanoGPT: rank s (c) CIFAR-10: optimizer (d) CIFAR-10: rank s

Figure 1: Convergence trajectories on the two benchmarks. (a,b) nanoGPT: optimizer comparison and randomized Muon (Nesterov) at varying randomized rank s, reporting validation perplexity. (c,d) CIFAR-10: same comparison reporting validation loss.

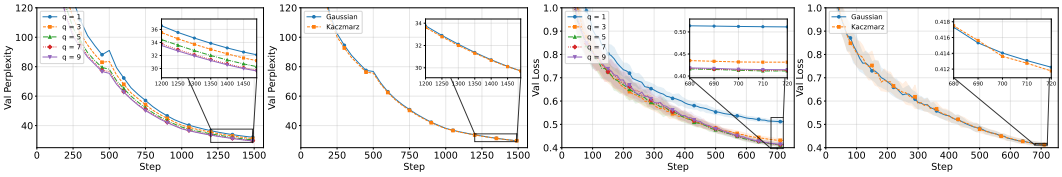
Figure 1a shows that all Muon variants clearly outperform AdamW and SGD-Nesterov. Moreover, the randomized Muon variants perform comparably to their full-space counterparts under both Polyak and Nesterov momentum. As reported in Table 2, the gap between full and randomized Muon is reasonable under both Polyak and Nesterov momentum. This comparable result is obtained with a randomized rank of only 200, about 26% of the corresponding full rank. Figure 1b further shows that validation perplexity decreases monotonically as the randomized rank grows. Even at rank 400, the sketch uses only about 52% of the corresponding full rank while further reducing validation perplexity. Figure 2a additionally shows that increasing the Newton-Schulz or PolarExpress inner-iteration count q improves final perplexity monotonically, though the gains diminish beyond $q = 7$.

Despite the comparable perplexity gap, the reduction in optimizer FLOPs is substantial. To quantify this, we report per-step FLOPs [16; 34; 8] in Table 2. Since the forward and backward passes are identical across methods, we focus on the optimizer-update components, which capture all optimizer-specific steps. Full-space Muon requires roughly 2136 GFLOPs per step (where 1 GFLOP $= 10^9$ FLOPs), whereas randomized Muon requires only about 251 GFLOPs, which is nearly an order-of-magnitude reduction. The Polyak variant is marginally cheaper than Nesterov since Nesterov adds one extra vector combination in the momentum update.

The cost can be further reduced by using a canonical-basis sketch rather than a dense Gaussian sketch, so that each column of $\boldsymbol{\Omega}$ has a single nonzero entry. Then $\boldsymbol{M}\boldsymbol{\Omega}$ reduces to a rescaled column selection rather than a dense matrix-matrix multiplication. With this Kaczmarz-style sparse sketch, the per-step optimizer cost drops to 217 GFLOPs while keeping almost the same validation perplexity. As we can see from Figure 2b and Table 2, the Kaczmarz-style drawing achieves nearly identical final perplexity

Table 2: Final validation perplexity for nanoGPT, test accuracy for CIFAR-10, wall-clock time, and per-step optimizer GFLOPs. Best metric, the lowest wall-clock time and GFLOPs among Muon variants are in **bold**.

Method	nanoGPT			CIFAR-10		
	Val. Perplexity	Time (s)	GFLOPs	Test Accuracy	Time (s)	GFLOPs
AdamW	35.3749 ± 0.4474	1227.8989 ± 4.4531	7.2621	0.9179 ± 0.0025	1.7958 ± 0.0066	0.0237
SGD (Nes.)	71.0208 ± 1.8195	1669.6667 ± 17.5563	3.1123	0.9212 ± 0.0014	1.7373 ± 0.0054	0.0118
Muon (Polyak)	$\mathbf{27.1115 \pm 0.0246}$	$\mathbf{1205.9631 \pm 5.0362}$	2135.5947	$\mathbf{0.9376 \pm 0.0012}$	$\mathbf{4.8547 \pm 0.0363}$	14.7920
Rand Muon (Polyak)	29.5218 ± 0.0287	1229.1785 ± 4.5745	250.8060	0.9363 ± 0.0015	7.0608 ± 0.0284	8.0723
Muon (Nes.)	27.2457 ± 0.0201	1208.5847 ± 5.8304	2135.7551	0.9368 ± 0.0012	4.9300 ± 0.0225	14.7960
Rand Muon (Nes.-Gauss.)	29.4565 ± 0.0482	1216.2849 ± 13.4836	250.9665	0.9364 ± 0.0013	7.1153 ± 0.0230	8.0763
Rand Muon (Nes.-Kacz.)	29.2445 ± 0.0348	1237.2930 ± 4.8522	**217.4581**	0.9366 ± 0.0011	8.1434 ± 0.0515	**7.5373**



(a) nanoGPT: q (b) nanoGPT: Sketching (c) CIFAR-10: q (d) CIFAR-10: Sketching

Figure 2: (a,b) nanoGPT: validation perplexity when varying the Newton-Schulz or PolarExpress inner-iteration count q and the randomized sketching. (c,d) CIFAR-10: validation loss for the same two ablations.

to the Gaussian drawing and also has almost the same rate of decreasing while further reducing per-step computational cost. The sparse drawing scheme is described in detail in Algorithm 2.

4.2 CNN ON CIFAR-10

To assess randomized Muon on a different data modality, we train a small CNN on CIFAR-10 [26] using the CIFAR-10 Airbench codebase of [19]. Each configuration is averaged over 50 random seeds. Table 2 reports the experimental results, and Figure 1 plots the validation-loss convergence trajectories. Full implementation details are deferred to Appendix E.3.

Figure 1c and Table 2 show that the Muon variants reach essentially the same test accuracy and outperform AdamW and SGD-Nesterov. The full and randomized variants are nearly indistinguishable on this benchmark. The convergence curves of Figure 1c are correspondingly indistinguishable across Muon variants. This narrow separation may be due to the limited capacity of the compact CNN, which restricts the potential gains from orthogonalization-based optimization over well-tuned baselines. Figure 1d shows that validation loss decreases monotonically as the randomized rank grows from $s = 16$ to $s = 128$, with diminishing returns at larger ranks. Figure 2c similarly shows that increasing the Newton-Schulz or PolarExpress inner-iteration count q improves final loss, though the gains are marginal beyond $q = 5$.

Following the same FLOP-counting convention as in Section 4.1, we focus on the per-step optimizer GFLOPs reported in Table 2. Full-space Muon spends roughly 14.79 GFLOPs per step, whereas randomized Muon drops to about 8.07 GFLOPs, approximately a 45% reduction. The canonical-basis (Kaczmarz) sketch lowers the cost further to 7.54 GFLOPs while attaining a marginally higher test accuracy than the Gaussian sketch. Figure 2d and Table 2 show that the Kaczmarz-style drawing not only reduces computational cost but also attains a marginally lower final validation loss than the Gaussian drawing. As in the nanoGPT experiment, the Polyak variant is marginally cheaper than Nesterov.

5 CONCLUSION

We developed a convergence theory for Muon that incorporates Nesterov momentum, inexact polar decomposition, and heavy-tailed stochastic gradients in non-convex matrix optimization. Our analysis introduces a unified inexact-polar framework that captures practical approximations such as Newton-Schulz, while quantifying how approximation errors influence the optimization dynamics. Under generalized heavy-tailed noise, uniform polar-error bounds, and uniformly bounded initial error, we established an optimal $\mathcal{O}(\varepsilon^{-(3\alpha-2)/(\alpha-1)})$ iteration and sample complexity for finding an ε-stationary point. Beyond full-space polar decomposition, we further analyzed a randomized low-rank polar procedure that remains compatible with our theory and substantially reduces optimizer-side computational cost. Experiments on nanoGPT pretraining and CIFAR-10 demonstrated the practical effectiveness of the proposed inexact and randomized variants.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript. We have reviewed all AI-assisted text and take responsibility for the final manuscript, including text, claims, proofs, and experimental results.

REFERENCES

[1] K. Ahn, X. Cheng, M. Song, C. Yun, A. Jadbabaie, and S. Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *International Conference on Learning Representations*, 2024.

[2] N. Amsel, D. Persson, C. Musco, and R. M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *International Conference on Learning Representations*, 2026.

[3] B. Battash, L. Wolf, and O. Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788. PMLR, 2024.

[4] J. Bernstein and L. Newhouse. Old optimizer, new norm: An anthology. *NeurIPS Workshop on Optimization for Machine Learning*, 2024.

[5] D. Chang, Y. Liu, and G. Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

[6] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

[7] S. Chezhegov, Y. Klyukin, A. Semenov, A. Beznosikov, A. Gasnikov, S. Horváth, M. Takáč, and E. Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *International Conference on Machine Learning*, 2025.

[8] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[9] P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, Jan. 2006.

[10] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

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

[12] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[13] Y. Hao, Y. Cao, and L. Mou. Flora: Low-rank adapters are secretly gradient compressors. *International Conference on Machine Learning*, 2024.

[14] C. He, Z. Deng, and Z. Lu. Low-rank orthogonalization for large-scale matrix optimization with applications to foundation model training. *arXiv preprint arXiv:2509.11983*, 2025.

[15] N. J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

[16] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.

[18] F. Hübler, I. Fatkhullin, and N. He. From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*, 2024.

[19] K. Jordan. 94% on CIFAR-10 in 3.29 seconds on a single gpu. *arXiv preprint arXiv:2404.00498*, 2024.

[20] K. Jordan, J. Bernstein, B. Rappazzo, V. Boza, J. You, F. Cesista, and B. Koszarsky. Modded-NanoGPT: Speedrunning the NanoGPT baseline. https://github.com/KellerJordan/modded-nanogpt, 2024. GitHub repository, accessed 2026-05-01.

[21] K. Jordan, Y. Jin, V. Boza, Y. Jiacheng, F. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

[22] G. Y. Kim and M.-h. Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pages 45672–45718, 2026.

[23] D. P. Kingma. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[24] N. Kornilov, O. Shamir, A. Lobanov, D. Dvinskikh, A. Gasnikov, I. Shibaev, E. Gorbunov, and S. Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36: 64083–64102, 2023.

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

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[27] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[28] G. Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[29] J. Li and M. Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

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

[31] Z. Liu and Z. Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. *International Conference on Learning Representations*, 2025.

[32] N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International conference on artificial intelligence and statistics*, pages 1306–1314. PMLR, 2021.

[33] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

[34] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. Efficient large-scale language model training on GPU clusters using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

[35] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[37] G. Penedo, H. Kydlíček, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, T. Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[38] T. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained lmos. *International Conference on Machine Learning*, 2025.

[39] PyTorch Contributors. PyTorch documentation: torch.optim.Muon, 2026. URL https://docs.pytorch.org/docs/stable/generated/torch.optim.Muon.html. Accessed: 2026-04-28.

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] A. Riabinin, E. Shulgin, K. Gruntkowska, and P. Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

[42] N. Sato, H. Naganuma, and H. Iiduka. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

[43] W. Shen, R. Huang, M. Huang, C. Shen, and J. Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

[44] E. Shulgin, S. AlRashed, F. Orabona, and P. Richtárik. Beyond the ideal: Analyzing the inexact muon update. *arXiv preprint arXiv:2510.19933*, 2025.

[45] U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[46] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, Apr. 2008. ISSN 1531-5851. doi: 10.1007/s00041-008-9030-4.

[47] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

[49] N. Vyas, D. Morwani, R. Zhao, M. Kwun, I. Shapira, D. Brandfonbrener, L. Janson, and S. Kakade. Soap: Improving and stabilizing shampoo using adam. *International Conference on Learning Representations*, 2025.

[50] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

[51] D. Yu, H. Tao, Y. Wan, L. Luo, and L. Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026.

[52] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33: 15383–15393, 2020.

[53] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Appendix

CONTENTS

1 Introduction **1**

2 (Randomized) Inexact Polar Decomposition **3**

3 Main Theoretical Results **5**

3.1 Warm-up: Analysis for Exact Polar Decomposition 6

3.2 Analysis for Inexact Polar Decomposition 7

4 Numerical Experiments **8**

4.1 Pretraining of nanoGPT on FineWeb10B 8

4.2 CNN on CIFAR-10 9

5 Conclusion **9**

A Preparation Lemmas **15**

B Missing Proofs and Discussions of Section 2 **21**

B.1 Proof of Proposition 2.2 21

B.2 Proof of Proposition 2.3 22

B.3 Uniform Bounds for Inexact Polar Decompositions 23

B.4 Discussion on Proposition 2.3 24

C Auxiliary Lemmas **25**

C.1 Lemmas for Analysis of Exact Polar Decomposition 25

C.2 Lemmas for Analysis of Inexact Polar Decomposition 29

D Missing Proofs of Section 3 **30**

D.1 Proof of Lemma 3.3 30

D.2 Proof of Theorem 3.4 31

D.3 Proof of Theorem 3.5 31

D.4 Proof of Corollary 3.6 33

E Experiment Details and Ablation Study **33**

E.1 Randomized Kaczmarz-Style Sparse Sketching 33

E.2 Pretraining of nanoGPT on FineWeb 33

E.3 CNN on CIFAR-10 34

E.4 nanoGPT Benchmark Ablation Study. 35

E.5 CIFAR-10 Benchmark Ablation Study. 35

E.6 Empirical Study of the Inexact-Polar Quantities on nanoGPT 36

F Limitations **37**

A PREPARATION LEMMAS

Sampling history. The two histories used in the analysis correspond to different stages of iteration k. The σ-field $\mathcal{F}_k$ contains the initialization and all randomness from earlier iterations, including earlier mini-batches and polar draws. Thus $\boldsymbol{X}_k$ and $\boldsymbol{C}_{k-1}$ are $\mathcal{F}_k$-measurable, whereas the current stochastic gradients have not yet been observed. After drawing $\boldsymbol{G}_k^1, \dots, \boldsymbol{G}_k^B$ and forming $\boldsymbol{G}_k$, $\boldsymbol{C}_k$, and $\boldsymbol{M}_k$, the history is

$$\mathcal{G}_k = \sigma(\mathcal{F}_k, \boldsymbol{G}_k^1, \dots, \boldsymbol{G}_k^B), \qquad \mathcal{F}_k \subseteq \mathcal{G}_k \subseteq \mathcal{F}_{k+1}.$$

In particular, $\boldsymbol{M}_k$, $\boldsymbol{S}_k = \widetilde{\boldsymbol{M}}_k - \nabla f(\boldsymbol{X}_k)$, and $\boldsymbol{\xi}_k = \boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)$ are $\mathcal{G}_k$-measurable. The polar seed ζ_k is then drawn independently of $\mathcal{G}_k$ as specified in the oracle model preceding Assumption 2.1, and the polar update forms $\boldsymbol{X}_{k+1}$. The next history $\mathcal{F}_{k+1}$ includes both the current mini-batch and the polar draw.

Conditioning on $\mathcal{F}_k$ therefore averages over the current gradient noise: Assumption 3.2 gives $\mathbb{E}[\boldsymbol{\xi}_k \mid \mathcal{F}_k] = \boldsymbol{0}$. In contrast, $\mathbb{E}[\boldsymbol{\xi}_k \mid \mathcal{G}_k] = \boldsymbol{\xi}_k$, since that noise is already known. The polar descent proof conditions on $\mathcal{G}_k$ to keep $\boldsymbol{M}_k$ and $\boldsymbol{S}_k$ fixed while averaging over the fresh polar randomness. These conditioning steps cannot be interchanged. Even when the polar map is deterministic, $\mathcal{F}_k$ and $\mathcal{G}_k$ generally remain distinct because of the current mini-batch. Natural filtrations of generic martingale sequences below are denoted by $\mathcal{N}_t$ to distinguish them from these algorithmic histories.

Lemma A.1. For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, let $d_0 := \min\{m, n\}$. Then

$$\|\boldsymbol{A}\|_F \le \|\boldsymbol{A}\|_* \le \sqrt{\operatorname{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|_F \qquad \text{and} \qquad \|\boldsymbol{A}\|_F \le \sqrt{d_0}\|\boldsymbol{A}\|_{\mathrm{op}}.$$

Lemma A.2. For matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, we have $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \le \|\boldsymbol{A}\|_F \, \|\boldsymbol{B}\|_F$.

Lemma A.3 (*Jensen's Inequality*)**.** Let X be a random variable for which the following expectations exist. If φ is convex, then

$$\varphi(\mathbb{E}[X]) \le \mathbb{E}[\varphi(X)].$$

If φ is concave, then

$$\mathbb{E}[\varphi(X)] \le \varphi(\mathbb{E}[X]).$$

Lemma A.4. For $x, y \ge 0$ and $\alpha \in (1, 2]$, we have $(x+y)^{1/\alpha} \le x^{1/\alpha} + y^{1/\alpha}$.

Lemma A.5. For $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{g} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I})$, we have

$$\mathbb{E}\left[\left\|\boldsymbol{g}^\top \boldsymbol{M}\right\|\right] \ge \sqrt{\frac{2}{\pi}}\|\boldsymbol{M}\|_F. \tag{3}$$

Proof. If $\boldsymbol{M} = \boldsymbol{0}$, the result is immediate. Otherwise, let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the compact singular value decomposition of $\boldsymbol{M}$, with rank r and singular values $\sigma_1, \dots, \sigma_r > 0$. Define $\widehat{\boldsymbol{g}} := \boldsymbol{U}^\top \boldsymbol{g}$. Since $\boldsymbol{U}$ has orthonormal columns, $\widehat{\boldsymbol{g}} \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I})$. Hence

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{g}^\top \boldsymbol{M}\right\|\right] &= \mathbb{E}\left[\sqrt{\sum_{i=1}^r \sigma_i^2 \widehat{\boldsymbol{g}}_i^2}\right] = \sqrt{\sum_{j=1}^r \sigma_j^2}\,\mathbb{E}\left[\sqrt{\sum_{i=1}^r \frac{\sigma_i^2}{\sum_{j=1}^r \sigma_j^2}\widehat{\boldsymbol{g}}_i^2}\right] \\
&\ge \sqrt{\sum_{j=1}^r \sigma_j^2}\sum_{i=1}^r \frac{\sigma_i^2}{\sum_{j=1}^r \sigma_j^2}\mathbb{E}\left[|\widehat{\boldsymbol{g}}_i|\right] = \sqrt{\frac{2}{\pi}}\sqrt{\sum_{j=1}^r \sigma_j^2} = \sqrt{\frac{2}{\pi}}\|\boldsymbol{M}\|_F,
\end{aligned}$$

where the inequality uses the concavity of $\sqrt{\cdot}$ with weights $\sigma_i^2 / \sum_{j=1}^r \sigma_j^2$. □

Lemma A.6. For $\varphi(x) := \frac{3}{2}x - \frac{1}{2}x^3$, we have, for every $x \in [0,1]$,

$$0 \le \varphi(x) \le 1, \qquad \varphi(x) \ge x, \qquad \varphi'(x) \ge 0.$$

Proof. Note that for $x \in [0,1]$,

$$\varphi(x) - x = \frac{1}{2}x(1-x^2) \ge 0,$$

so $\varphi(x) \ge x$. Also,

$$1 - \varphi(x) = 1 - \frac{3}{2}x + \frac{1}{2}x^3 = \frac{1}{2}(1-x)^2(2+x) \ge 0,$$

using $x \in [0,1]$. Therefore $\varphi(x) \le 1$. Finally,

$$\varphi'(x) = \frac{3}{2}(1-x^2) \ge 0 \qquad \text{for } x \in [0,1].$$

□

Lemma A.7. For $\varphi(x) := \frac{1}{8}\left(15x - 10x^3 + 3x^5\right)$, we have, for every $x \in [0,1]$,

$$0 \le \varphi(x) \le 1, \qquad \varphi(x) \ge x, \qquad \varphi'(x) \ge 0.$$

Proof. First, we show that $\varphi(x) \ge x$. Indeed,

$$\varphi(x) - x = \frac{1}{8}\left(15x - 10x^3 + 3x^5\right) - x = \frac{1}{8}x\left(7 - 10x^2 + 3x^4\right) = \frac{1}{8}x(1-x^2)(7-3x^2).$$

Since $x \in [0,1]$, we have $x \ge 0$, $1 - x^2 \ge 0$, and $7 - 3x^2 \ge 4 > 0$. Hence $\varphi(x) \ge x$. In particular, since $x \ge 0$, this also implies $\varphi(x) \ge 0$.

Next,

$$1 - \varphi(x) = 1 - \frac{1}{8}\left(15x - 10x^3 + 3x^5\right) = \frac{1}{8}\left(8 - 15x + 10x^3 - 3x^5\right) = \frac{1}{8}(1-x)^3(8+9x+3x^2).$$

Since $x \in [0,1]$, we have $(1-x)^3 \ge 0$ and $8 + 9x + 3x^2 > 0$. Thus $\varphi(x) \le 1$. Finally,

$$\varphi'(x) = \frac{1}{8}\left(15 - 30x^2 + 15x^4\right) = \frac{15}{8}(1-x^2)^2.$$

Hence $\varphi'(x) \ge 0$ for every $x \in [0,1]$. □

Lemma A.8. Let $p \in (1,2]$, and let $\boldsymbol{W}_t \in \mathbb{R}^{m \times n}$ be a matrix martingale difference sequence with finite p-th moments with respect to the natural filtration $\mathcal{N}_t = \sigma(\boldsymbol{W}_1, \dots, \boldsymbol{W}_t)$, with $\mathcal{N}_0$ trivial, meaning

$$\mathbb{E}\left[\boldsymbol{W}_t \mid \mathcal{N}_{t-1}\right] = \boldsymbol{0}.$$

Then, for any $T \ge 1$,

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T} \boldsymbol{W}_t\right\|_F\right] \le 2\sqrt{\pi}\,\mathbb{E}\left[\left(\sum_{t=1}^{T} \|\boldsymbol{W}_t\|_F^p\right)^{1/p}\right].$$

Proof. Following [24], let $\boldsymbol{g} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I})$ be independent of $\boldsymbol{W}_1, \dots, \boldsymbol{W}_T$, and define $\boldsymbol{v}_t := \boldsymbol{g}^\top \boldsymbol{W}_t$. Since $\sigma(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{t-1}) \subseteq \sigma(\boldsymbol{W}_1, \dots, \boldsymbol{W}_{t-1}, \boldsymbol{g})$, the tower property and the independence of $\boldsymbol{g}$ give

$$\mathbb{E}\left[\boldsymbol{v}_t \mid \boldsymbol{v}_1, \dots, \boldsymbol{v}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{g}^\top \boldsymbol{W}_t \mid \boldsymbol{W}_1, \dots, \boldsymbol{W}_{t-1}, \boldsymbol{g}\right] \mid \boldsymbol{v}_1, \dots, \boldsymbol{v}_{t-1}\right] = \boldsymbol{0}.$$

Thus, by [31, Lemma 4.3],

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{v}_t\right\|\right] \le 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^p\right)^{1/p}\right].$$

On the other hand, using (3) conditionally on $\sum_{t=1}^{T}\boldsymbol{W}_t$,

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{v}_t\right\|\right] = \mathbb{E}\left[\left\|\boldsymbol{g}^\top\left(\sum_{t=1}^{T}\boldsymbol{W}_t\right)\right\|\right] \ge \sqrt{\frac{2}{\pi}}\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{W}_t\right\|_F\right].$$

It remains to bound the right-hand side. Since $(z_1+\cdots+z_T)^{1/p}$ is concave on $\mathbb{R}_+^T$, Jensen's inequality conditional on $\mathcal{N}_T$ gives

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^p\right)^{1/p}\right] \le \mathbb{E}\left[\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{g}^\top\boldsymbol{W}_t\right\|^p \mid \mathcal{N}_T\right]\right)^{1/p}\right] = \mathbb{E}\left[\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{g}^\top\boldsymbol{W}_t\right\|^p \mid \boldsymbol{W}_t\right]\right)^{1/p}\right]$$

where the last equality uses the independence of $\boldsymbol{g}$ from $\mathcal{N}_T$. For each t, an SVD of $\boldsymbol{W}_t$ and the concavity of $x \mapsto x^{p/2}$ for $p \le 2$ imply

$$\mathbb{E}\left[\left\|\boldsymbol{g}^\top\boldsymbol{W}_t\right\|^p \mid \boldsymbol{W}_t\right] \le \|\boldsymbol{W}_t\|_F^p.$$

Therefore,

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^p\right)^{1/p}\right] \le \mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{W}_t\|_F^p\right)^{1/p}\right].$$

Combining the previous displays yields

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T}\boldsymbol{W}_t\right\|_F\right] \le 2\sqrt{\pi}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\boldsymbol{W}_t\|_F^p\right)^{1/p}\right].$$

This completes the proof. □

Lemma A.9 (*Weighted stochastic-gradient noise bound*)**.** Under Assumption 3.2, let $\boldsymbol{\xi}_s = \boldsymbol{G}_s - \nabla f(\boldsymbol{X}_s)$ and suppose $\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F^\alpha\right] < \infty$ for $s = 0,\dots,k$. Then, for deterministic real coefficients $a_0,\dots,a_k$,

$$\mathbb{E}\left[\left\|\sum_{s=0}^{k}a_s\boldsymbol{\xi}_s\right\|_F\right] \le \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{(\alpha-1)/\alpha}}\left[\sigma_0\left(\sum_{s=0}^{k}|a_s|^\alpha\right)^{1/\alpha} + \sigma_1\sum_{s=0}^{k}|a_s|\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right]\right].$$

Here $\mathcal{C}_\alpha \in (0,2]$ is the constant from Lemma A.11.

Proof. By the sampling order above, $\boldsymbol{\xi}_s$ is $\mathcal{G}_s$-measurable, hence $\mathcal{F}_{s+1}$-measurable. Assumption 3.2 gives $\mathbb{E}\left[\boldsymbol{\xi}_s \mid \mathcal{F}_s\right] = \mathbf{0}$. Lemma A.11 gives

$$\mathbb{E}\left[\|\boldsymbol{\xi}_s\|_F^\alpha \mid \mathcal{F}_s\right] \le \frac{\mathcal{C}_\alpha}{B^{\alpha-1}}\left(\sigma_0^\alpha + \sigma_1^\alpha\|\nabla f(\boldsymbol{X}_s)\|_F^\alpha\right),$$

which also ensures finite α-th moments of $\boldsymbol{\xi}_s$. Put $\boldsymbol{W}_{s+1} = a_s\boldsymbol{\xi}_s$ and $\mathcal{N}_s = \sigma(\boldsymbol{W}_1,\dots,\boldsymbol{W}_s)$, with $\mathcal{N}_0$ trivial. Since $\mathcal{N}_s \subseteq \mathcal{F}_s$, the tower property gives

$$\mathbb{E}\left[\boldsymbol{W}_{s+1} \mid \mathcal{N}_s\right] = \mathbb{E}\left[a_s\mathbb{E}\left[\boldsymbol{\xi}_s \mid \mathcal{F}_s\right] \mid \mathcal{N}_s\right] = \mathbf{0}.$$

Thus Lemma A.8, with $p = \alpha$, yields

$$\mathbb{E}\left[\left\|\sum_{s=0}^{k}a_s\boldsymbol{\xi}_s\right\|_F\right] \le 2\sqrt{\pi}\mathbb{E}\left[\left(\sum_{s=0}^{k}|a_s|^\alpha\|\boldsymbol{\xi}_s\|_F^\alpha\right)^{1/\alpha}\right]. \tag{4}$$

To bound the realized moment on the right, set $\tau = 1/\alpha$, $c_s = |a_s|^\alpha$, and define

$$Z_r = \sum_{s=0}^{r} c_s \|\boldsymbol{\xi}_s\|_F^\alpha + \frac{\mathcal{C}_\alpha \sigma_0^\alpha}{B^{\alpha-1}} \sum_{s=r+1}^{k} c_s \quad (0 \le r \le k),$$

$$Z_{-1} = \frac{\mathcal{C}_\alpha \sigma_0^\alpha}{B^{\alpha-1}} \sum_{s=0}^{k} c_s.$$

For each r, all terms involving $\boldsymbol{\xi}_0, \dots, \boldsymbol{\xi}_{r-1}$ are $\mathcal{F}_r$-measurable. Conditional Jensen's inequality for the concave function $x \mapsto x^\tau$, the conditional moment bound, and $(x+y)^\tau \le x^\tau + y^\tau$ give

$$\begin{aligned}
\mathbb{E}\left[Z_r^\tau \mid \mathcal{F}_r\right] &\le \left(\sum_{s=0}^{r-1} c_s \|\boldsymbol{\xi}_s\|_F^\alpha + c_r \mathbb{E}\left[\|\boldsymbol{\xi}_r\|_F^\alpha \mid \mathcal{F}_r\right] + \frac{\mathcal{C}_\alpha \sigma_0^\alpha}{B^{\alpha-1}} \sum_{s=r+1}^{k} c_s \right)^\tau \\
&\le \left(Z_{r-1} + \frac{\mathcal{C}_\alpha \sigma_1^\alpha c_r}{B^{\alpha-1}} \|\nabla f(\boldsymbol{X}_r)\|_F^\alpha \right)^\tau \\
&\le Z_{r-1}^\tau + \frac{\mathcal{C}_\alpha^{1/\alpha} \sigma_1 |a_r|}{B^{(\alpha-1)/\alpha}} \|\nabla f(\boldsymbol{X}_r)\|_F.
\end{aligned}$$

Taking expectations and iterating this inequality from $r = k$ down to $r = 0$ yields

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{s=0}^{k} |a_s|^\alpha \|\boldsymbol{\xi}_s\|_F^\alpha \right)^{1/\alpha} \right] &= \mathbb{E}\left[Z_k^\tau\right] \\
&\le \frac{\mathcal{C}_\alpha^{1/\alpha}}{B^{(\alpha-1)/\alpha}} \left[\sigma_0 \left(\sum_{s=0}^{k} |a_s|^\alpha \right)^{1/\alpha} + \sigma_1 \sum_{s=0}^{k} |a_s| \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right] \right].
\end{aligned}$$

Combining this with (4) proves the claim. □

The proof conditions one time index at a time, keeping the gradient-dependent contribution in terms of $\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right]$, as required by the descent argument. It does not replace all realized moments inside the fractional power by their conditional moments.

Lemma A.10. Let $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ have rank r and compact singular value decomposition

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix}^\top,$$

where $\sigma_1 \ge \cdots \ge \sigma_r > 0$ and

$$\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_1, \dots, \sigma_s), \qquad \boldsymbol{\Sigma}_2 = \mathrm{diag}(\sigma_{s+1}, \dots, \sigma_r),$$

for some $1 \le s < r$. Define

$$\varrho_s := \frac{\sigma_{s+1}}{\sigma_s}.$$

Let $p \ge 0$ be an integer, set $\ell = s + p$, and let $\boldsymbol{\Omega} \in \mathbb{R}^{n \times \ell}$. Define

$$\boldsymbol{\Omega}_1 := \boldsymbol{V}_1^\top \boldsymbol{\Omega} \in \mathbb{R}^{s \times \ell}, \qquad \boldsymbol{\Omega}_2 := \boldsymbol{V}_2^\top \boldsymbol{\Omega} \in \mathbb{R}^{(r-s) \times \ell}.$$

Assume that $\boldsymbol{\Omega}_1$ has full row rank. For an integer $h \ge 0$, form

$$\boldsymbol{Y} := (\boldsymbol{M}\boldsymbol{M}^\top)^h \boldsymbol{M}\boldsymbol{\Omega}, \qquad \boldsymbol{Q} := \mathrm{orth}(\boldsymbol{Y}), \qquad \boldsymbol{P}_{\boldsymbol{Q}} := \boldsymbol{Q}\boldsymbol{Q}^\top,$$

where $\mathrm{orth}(\boldsymbol{Y})$ denotes any matrix whose columns form an orthonormal basis for $\mathrm{range}(\boldsymbol{Y})$. Then

$$\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{M}\|_F^2 \le \|\boldsymbol{\Sigma}_2\|_F^2 + \varrho_s^{4h} \|\boldsymbol{\Sigma}_2 \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger\|_F^2. \tag{5}$$

Consequently,

$$\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{M}\|_F^2 \ge \|\boldsymbol{\Sigma}_1\|_F^2 - \varrho_s^{4h} \|\boldsymbol{\Sigma}_2 \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger\|_F^2. \tag{6}$$

Moreover, if $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times \ell}$ with $\ell = s + p$ and $p \geq 2$, then $\boldsymbol{\Omega}_1$ has full row rank almost surely and

$$\mathbb{E}_{\boldsymbol{\Omega}} \|\boldsymbol{P}_{\boldsymbol{Q}} \boldsymbol{M}\|_F^2 \geq \left[\sum_{j=1}^{s} \sigma_j^2 - \frac{s}{p-1} \varrho_s^{4h} \sum_{j>s} \sigma_j^2 \right]_+ . \tag{7}$$

Proof. Set

$$\boldsymbol{U}_r := [\boldsymbol{U}_1 \quad \boldsymbol{U}_2], \qquad \boldsymbol{V}_r := [\boldsymbol{V}_1 \quad \boldsymbol{V}_2], \qquad \boldsymbol{\Sigma} := \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}.$$

Then $\boldsymbol{M} = \boldsymbol{U}_r \boldsymbol{\Sigma} \boldsymbol{V}_r^\top$. Since

$$(\boldsymbol{M}\boldsymbol{M}^\top)^h \boldsymbol{M} = \boldsymbol{U}_r \boldsymbol{\Sigma}^{2h+1} \boldsymbol{V}_r^\top,$$

we have

$$\boldsymbol{Y} = \boldsymbol{U}_r \begin{bmatrix} \boldsymbol{\Sigma}_1^{2h+1} \boldsymbol{\Omega}_1 \\ \boldsymbol{\Sigma}_2^{2h+1} \boldsymbol{\Omega}_2 \end{bmatrix}.$$

Define

$$\boldsymbol{G} := \begin{bmatrix} \boldsymbol{\Sigma}_1^{2h+1} \boldsymbol{\Omega}_1 \\ \boldsymbol{\Sigma}_2^{2h+1} \boldsymbol{\Omega}_2 \end{bmatrix} \in \mathbb{R}^{r\times \ell}.$$

Then $\boldsymbol{Y} = \boldsymbol{U}_r \boldsymbol{G}$. Let $\boldsymbol{P}_{\boldsymbol{G}}$ denote the orthogonal projector onto $\operatorname{range}(\boldsymbol{G})$. Since $\boldsymbol{U}_r$ has orthonormal columns, the orthogonal projector onto $\operatorname{range}(\boldsymbol{Y}) = \boldsymbol{U}_r \operatorname{range}(\boldsymbol{G})$ is

$$\boldsymbol{P}_{\boldsymbol{Q}} = \boldsymbol{U}_r \boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{U}_r^\top.$$

Therefore,

$$\begin{aligned} \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{M}\|_F &= \|(\boldsymbol{I} - \boldsymbol{U}_r \boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{U}_r^\top) \boldsymbol{U}_r \boldsymbol{\Sigma} \boldsymbol{V}_r^\top\|_F \\ &= \|\boldsymbol{U}_r (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}}) \boldsymbol{\Sigma} \boldsymbol{V}_r^\top\|_F = \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}})\boldsymbol{\Sigma}\|_F. \end{aligned} \tag{8}$$

Since $\boldsymbol{\Omega}_1$ has full row rank, $\boldsymbol{\Omega}_1 \boldsymbol{\Omega}_1^\dagger = \boldsymbol{I}_s$. Define

$$\boldsymbol{F} := \boldsymbol{\Sigma}_2^{2h+1} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger \boldsymbol{\Sigma}_1^{-(2h+1)} \in \mathbb{R}^{(r-s)\times s}.$$

Then

$$\boldsymbol{G} \boldsymbol{\Omega}_1^\dagger \boldsymbol{\Sigma}_1^{-(2h+1)} = \begin{bmatrix} \boldsymbol{\Sigma}_1^{2h+1} \boldsymbol{\Omega}_1 \\ \boldsymbol{\Sigma}_2^{2h+1} \boldsymbol{\Omega}_2 \end{bmatrix} \boldsymbol{\Omega}_1^\dagger \boldsymbol{\Sigma}_1^{-(2h+1)} = \begin{bmatrix} \boldsymbol{I}_s \\ \boldsymbol{F} \end{bmatrix}. \tag{9}$$

Thus

$$\operatorname{range} \begin{bmatrix} \boldsymbol{I}_s \\ \boldsymbol{F} \end{bmatrix} \subseteq \operatorname{range}(\boldsymbol{G}).$$

By the least-squares characterization of the orthogonal projector,

$$\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}})\boldsymbol{\Sigma}\|_F = \min_{\boldsymbol{X} \in \mathbb{R}^{\ell\times r}} \|\boldsymbol{\Sigma} - \boldsymbol{G}\boldsymbol{X}\|_F.$$

Using (9), for any $\boldsymbol{C} \in \mathbb{R}^{s\times r}$,

$$\begin{bmatrix} \boldsymbol{I}_s \\ \boldsymbol{F} \end{bmatrix} \boldsymbol{C} = \boldsymbol{G} \boldsymbol{\Omega}_1^\dagger \boldsymbol{\Sigma}_1^{-(2h+1)} \boldsymbol{C}$$

lies in the class of admissible approximants $\boldsymbol{G}\boldsymbol{X}$. Hence

$$\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}})\boldsymbol{\Sigma}\|_F \leq \left\| \boldsymbol{\Sigma} - \begin{bmatrix} \boldsymbol{I}_s \\ \boldsymbol{F} \end{bmatrix} \boldsymbol{C} \right\|_F$$

for every $\boldsymbol{C} \in \mathbb{R}^{s\times r}$. Taking

$$\boldsymbol{C} := [\boldsymbol{\Sigma}_1 \quad \boldsymbol{0}],$$

we obtain

$$\begin{aligned} \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{G}})\boldsymbol{\Sigma}\|_F^2 &\leq \left\| \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{I}_s \\ \boldsymbol{F} \end{bmatrix} [\boldsymbol{\Sigma}_1 \quad \boldsymbol{0}] \right\|_F^2 \\ &= \left\| \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{F}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 \end{bmatrix} \right\|_F^2 = \|\boldsymbol{\Sigma}_2\|_F^2 + \|\boldsymbol{F}\boldsymbol{\Sigma}_1\|_F^2. \end{aligned} \tag{10}$$

It remains to bound $\|\boldsymbol{F}\boldsymbol{\Sigma}_1\|_F$. By the definition of $\boldsymbol{F}$,

$$\boldsymbol{F}\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2^{2h}\left(\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right)\boldsymbol{\Sigma}_1^{-2h}.$$

Using $\|\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}\|_F \le \|\boldsymbol{A}\|_{\mathrm{op}}\|\boldsymbol{X}\|_F\|\boldsymbol{B}\|_{\mathrm{op}}$, we get

$$\|\boldsymbol{F}\boldsymbol{\Sigma}_1\|_F \le \|\boldsymbol{\Sigma}_2^{2h}\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F\|\boldsymbol{\Sigma}_1^{-2h}\|_{\mathrm{op}} = \varrho_s^{2h}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F. \tag{11}$$

Squaring (11) and substituting into (10) gives

$$\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}})\boldsymbol{\Sigma}\|_F^2 \le \|\boldsymbol{\Sigma}_2\|_F^2 + \varrho_s^{4h}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2.$$

Together with (8), this proves (5).

Next, since $\boldsymbol{P}_{\boldsymbol{Q}}$ is an orthogonal projector,

$$\|\boldsymbol{M}\|_F^2 = \|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{M}\|_F^2 + \|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{M}\|_F^2.$$

Also,

$$\|\boldsymbol{M}\|_F^2 = \|\boldsymbol{\Sigma}_1\|_F^2 + \|\boldsymbol{\Sigma}_2\|_F^2.$$

Combining these identities with (5) yields

$$\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{M}\|_F^2 \ge \|\boldsymbol{\Sigma}_1\|_F^2 - \varrho_s^{4h}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2,$$

which proves (6).

It remains to prove the Gaussian expectation bound. Suppose $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times\ell}$ with $\ell = s+p$ and $p \ge 2$. By rotational invariance of the Gaussian distribution, after extending $[\boldsymbol{V}_1 \quad \boldsymbol{V}_2]$ to an orthogonal matrix, the blocks $\boldsymbol{\Omega}_1 = \boldsymbol{V}_1^\top\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_2 = \boldsymbol{V}_2^\top\boldsymbol{\Omega}$ are independent standard Gaussian matrices. Moreover, $\boldsymbol{\Omega}_1 \in \mathbb{R}^{s\times(s+p)}$ has full row rank almost surely.

Conditioning on $\boldsymbol{\Omega}_1$, we have

$$\mathbb{E}_{\boldsymbol{\Omega}_2}\left[\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2 \mid \boldsymbol{\Omega}_1\right] = \mathrm{tr}\left((\boldsymbol{\Omega}_1^\dagger)^\top\mathbb{E}_{\boldsymbol{\Omega}_2}\left[\boldsymbol{\Omega}_2^\top\boldsymbol{\Sigma}_2^2\boldsymbol{\Omega}_2\right]\boldsymbol{\Omega}_1^\dagger\right).$$

Since $\boldsymbol{\Omega}_2$ has independent standard Gaussian entries,

$$\mathbb{E}_{\boldsymbol{\Omega}_2}\left[\boldsymbol{\Omega}_2^\top\boldsymbol{\Sigma}_2^2\boldsymbol{\Omega}_2\right] = \mathrm{tr}(\boldsymbol{\Sigma}_2^2)\boldsymbol{I}_\ell = \|\boldsymbol{\Sigma}_2\|_F^2\boldsymbol{I}_\ell.$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\Omega}_2}\left[\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2 \mid \boldsymbol{\Omega}_1\right] = \|\boldsymbol{\Sigma}_2\|_F^2\|\boldsymbol{\Omega}_1^\dagger\|_F^2.$$

Taking expectation over $\boldsymbol{\Omega}_1$ gives

$$\mathbb{E}_{\boldsymbol{\Omega}}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2 = \|\boldsymbol{\Sigma}_2\|_F^2\mathbb{E}_{\boldsymbol{\Omega}_1}\|\boldsymbol{\Omega}_1^\dagger\|_F^2.$$

Since $\boldsymbol{\Omega}_1$ has full row rank almost surely,

$$\boldsymbol{\Omega}_1^\dagger = \boldsymbol{\Omega}_1^\top(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\top)^{-1},$$

and hence

$$\|\boldsymbol{\Omega}_1^\dagger\|_F^2 = \mathrm{tr}\left((\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\top)^{-1}\right).$$

The matrix $\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\top$ has the Wishart distribution $W_s(\ell, \boldsymbol{I}_s)$. Since $\ell = s+p$ and $p \ge 2$, the inverse moment exists and

$$\mathbb{E}\left[(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^\top)^{-1}\right] = \frac{1}{\ell-s-1}\boldsymbol{I}_s = \frac{1}{p-1}\boldsymbol{I}_s.$$

Thus

$$\mathbb{E}_{\boldsymbol{\Omega}_1}\|\boldsymbol{\Omega}_1^\dagger\|_F^2 = \frac{s}{p-1},$$

and therefore

$$\mathbb{E}_{\boldsymbol{\Omega}}\|\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_F^2 = \frac{s}{p-1}\|\boldsymbol{\Sigma}_2\|_F^2.$$

Taking expectations in (6), we get

$$\mathbb{E}_{\boldsymbol{\Omega}}\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{M}\|_F^2 \ge \|\boldsymbol{\Sigma}_1\|_F^2 - \frac{s}{p-1}\varrho_s^{4h}\|\boldsymbol{\Sigma}_2\|_F^2.$$

Finally, using

$$\|\mathbf{\Sigma}_1\|_F^2 = \sum_{j=1}^{s} \sigma_j^2, \qquad \|\mathbf{\Sigma}_2\|_F^2 = \sum_{j>s} \sigma_j^2,$$

and the trivial bound $\|\boldsymbol{P_Q M}\|_F^2 \geq 0$, we obtain

$$\mathbb{E}_{\mathbf{\Omega}} \|\boldsymbol{P_Q M}\|_F^2 \geq \left[\sum_{j=1}^{s} \sigma_j^2 - \frac{s}{p-1} \varrho_s^{4h} \sum_{j>s} \sigma_j^2 \right]_+ .$$

This proves (7). □

Lemma A.11. Under Assumption 3.2, we have

$$\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \mathcal{F}_k\right] \leq \frac{\mathcal{C}_\alpha}{B^{\alpha-1}} \left(\sigma_0^\alpha + \sigma_1^\alpha \|\nabla f(\boldsymbol{X}_k)\|_F^\alpha\right)$$

for some $0 < \mathcal{C}_\alpha \leq 2$ depending only on α and independent of B. Consequently, the same bound holds with conditioning on $\boldsymbol{X}_k$ in place of $\mathcal{F}_k$.

Proof. From the definition $\boldsymbol{G}_k := B^{-1} \sum_{i=1}^{B} \boldsymbol{G}_k^i$, conditional on $\mathcal{F}_k$ we have

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \mathcal{F}_k\right] &= \frac{1}{B^\alpha} \mathbb{E}\left[\left\| \sum_{i=1}^{B} \left(\boldsymbol{G}_k^i - \nabla f(\boldsymbol{X}_k)\right) \right\|_F^\alpha \mid \mathcal{F}_k \right] \\
&\leq \frac{\mathcal{C}_\alpha}{B^\alpha} \sum_{i=1}^{B} \mathbb{E}\left[\left\|\boldsymbol{G}_k^i - \nabla f(\boldsymbol{X}_k)\right\|_F^\alpha \mid \mathcal{F}_k \right],
\end{aligned}$$

where the inequality follows from the von Bahr-Esseen inequality, applied conditionally on $\mathcal{F}_k$ to the independent mean-zero random matrices $\boldsymbol{G}_k^i - \nabla f(\boldsymbol{X}_k)$, viewed as elements of the Hilbert space equipped with the Frobenius norm. For any $\alpha \in (1, 2]$, one may take $\mathcal{C}_\alpha \leq 2$. Using Assumption 3.2, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k)\|_F^\alpha \mid \mathcal{F}_k\right] &\leq \frac{\mathcal{C}_\alpha}{B^\alpha} \sum_{i=1}^{B} \left(\sigma_0^\alpha + \sigma_1^\alpha \|\nabla f(\boldsymbol{X}_k)\|_F^\alpha\right) \\
&= \frac{\mathcal{C}_\alpha}{B^{\alpha-1}} \left(\sigma_0^\alpha + \sigma_1^\alpha \|\nabla f(\boldsymbol{X}_k)\|_F^\alpha\right).
\end{aligned}$$

Since $\boldsymbol{X}_k$ is $\mathcal{F}_k$-measurable, taking conditional expectation of the preceding bound with respect to $\boldsymbol{X}_k$ yields the version conditioned on $\boldsymbol{X}_k$. □

B Missing Proofs and Discussions of Section 2

B.1 Proof of Proposition 2.2

Proposition 2.2. Suppose $\varphi(x) = \frac{1}{2}(3x - x^3)$ or $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$. Let $p_0(x) = x$ and $p_{t+1}(x) = \varphi(p_t(x))$ for $t = 0, 1, \ldots, q-1$. Let $\boldsymbol{M} = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^\top$ be a compact SVD of a rank-r matrix with singular values $\sigma_1, \ldots, \sigma_r > 0$, and let $\delta \geq \|\boldsymbol{M}\|_{\mathrm{op}}$. Define

$$\mathcal{T}(\boldsymbol{M}) := p_q\left(\boldsymbol{M}/\delta\right) = \boldsymbol{U}\,\mathrm{diag}\left(p_q\left(\sigma_1/\delta\right), \cdots, p_q\left(\sigma_r/\delta\right)\right)\boldsymbol{V}^\top.$$

Then, $\mathcal{T}(\boldsymbol{M})$ satisfies Assumption 2.1 with $\nu = 0$ and $\gamma = 1 - \frac{\sum_{i=1}^{r} \sigma_i p_q(\sigma_i/\delta)}{\sum_{i=1}^{r} \sigma_i}$.

Proof. We prove the result for $\varphi(x) = \frac{3}{2}x - \frac{1}{2}x^3$. From the Newton-Schulz iterations, we have $\boldsymbol{Z}_0 = \boldsymbol{M}/\delta$ and $\boldsymbol{Z}_{t+1} = \varphi\left(\boldsymbol{Z}_t\right)$ for $t = 0, 1, \cdots, q-1$. Define $s_i := \frac{\sigma_i}{\delta}$ for $i = 1, \cdots r$. Since $\delta \geq \|\boldsymbol{M}\|_{\mathrm{op}}$, we have $s_i \in (0, 1]$. First, we show for every $t \geq 0$,

$$\boldsymbol{Z}_t = \boldsymbol{U}\,\mathrm{diag}\big(p_t(s_1), \ldots, p_t(s_r)\big)\boldsymbol{V}^\top.$$

We prove this by induction on t. For $t = 0$, this is immediate from the SVD of $\boldsymbol{M}$:

$$\boldsymbol{Z}_0 = \frac{\boldsymbol{M}}{\delta} = \boldsymbol{U}\,\mathrm{diag}\left(\frac{\sigma_1}{\delta}, \ldots, \frac{\sigma_r}{\delta}\right)\boldsymbol{V}^\top = \boldsymbol{U}\,\mathrm{diag}(s_1, \ldots, s_r)\boldsymbol{V}^\top.$$

Assume the representation holds at time t, i.e., $\boldsymbol{Z}_t = \boldsymbol{U}\boldsymbol{D}_t\boldsymbol{V}^\top$, where

$$\boldsymbol{D}_t := \mathrm{diag}\big(p_t(s_1), \ldots, p_t(s_r)\big).$$

Then $\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\boldsymbol{Z}_t = (\boldsymbol{U}\boldsymbol{D}_t\boldsymbol{V}^\top)(\boldsymbol{V}\boldsymbol{D}_t\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{D}_t\boldsymbol{V}^\top) = \boldsymbol{U}\boldsymbol{D}_t^3\boldsymbol{V}^\top$. Therefore

$$\boldsymbol{Z}_{t+1} = \frac{3}{2}\boldsymbol{U}\boldsymbol{D}_t\boldsymbol{V}^\top - \frac{1}{2}\boldsymbol{U}\boldsymbol{D}_t^3\boldsymbol{V}^\top = \boldsymbol{U}\left(\frac{3}{2}\boldsymbol{D}_t - \frac{1}{2}\boldsymbol{D}_t^3\right)\boldsymbol{V}^\top.$$

Since $\boldsymbol{D}_t$ is diagonal,

$$\frac{3}{2}\boldsymbol{D}_t - \frac{1}{2}\boldsymbol{D}_t^3 = \mathrm{diag}\big(\varphi(p_t(s_1)), \ldots, \varphi(p_t(s_r))\big) = \mathrm{diag}\big(p_{t+1}(s_1), \ldots, p_{t+1}(s_r)\big).$$

The last line follows from the definition $p_{t+1}(x) = \varphi(p_t(x))$. Thus, the claim follows by induction.

By induction and Lemma A.6, we have $0 \le p_q(x) \le 1$ and $p_q(x) \ge x$ for all $x \in [0, 1]$. Hence, for $s_i = \sigma_i/\delta \in (0, 1]$, we obtain $0 < p_q(s_i) \le 1$. Therefore,

$$\|\mathcal{T}(\boldsymbol{M})\|_{\mathrm{op}} = \|p_q\left(\boldsymbol{M}/\delta\right)\|_{\mathrm{op}} = \max_{1\le i\le r} p_q(s_i) \le 1.$$

Thus, the second condition of Assumption 2.1 holds with $\nu = 0$. For the first condition, we have

$$\begin{aligned}\langle \boldsymbol{M}, \mathcal{T}(\boldsymbol{M})\rangle &= \mathrm{tr}(\boldsymbol{M}^\top\mathcal{T}(\boldsymbol{M})) = \mathrm{tr}\left((\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{D}_q\boldsymbol{V}^\top)\right) = \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{D}_q\boldsymbol{V}^\top)\\ &= \mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{D}_q) = \sum_{i=1}^r \sigma_i p_q(s_i) = \left(\frac{\sum_{i=1}^r \sigma_i p_q(s_i)}{\sum_{i=1}^r \sigma_i}\right)\|\boldsymbol{M}\|_*.\end{aligned}$$

Since $0 < p_q(s_i) \le 1$, the stated value of γ belongs to $[0, 1)$. Because $\mathcal{T}(\boldsymbol{M})$ is deterministic in this proposition, the same bounds hold for the conditional expectations in Assumption 2.1. This concludes the proof for the cubic polynomial. The proof for $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$ follows similarly using Lemma A.7. □

B.2 Proof of Proposition 2.3

Proposition 2.3. Let $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ have rank r with singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$. Let $1 \le s < r$, let $p \ge 2$ and $\ell = s + p \le \min(m, n)$, and let h, q be nonnegative integers. Define $\varrho_s := \sigma_{s+1}/\sigma_s \in [0, 1]$. For any $\delta \ge \|\boldsymbol{M}\|_{\mathrm{op}}$, the output $\mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega})$ from Algorithm 1, with $\boldsymbol{\Omega} \sim \mathcal{N}(0, 1)^{n\times\ell}$ and $\varphi(x) = \frac{1}{2}(3x - x^3)$ or $\varphi(x) = \frac{1}{8}(15x - 10x^3 + 3x^5)$, satisfies

$$\mathbb{E}_{\boldsymbol{\Omega}}\left[\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega})\rangle\right] \ge \frac{1}{\delta}\left[\sum_{j=1}^s \sigma_j^2 - \frac{s}{p-1}\varrho_s^{4h}\sum_{j>s}\sigma_j^2\right]_+.$$

Moreover, almost surely, $\|\mathcal{T}_{h,q}\left(\boldsymbol{M}; \boldsymbol{\Omega}\right)\|_{\mathrm{op}} \le 1$.

Proof. Let $p_0(x) = x$ and $p_{t+1}(x) = \varphi(p_t(x))$. Fix a realization of $\boldsymbol{\Omega}$ and write $\boldsymbol{B} = \boldsymbol{Q}^\top\boldsymbol{M} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top$, where $\widehat{r} = \mathrm{rank}(\boldsymbol{B})$ and $\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{\widehat{r}})$. Since $\boldsymbol{Q}$ has orthonormal columns, $\|\boldsymbol{B}\|_{\mathrm{op}} \le \|\boldsymbol{M}\|_{\mathrm{op}}$, and hence $\delta \ge \|\boldsymbol{B}\|_{\mathrm{op}}$. The scalar bounds in Lemmas A.6 and A.7 imply $0 \le p_q(x) \le 1$ for all $x \in [0, 1]$. Therefore,

$$\|\boldsymbol{Z}_q\|_{\mathrm{op}} = \|p_q\left(\boldsymbol{B}/\delta\right)\|_{\mathrm{op}} = \max_{1\le i\le\widehat{r}} p_q\left(\widehat{\sigma}_i/\delta\right) \le 1.$$

Since $\mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega}) = \boldsymbol{Q}\boldsymbol{Z}_q$ and $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$, this gives $\|\mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega})\|_{\mathrm{op}} \le 1$.

For the alignment term, we have

$$\begin{aligned}\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega})\rangle &= \langle \boldsymbol{M}, \boldsymbol{Q}\boldsymbol{Z}_q\rangle = \langle \boldsymbol{Q}^\top\boldsymbol{M}, \boldsymbol{Z}_q\rangle = \langle \boldsymbol{B}, \boldsymbol{Z}_q\rangle\\ &= \sum_{i=1}^{\widehat{r}} \widehat{\sigma}_i p_q\left(\widehat{\sigma}_i/\delta\right) \ge \sum_{i=1}^{\widehat{r}} \widehat{\sigma}_i^2/\delta = \|\boldsymbol{B}\|_F^2/\delta,\end{aligned}$$

where the inequality uses $p_q(x) \geq x$ on $[0,1]$. Let $\boldsymbol{P_Q} := \boldsymbol{Q}\boldsymbol{Q}^\top$. Since $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$,

$$\|\boldsymbol{B}\|_F^2 = \|\boldsymbol{Q}^\top\boldsymbol{M}\|_F^2 = \|\boldsymbol{P_Q}\boldsymbol{M}\|_F^2.$$

Thus,

$$\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})\rangle \geq \frac{\|\boldsymbol{P_Q}\boldsymbol{M}\|_F^2}{\delta}.$$

Taking expectation with respect to $\boldsymbol{\Omega}$ and applying Lemma A.10 gives

$$\mathbb{E}_{\boldsymbol{\Omega}}\left[\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})\rangle\right] \geq \frac{1}{\delta}\left[\sum_{j=1}^{s}\sigma_j^2 - \frac{s}{p-1}\varrho_s^{4h}\sum_{j>s}\sigma_j^2\right]_+.$$

This completes the proof. □

B.3 Uniform Bounds for Inexact Polar Decompositions

The following bounds provide deterministic oracle constants that are independent of the momentum matrices, their ranks, and the optimization horizon.

Lemma B.1 (*Uniform polar-oracle bounds*)**.** Let $d_0 = \min\{m,n\} \geq 1$, let h, q be nonnegative integers, and let p_q be the q-fold iterate of either the cubic or the theoretical quintic Newton–Schulz polynomial in Section 2, with $p_0(x) = x$. For $\boldsymbol{M} \neq \boldsymbol{0}$, set $\delta = \|\boldsymbol{M}\|_F$ and define

$$\mathcal{T}_{\text{full}}(\boldsymbol{M}) = p_q(\boldsymbol{M}/\delta), \qquad \mathcal{T}_{\text{G}}(\boldsymbol{M};\boldsymbol{\Omega}) = \boldsymbol{Q}p_q(\boldsymbol{Q}^\top\boldsymbol{M}/\delta),$$

where $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times\ell}$ for any integer $\ell \geq 1$ and $\boldsymbol{Q}$ is an orthonormal basis of the actual range of $(\boldsymbol{M}\boldsymbol{M}^\top)^h\boldsymbol{M}\boldsymbol{\Omega}$. Define both maps to be zero at $\boldsymbol{M} = \boldsymbol{0}$. With fresh independent Gaussian sketches at each iteration, these maps satisfy Assumption 2.1 with the following uniform choices:

	$\bar{\gamma}$	$\bar{\nu}$
$\mathcal{T}_{\text{full}}$	$1 - d_0^{-1/2}$	0
$\mathcal{T}_{\text{G}}$	$1 - d_0^{-3/2}$	$0.$

These choices hold for every input and every h, q, ℓ as above, including when the input rank changes or is smaller than ℓ.

Proof. Both scalar iterations preserve $[0,1]$ and satisfy $\varphi(x) \geq x$ there (Lemmas A.6 and A.7). Hence $x \leq p_q(x) \leq 1$ for $x \in [0,1]$, including $q = 0$. Every singular value of $\boldsymbol{M}/\delta$ and of $\boldsymbol{Q}^\top\boldsymbol{M}/\delta$ lies in this interval. It follows that both output operator norms are at most one. For the full-space map, writing $\sigma_1, \ldots, \sigma_r$ for the positive singular values of $\boldsymbol{M}$,

$$\langle \boldsymbol{M}, \mathcal{T}_{\text{full}}(\boldsymbol{M})\rangle = \sum_{i=1}^{r}\sigma_i p_q(\sigma_i/\delta) \geq \frac{\|\boldsymbol{M}\|_F^2}{\delta} = \|\boldsymbol{M}\|_F \geq d_0^{-1/2}\|\boldsymbol{M}\|_*.$$

For the Gaussian map, set $\boldsymbol{B} = \boldsymbol{Q}^\top\boldsymbol{M}$. Its number of rows is the dimension of the sketch range, which is $\min\{r,\ell\}$ almost surely; no full-rank assumption on the sketch is needed. Orthogonality of $\boldsymbol{Q}$ and the same singular-value calculation give

$$\langle \boldsymbol{M}, \mathcal{T}_{\text{G}}(\boldsymbol{M};\boldsymbol{\Omega})\rangle = \langle \boldsymbol{B}, p_q(\boldsymbol{B}/\delta)\rangle \geq \frac{\|\boldsymbol{Q}^\top\boldsymbol{M}\|_F^2}{\delta}. \tag{12}$$

We next show that a single Gaussian column already guarantees

$$\mathbb{E}_{\boldsymbol{\Omega}}\|\boldsymbol{Q}^\top\boldsymbol{M}\|_F^2 \geq \frac{\|\boldsymbol{M}\|_F^2}{r}. \tag{13}$$

Take a thin SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and the first column $\boldsymbol{\omega}$ of $\boldsymbol{\Omega}$. Then $\boldsymbol{g} = \boldsymbol{V}^\top\boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{I}_r)$ and $\boldsymbol{y} = (\boldsymbol{M}\boldsymbol{M}^\top)^h\boldsymbol{M}\boldsymbol{\omega}$ belongs to the range of $\boldsymbol{Q}$. Put $a_i = \sigma_i^2$, $w_i = g_i^2$, and $t = 2h+1$. Almost surely $\boldsymbol{y} \neq 0$, and projection onto its span gives

$$\|\boldsymbol{Q}^\top\boldsymbol{M}\|_F^2 \geq \frac{\|\boldsymbol{M}^\top\boldsymbol{y}\|_2^2}{\|\boldsymbol{y}\|_2^2} = \frac{\sum_i w_i a_i^{t+1}}{\sum_i w_i a_i^t} \geq \frac{\sum_i w_i a_i}{\sum_i w_i}.$$

The last inequality follows by multiplying the positive denominators and using

$$\left(\sum_i w_i a_i^{t+1}\right)\left(\sum_i w_i\right) - \left(\sum_i w_i a_i^t\right)\left(\sum_i w_i a_i\right)$$
$$= \frac{1}{2}\sum_{i,j} w_i w_j (a_i^t - a_j^t)(a_i - a_j) \geq 0.$$

Since the g_i are identically distributed, $\mathbb{E}[w_i / \sum_j w_j] = 1/r$. Taking expectations proves (13). Combining it with (12) and $\delta = \|\boldsymbol{M}\|_F$ yields

$$\mathbb{E}_{\boldsymbol{\Omega}}\langle \boldsymbol{M}, \mathcal{T}_{\mathrm{G}}(\boldsymbol{M};\boldsymbol{\Omega})\rangle \geq \frac{\|\boldsymbol{M}\|_F}{r} \geq \frac{\|\boldsymbol{M}\|_*}{r^{3/2}} \geq d_0^{-3/2}\|\boldsymbol{M}\|_*.$$

At $\boldsymbol{M} = \mathbf{0}$, both alignment bounds and the operator-norm bounds hold by definition. At each optimizer iteration, conditioning on $\mathcal{G}_k$ fixes $\boldsymbol{M}_k$ and leaves a fresh standard Gaussian sketch. Thus the bounds hold conditionally on $\mathcal{G}_k$, and hence on $\boldsymbol{M}_k$, with the stated deterministic constants. The full-space map is deterministic given $\boldsymbol{M}_k$. □

These conservative dimension bounds complement the spectrum-dependent bound in Proposition 2.3, which can provide better alignment constants for particular inputs. The uniform bounds above require Frobenius normalization and the two theoretical polynomials; they do not establish the same constants for the empirical quintic, PolarExpress, or non-Gaussian sketches. More generally, $\|\boldsymbol{M}\|_{\mathrm{op}} \leq \delta \leq C\|\boldsymbol{M}\|_F$ with a fixed $C \geq 1$ divides the displayed alignment constants by C. An arbitrary larger scaling, or an additive normalization constant, need not satisfy this uniform condition.

B.4 Discussion on Proposition 2.3

Define

$$A_{s,h} := \sum_{j=1}^{s} \sigma_j^2 - \frac{s}{p-1}\varrho_s^{4h}\sum_{j>s}\sigma_j^2, \qquad \theta_{s,h} := \frac{[A_{s,h}]_+}{\delta\|\boldsymbol{M}\|_*}.$$

Then Proposition 2.3 gives

$$\mathbb{E}_{\boldsymbol{\Omega}}\left[\langle \boldsymbol{M}, \mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})\rangle\right] \geq \theta_{s,h}\|\boldsymbol{M}\|_*, \qquad \|\mathcal{T}_{h,q}(\boldsymbol{M};\boldsymbol{\Omega})\|_{\mathrm{op}} \leq 1.$$

For this fixed input, whenever $\theta_{s,h} > 0$, the alignment bound has the form required in Assumption 2.1, with

$$\gamma = 1 - \theta_{s,h} = 1 - \frac{A_{s,h}}{\delta\|\boldsymbol{M}\|_*},$$

where the second equality uses $A_{s,h} > 0$. The second condition holds with $\nu = 0$. It remains to verify that $\theta_{s,h} \leq 1$ and to discuss when $\theta_{s,h} > 0$. Positivity at each input, or over one finite trajectory, does not itself give a deterministic alignment constant bounded away from zero across runs and horizons. Theorem 3.5 uses such uniform constants. Under Frobenius normalization, Lemma B.1 supplies them independently of whether this particular spectral lower bound is positive.

Part I. Using $\delta \geq \|\boldsymbol{M}\|_{\mathrm{op}} = \sigma_1$ from Algorithm 1, we have

$$\delta\|\boldsymbol{M}\|_* = \delta\sum_{j=1}^{r}\sigma_j \geq \sigma_1\sum_{j=1}^{r}\sigma_j \geq \sum_{j=1}^{r}\sigma_j^2 \geq \sum_{j=1}^{s}\sigma_j^2 \geq [A_{s,h}]_+.$$

This proves $\theta_{s,h} \leq 1$.

Part II. We now characterize when the quantity $A_{s,h}$ is strictly positive. For a fixed target rank s, define the head and tail sums of squared singular values as

$$H_s := \sum_{j=1}^{s}\sigma_j^2, \qquad T_s := \sum_{j>s}\sigma_j^2.$$

Then

$$[A_{s,h}]_+ = \left[H_s - \frac{s}{p-1}\varrho_s^{4h} T_s\right]_+ .$$

Since $s < r$, we have $T_s > 0$. Hence $A_{s,h} > 0$ if and only if

$$H_s > \frac{s}{p-1}\varrho_s^{4h} T_s,$$

or equivalently,

$$\varrho_s^{4h} < \frac{(p-1)H_s}{sT_s}.$$

This condition shows two mechanisms that make the bound positive. The oversampling parameter p reduces the penalty through the factor $(p-1)^{-1}$, while power iteration reduces it through the factor ϱ_s^{4h}, which is small whenever there is a spectral gap at index s, namely

$$\varrho_s = \frac{\sigma_{s+1}}{\sigma_s} < 1.$$

Since h must be an integer, one may take

$$h = \begin{cases} 0, & \text{if } H_s > \dfrac{s}{p-1} T_s, \\ 1 + \left\lfloor \dfrac{\log\left(\dfrac{sT_s}{(p-1)H_s}\right)}{4\log(1/\varrho_s)} \right\rfloor, & \text{if } H_s \le \dfrac{s}{p-1} T_s \text{ and } \varrho_s < 1. \end{cases}$$

With this choice, $A_{s,h} = H_s - \frac{s}{p-1}\varrho_s^{4h} T_s$ is strictly positive. Finally, if $\varrho_s = 1$, then power iteration does not improve the positivity condition at that value of s, since $\varrho_s^{4h} = 1$ for all h. In this case, one must either choose a different target rank s at which a spectral gap is present, increase the oversampling parameter p, or rely on the no-power condition $H_s > \frac{s}{p-1} T_s$.

C Auxiliary Lemmas

C.1 Lemmas for Analysis of Exact Polar Decomposition

Throughout this subsection, let $d_0 := \min\{m, n\}$. Let $\mathcal{P}(\boldsymbol{M})$ denote an exact polar factor satisfying $\langle \boldsymbol{M}, \mathcal{P}(\boldsymbol{M})\rangle = \|\boldsymbol{M}\|_*$ and $\|\mathcal{P}(\boldsymbol{M})\|_{\mathrm{op}} \le 1$.

Lemma C.1. Define

$$\boldsymbol{S}_k := \widetilde{\boldsymbol{M}}_k - \nabla f(\boldsymbol{X}_k).$$

Then, under Assumption 3.1, Muon with exact polar decomposition satisfies

$$\eta \sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \le f(\boldsymbol{X}_0) - f_\star + \frac{Ld_0}{2} K\eta^2 + \eta\left(1 + \sqrt{d_0}\right) \sum_{k=0}^{K-1} \mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right]. \tag{14}$$

Proof. Since exact polar factors are invariant under multiplication by a positive scalar, the update can be written as $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta\mathcal{P}(\widetilde{\boldsymbol{M}}_k)$. Moreover, $\|\mathcal{P}(\widetilde{\boldsymbol{M}}_k)\|_F \le \sqrt{d_0}$, and hence

$$\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \le \eta^2 d_0.$$

By L-smoothness,

$$\begin{aligned} f(\boldsymbol{X}_{k+1}) &\le f(\boldsymbol{X}_k) - \eta\left\langle \nabla f(\boldsymbol{X}_k), \mathcal{P}(\widetilde{\boldsymbol{M}}_k)\right\rangle + \frac{Ld_0}{2}\eta^2 \\ &= f(\boldsymbol{X}_k) - \eta\left\langle \widetilde{\boldsymbol{M}}_k, \mathcal{P}(\widetilde{\boldsymbol{M}}_k)\right\rangle + \eta\left\langle \boldsymbol{S}_k, \mathcal{P}(\widetilde{\boldsymbol{M}}_k)\right\rangle + \frac{Ld_0}{2}\eta^2 \\ &= f(\boldsymbol{X}_k) - \eta\left\|\widetilde{\boldsymbol{M}}_k\right\|_* + \eta\left\langle \boldsymbol{S}_k, \mathcal{P}(\widetilde{\boldsymbol{M}}_k)\right\rangle + \frac{Ld_0}{2}\eta^2 \\ &\le f(\boldsymbol{X}_k) - \eta\left\|\widetilde{\boldsymbol{M}}_k\right\|_F + \eta\sqrt{d_0}\left\|\boldsymbol{S}_k\right\|_F + \frac{Ld_0}{2}\eta^2. \end{aligned}$$

Since $\|\widetilde{\boldsymbol{M}}_k\|_F = \|\nabla f(\boldsymbol{X}_k) + \boldsymbol{S}_k\|_F \geq \|\nabla f(\boldsymbol{X}_k)\|_F - \|\boldsymbol{S}_k\|_F$, we obtain

$$f(\boldsymbol{X}_{k+1}) \leq f(\boldsymbol{X}_k) - \eta \|\nabla f(\boldsymbol{X}_k)\|_F + \eta(1+\sqrt{d_0}) \|\boldsymbol{S}_k\|_F + \frac{Ld_0}{2}\eta^2.$$

Taking total expectation and summing from $k = 0$ to $K-1$ yields the claim. □

Lemma C.2. Define

$$\boldsymbol{D}_k := \nabla f(\boldsymbol{X}_{k-1}) - \nabla f(\boldsymbol{X}_k) \qquad \text{and} \qquad \boldsymbol{\xi}_k := \boldsymbol{G}_k - \nabla f(\boldsymbol{X}_k).$$

Then, for every $k \geq 1$,

$$\boldsymbol{S}_k = \beta^k \boldsymbol{S}_0 + \beta^2 \sum_{s=1}^{k} \beta^{k-s} \boldsymbol{D}_s + (1-\beta)\Big((1+\beta)\boldsymbol{\xi}_k + \sum_{s=1}^{k-1} \beta^{k+1-s}\boldsymbol{\xi}_s - \beta^k \boldsymbol{\xi}_0\Big).$$

Proof. Note that

$$\begin{aligned}
\boldsymbol{S}_k &= \widetilde{\boldsymbol{M}}_k - \nabla f(\boldsymbol{X}_k) \\
&= \beta\widetilde{\boldsymbol{M}}_{k-1} + (1-\beta)\big((1+\beta)\boldsymbol{G}_k - \beta\boldsymbol{G}_{k-1}\big) - \nabla f(\boldsymbol{X}_k) \\
&= \beta\left(\boldsymbol{S}_{k-1} + \nabla f(\boldsymbol{X}_{k-1})\right) + (1-\beta)\big((1+\beta)\boldsymbol{G}_k - \beta\boldsymbol{G}_{k-1}\big) - \nabla f(\boldsymbol{X}_k).
\end{aligned}$$

Using $\boldsymbol{G}_k = \nabla f(\boldsymbol{X}_k) + \boldsymbol{\xi}_k$ gives

$$\boldsymbol{S}_k = \beta\boldsymbol{S}_{k-1} + (1-\beta)\big((1+\beta)\boldsymbol{\xi}_k - \beta\boldsymbol{\xi}_{k-1}\big) + \beta^2\left(\nabla f(\boldsymbol{X}_{k-1}) - \nabla f(\boldsymbol{X}_k)\right).$$

Thus,

$$\boldsymbol{S}_k = \beta\boldsymbol{S}_{k-1} + \beta^2\boldsymbol{D}_k + (1-\beta)\big((1+\beta)\boldsymbol{\xi}_k - \beta\boldsymbol{\xi}_{k-1}\big).$$

Unrolling this recursion completes the proof. □

Lemma C.3. Under Assumptions 3.1 and 3.2, for every $0 \leq k \leq K-1$,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] \leq \beta^k \mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right] + \frac{\beta^2}{1-\beta} L\eta\sqrt{d_0} \\
+ \frac{2\sqrt{\pi}C_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Bigg[\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}} + \sigma_1(1-\beta)\Bigg((1+\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \\
+ \sum_{s=1}^{k-1}\beta^{k+1-s}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right] + \beta^k\|\nabla f(\boldsymbol{X}_0)\|_F\Bigg)\Bigg],
\end{aligned} \tag{15}$$

where $\Gamma_\alpha := (2^\alpha + 1)^{1/\alpha}$.

Proof. The case $k = 0$ is immediate. Let $k \geq 1$. By L-smoothness,

$$\|\boldsymbol{D}_s\|_F = \|\nabla f(\boldsymbol{X}_{s-1}) - \nabla f(\boldsymbol{X}_s)\|_F \leq L\|\boldsymbol{X}_s - \boldsymbol{X}_{s-1}\|_F \leq L\eta\sqrt{d_0}.$$

Therefore,

$$\beta^2 \sum_{s=1}^{k} \beta^{k-s}\mathbb{E}\left[\|\boldsymbol{D}_s\|_F\right] \leq \frac{\beta^2}{1-\beta} L\eta\sqrt{d_0}.$$

For this fixed k, define the deterministic coefficients

$$\begin{aligned}
a_0 &:= -(1-\beta)\beta^k, \\
a_s &:= (1-\beta)\beta^{k+1-s} \quad \text{for } 1 \leq s \leq k-1, \\
a_k &:= (1-\beta)(1+\beta).
\end{aligned} \tag{16}$$

For each finite s, the exact-polar step bound gives $\|\boldsymbol{X}_s - \boldsymbol{X}_0\|_F \le s\eta\sqrt{d_0}$; smoothness therefore ensures $\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F^{\alpha}\right] < \infty$. Applying Lemma A.9 with these coefficients gives

$$\mathbb{E}\left[\left\|\sum_{s=0}^{k} a_s \boldsymbol{\xi}_s\right\|_F\right] \le \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\left[\sigma_0\left(\sum_{s=0}^{k}|a_s|^\alpha\right)^{1/\alpha} + \sigma_1\sum_{s=0}^{k}|a_s|\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right]\right]. \tag{17}$$

Moreover, reindexing the middle sum and using $\beta^{\alpha k} \le \beta$ and $\beta^{\alpha j} \le \beta^j$ gives

$$\begin{aligned}\left(\sum_{s=0}^{k}|a_s|^\alpha\right)^{1/\alpha} &= (1-\beta)\left(\beta^{\alpha k} + \sum_{s=1}^{k-1}\beta^{\alpha(k+1-s)} + (1+\beta)^\alpha\right)^{1/\alpha} \\ &\le (1-\beta)\left(\sum_{j=1}^{k}\beta^j + 2^\alpha\right)^{1/\alpha} \le (1-\beta)\left(\frac{1}{1-\beta} + 2^\alpha\right)^{1/\alpha} \\ &= (1-\beta)^{\frac{\alpha-1}{\alpha}}\left(1 + (1-\beta)2^\alpha\right)^{1/\alpha} \le \Gamma_\alpha(1-\beta)^{\frac{\alpha-1}{\alpha}}.\end{aligned}$$

Here $\Gamma_\alpha = (1+2^\alpha)^{1/\alpha}$, consistently with the statement. Thus,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{s=0}^{k} a_s \boldsymbol{\xi}_s\right\|_F\right] \le \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Bigg[&\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}} + \sigma_1(1-\beta)\Bigg((1+\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \\ &+ \sum_{s=1}^{k-1}\beta^{k+1-s}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right] + \beta^k\|\nabla f(\boldsymbol{X}_0)\|_F\Bigg)\Bigg].\end{aligned}$$

Combining this bound with Lemma C.2 and the drift bound above yields (15). □

Lemma C.4. Under the conditions of Lemma C.3, we have

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] \le \frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta} + K\frac{\beta^2}{1-\beta}L\eta\sqrt{d_0} + K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}} \\ + \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\sigma_1\left(\|\nabla f(\boldsymbol{X}_0)\|_F + \sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]\right).\end{aligned} \tag{18}$$

Proof. We now sum each term of (15) for $k = 0, \dots, K-1$. Note that

$$\begin{aligned}\sum_{k=0}^{K-1}\beta^k\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right] &\le \frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta}, \\ \sum_{k=0}^{K-1}\frac{\beta^2}{1-\beta}L\eta\sqrt{d_0} &= K\frac{\beta^2}{1-\beta}L\eta\sqrt{d_0}, \\ \sum_{k=0}^{K-1}\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}} &= K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&(1-\beta)\sum_{k=0}^{K-1}\left((1+\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] + \sum_{s=1}^{k-1}\beta^{k+1-s}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right]\right) \\ &\le \sum_{k=0}^{K-1}(1-\beta^2)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] + \beta^2\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \\ &= \sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right],\end{aligned}$$

and

$$\sum_{k=0}^{K-1}(1-\beta)\beta^k\|\nabla f(\boldsymbol{X}_0)\|_F \le \|\nabla f(\boldsymbol{X}_0)\|_F.$$

Combining the above inequalities and using (15) gives (18). □

Theorem C.5. Suppose Assumptions 3.1 and 3.2 hold. Then, for any $\eta > 0$ and $\beta \in (0,1)$, Muon (1) with exact polar decomposition and batch size

$$B > \left(2\sqrt{\pi}(1+\sqrt{d_0})\mathcal{C}_\alpha^{1/\alpha}\sigma_1\right)^{\frac{\alpha}{\alpha-1}}$$

satisfies

$$\begin{aligned}\min_{0\le k\le K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \le{}& \frac{f(\boldsymbol{X}_0)-f_\star}{\eta\rho K} + \frac{Ld_0\eta}{2\rho} + (1+\sqrt{d_0})\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{(1-\beta)K\rho}\\ &+ \frac{\beta^2}{(1-\beta)\rho}L\eta\sqrt{d_0}(1+\sqrt{d_0})\\ &+ (1+\sqrt{d_0})\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{\rho B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}}\\ &+ (1+\sqrt{d_0})\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho B^{\frac{\alpha-1}{\alpha}}K},\end{aligned}$$

where

$$\rho := 1-(1+\sqrt{d_0})\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1}{B^{\frac{\alpha-1}{\alpha}}} > 0.$$

Proof. From (14) and (18),

$$\begin{aligned}\eta\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] &\le f(\boldsymbol{X}_0)-f_\star + \frac{Ld_0}{2}K\eta^2 + \eta(1+\sqrt{d_0})\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right]\\ &\le f(\boldsymbol{X}_0)-f_\star + \frac{Ld_0}{2}K\eta^2 + \eta(1+\sqrt{d_0})\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta}\\ &\quad + K\frac{\beta^2}{1-\beta}L\eta^2\sqrt{d_0}(1+\sqrt{d_0}) + \eta(1+\sqrt{d_0})K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}}\\ &\quad + \eta(1+\sqrt{d_0})\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\sigma_1\left(\|\nabla f(\boldsymbol{X}_0)\|_F + \sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]\right).\end{aligned}$$

Rearranging gives

$$\begin{aligned}\eta\rho\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] &\le f(\boldsymbol{X}_0)-f_\star + \frac{Ld_0}{2}K\eta^2 + \eta(1+\sqrt{d_0})\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta}\\ &\quad + K\frac{\beta^2}{1-\beta}L\eta^2\sqrt{d_0}(1+\sqrt{d_0}) + \eta(1+\sqrt{d_0})K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}}\\ &\quad + \eta(1+\sqrt{d_0})\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{B^{\frac{\alpha-1}{\alpha}}}.\end{aligned}$$

Dividing by $\eta\rho K$ and using

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \ge K\min_{0\le k\le K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]$$

completes the proof. □

C.2 Lemmas for Analysis of Inexact Polar Decomposition

Throughout this subsection, let $d_0 = \min\{m, n\}$ and let $\bar{\gamma} \in [0, 1)$ and $\bar{\nu} \geq 0$ be deterministic upper bounds satisfying

$$\gamma_k \leq \bar{\gamma} \qquad \text{and} \qquad \nu_k \leq \bar{\nu} \qquad (0 \leq k < K),$$

and set $\boldsymbol{S}_k := \widetilde{\boldsymbol{M}}_k - \nabla f(\boldsymbol{X}_k)$. The lemmas below only require these bounds over the current horizon. Theorem 3.5 applies them with constants that hold uniformly over all iterations and horizons.

Lemma C.6. Under Assumptions 3.1 and 2.1,

$$\eta\,(1-\bar{\gamma}) \sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq f(\boldsymbol{X}_0) - f_\star + \eta\left(1 + \sqrt{d_0}\,(1+\bar{\nu})\right) \sum_{k=0}^{K-1} \mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] + \frac{Ld_0}{2}\eta^2\,(1+\bar{\nu})^2\,K. \tag{19}$$

Proof. Let $\mathcal{G}_k = \sigma(\mathcal{F}_k, \boldsymbol{G}_k^1, \ldots, \boldsymbol{G}_k^B)$ be the history after the current mini-batch and before the polar draw, as defined in Section 2. Then $\boldsymbol{X}_k$, $\boldsymbol{M}_k$, $\widetilde{\boldsymbol{M}}_k$, and $\boldsymbol{S}_k$ are $\mathcal{G}_k$-measurable. By the fresh-seed oracle model preceding Assumption 2.1, for every integrable function ψ of the input and output,

$$\mathbb{E}\left[\psi(\boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)) \mid \mathcal{G}_k\right] = \mathbb{E}\left[\psi(\boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)) \mid \boldsymbol{M}_k\right].$$

Consequently, both the alignment and operator-norm bounds in Assumption 2.1 hold conditionally on $\mathcal{G}_k$.

By L-smoothness and the update rule,

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \mathcal{G}_k\right] &\leq f(\boldsymbol{X}_k) - \eta\mathbb{E}\left[\langle \nabla f(\boldsymbol{X}_k), \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \mathcal{G}_k\right] + \frac{L\eta^2}{2}\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_F^2 \mid \mathcal{G}_k\right] \\
&= f(\boldsymbol{X}_k) - \eta\mathbb{E}\left[\left\langle \widetilde{\boldsymbol{M}}_k, \mathcal{T}(\boldsymbol{M}_k)\right\rangle \mid \mathcal{G}_k\right] + \eta\mathbb{E}\left[\langle \boldsymbol{S}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \mathcal{G}_k\right] \\
&\quad + \frac{L\eta^2}{2}\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_F^2 \mid \mathcal{G}_k\right].
\end{aligned}$$

For the alignment term, using $\widetilde{\boldsymbol{M}}_k = (1-\beta)\boldsymbol{M}_k$ gives

$$\begin{aligned}
\mathbb{E}\left[\left\langle \widetilde{\boldsymbol{M}}_k, \mathcal{T}(\boldsymbol{M}_k)\right\rangle \mid \mathcal{G}_k\right] &= (1-\beta)\mathbb{E}\left[\langle \boldsymbol{M}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \mathcal{G}_k\right] \\
&\geq (1-\beta)(1-\gamma_k)\|\boldsymbol{M}_k\|_* = (1-\gamma_k)\|\widetilde{\boldsymbol{M}}_k\|_*.
\end{aligned}$$

For the error term, Cauchy-Schwarz, Jensen's inequality, and Assumption 2.1 imply

$$\begin{aligned}
\mathbb{E}\left[\langle \boldsymbol{S}_k, \mathcal{T}(\boldsymbol{M}_k)\rangle \mid \mathcal{G}_k\right] &\leq \|\boldsymbol{S}_k\|_F\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_F \mid \mathcal{G}_k\right] \\
&\leq \|\boldsymbol{S}_k\|_F\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_F^2 \mid \mathcal{G}_k\right]^{1/2} \leq \sqrt{d_0}(1+\nu_k)\|\boldsymbol{S}_k\|_F.
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_F^2 \mid \mathcal{G}_k\right] \leq d_0\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_k)\|_{\mathrm{op}}^2 \mid \mathcal{G}_k\right] \leq d_0(1+\nu_k)^2.$$

Combining these bounds,

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{X}_{k+1}) \mid \mathcal{G}_k\right] &\leq f(\boldsymbol{X}_k) - \eta(1-\gamma_k)\|\widetilde{\boldsymbol{M}}_k\|_* + \eta\sqrt{d_0}(1+\nu_k)\|\boldsymbol{S}_k\|_F + \frac{Ld_0}{2}\eta^2(1+\nu_k)^2 \\
&\leq f(\boldsymbol{X}_k) - \eta(1-\gamma_k)\|\widetilde{\boldsymbol{M}}_k\|_F + \eta\sqrt{d_0}(1+\nu_k)\|\boldsymbol{S}_k\|_F + \frac{Ld_0}{2}\eta^2(1+\nu_k)^2 \\
&\leq f(\boldsymbol{X}_k) - \eta(1-\gamma_k)\|\nabla f(\boldsymbol{X}_k)\|_F + \eta\left((1-\gamma_k) + \sqrt{d_0}(1+\nu_k)\right)\|\boldsymbol{S}_k\|_F \\
&\quad + \frac{Ld_0}{2}\eta^2(1+\nu_k)^2 \\
&\leq f(\boldsymbol{X}_k) - \eta(1-\bar{\gamma})\|\nabla f(\boldsymbol{X}_k)\|_F + \eta\left(1 + \sqrt{d_0}(1+\bar{\nu})\right)\|\boldsymbol{S}_k\|_F \\
&\quad + \frac{Ld_0}{2}\eta^2(1+\bar{\nu})^2.
\end{aligned}$$

Taking total expectation, summing over $k = 0, \ldots, K-1$, and using $f(\boldsymbol{X}_K) \geq f_\star$ gives (19). □

Lemma C.7. Under Assumptions 3.1, 2.1, and 3.2,

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] \le \frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta} + K\frac{\beta^2}{1-\beta}L\eta\sqrt{d_0}(1+\bar{\nu}) + K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{(\alpha-1)/\alpha}}\Gamma_\alpha\sigma_0(1-\beta)^{(\alpha-1)/\alpha} + \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{(\alpha-1)/\alpha}}\sigma_1\left(\|\nabla f(\boldsymbol{X}_0)\|_F + \sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]\right), \tag{20}$$

where $\Gamma_\alpha := (2^\alpha+1)^{1/\alpha}$.

Proof. Let

$$\boldsymbol{D}_s := \nabla f(\boldsymbol{X}_{s-1}) - \nabla f(\boldsymbol{X}_s), \qquad \boldsymbol{\xi}_s := \boldsymbol{G}_s - \nabla f(\boldsymbol{X}_s).$$

By Lemma C.2, for every $k \ge 1$,

$$\boldsymbol{S}_k = \beta^k\boldsymbol{S}_0 + \beta^2\sum_{s=1}^{k}\beta^{k-s}\boldsymbol{D}_s + (1-\beta)\left((1+\beta)\boldsymbol{\xi}_k + \sum_{s=1}^{k-1}\beta^{k+1-s}\boldsymbol{\xi}_s - \beta^k\boldsymbol{\xi}_0\right).$$

For $s \ge 1$, L-smoothness and the update $\boldsymbol{X}_s = \boldsymbol{X}_{s-1} - \eta\mathcal{T}(\boldsymbol{M}_{s-1})$ give

$$\mathbb{E}\left[\|\boldsymbol{D}_s\|_F\right] \le L\mathbb{E}\left[\|\boldsymbol{X}_s - \boldsymbol{X}_{s-1}\|_F\right] = L\eta\mathbb{E}\left[\|\mathcal{T}(\boldsymbol{M}_{s-1})\|_F\right] \le L\eta\sqrt{d_0}(1+\bar{\nu}),$$

where the last step uses Jensen's inequality and Assumption 2.1. Hence

$$\beta^2\sum_{s=1}^{k}\beta^{k-s}\mathbb{E}\left[\|\boldsymbol{D}_s\|_F\right] \le \frac{\beta^2}{1-\beta}L\eta\sqrt{d_0}(1+\bar{\nu}).$$

For the noise term, apply Lemma A.9 with the deterministic coefficients in (16) and the filtration $\mathcal{F}_s$ including all previous polar randomness. To verify its gradient-moment condition, note that, for each finite s,

$$\mathbb{E}\left[\|\boldsymbol{X}_s - \boldsymbol{X}_0\|_F^2\right] \le s\sum_{j=0}^{s-1}\mathbb{E}\left[\|\boldsymbol{X}_{j+1} - \boldsymbol{X}_j\|_F^2\right] \le s^2\eta^2 d_0(1+\bar{\nu})^2,$$

so smoothness gives $\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F^\alpha\right] < \infty$ for $\alpha \le 2$. Thus (17) also holds here. Combining it with the coefficient bound from Lemma C.3 and the drift bound above yields

$$\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] \le \beta^k\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right] + \frac{\beta^2}{1-\beta}L\eta\sqrt{d_0}(1+\bar{\nu}) + \frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{(\alpha-1)/\alpha}}\left[\Gamma_\alpha\sigma_0(1-\beta)^{(\alpha-1)/\alpha} + \sigma_1(1-\beta)\left((1+\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] + \sum_{s=1}^{k-1}\beta^{k+1-s}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_s)\|_F\right] + \beta^k\|\nabla f(\boldsymbol{X}_0)\|_F\right)\right].$$

Summing this inequality over $k = 0, \dots, K-1$ and using the same geometric-series bounds as in Lemma C.4 proves (20). □

D Missing Proofs of Section 3

D.1 Proof of Lemma 3.3

Lemma 3.3. For $\widetilde{\boldsymbol{C}}_k := (1-\beta)\boldsymbol{C}_k$ and $\widetilde{\boldsymbol{M}}_k := (1-\beta)\boldsymbol{M}_k$, we have, for all $k \ge 1$,

$$\widetilde{\boldsymbol{C}}_k = \beta\widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)\boldsymbol{G}_k \qquad \text{and} \qquad \widetilde{\boldsymbol{M}}_k = \beta\widetilde{\boldsymbol{M}}_{k-1} + (1-\beta)\big((1+\beta)\boldsymbol{G}_k - \beta\boldsymbol{G}_{k-1}\big).$$

Proof. From $\widetilde{\boldsymbol{C}}_k := (1-\beta)\boldsymbol{C}_k$, we obtain

$$\widetilde{\boldsymbol{C}}_k = (1-\beta)\boldsymbol{C}_k = (1-\beta)(\beta\boldsymbol{C}_{k-1} + \boldsymbol{G}_k) = \beta\widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)\boldsymbol{G}_k,$$

and from $\widetilde{\boldsymbol{M}}_k := (1-\beta)\boldsymbol{M}_k$, we get

$$\widetilde{\boldsymbol{M}}_k = (1-\beta)\boldsymbol{M}_k = (1-\beta)(\beta\boldsymbol{C}_k + \boldsymbol{G}_k) = \beta\widetilde{\boldsymbol{C}}_k + (1-\beta)\boldsymbol{G}_k.$$

For $k \geq 1$, we want to express $\widetilde{\boldsymbol{M}}_k$ in terms of $\widetilde{\boldsymbol{M}}_{k-1}$. Note that

$$\begin{aligned}\widetilde{\boldsymbol{M}}_k &= \beta\widetilde{\boldsymbol{C}}_k + (1-\beta)\boldsymbol{G}_k = \beta(\beta\widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)\boldsymbol{G}_k) + (1-\beta)\boldsymbol{G}_k \\ &= \beta^2\widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)(1+\beta)\boldsymbol{G}_k.\end{aligned}$$

Also $\widetilde{\boldsymbol{M}}_{k-1} = \beta\widetilde{\boldsymbol{C}}_{k-1} + (1-\beta)\boldsymbol{G}_{k-1}$, so

$$\beta^2\widetilde{\boldsymbol{C}}_{k-1} = \beta\widetilde{\boldsymbol{M}}_{k-1} - \beta(1-\beta)\boldsymbol{G}_{k-1}.$$

Therefore

$$\widetilde{\boldsymbol{M}}_k = \beta\widetilde{\boldsymbol{M}}_{k-1} + (1-\beta)\big((1+\beta)\boldsymbol{G}_k - \beta\boldsymbol{G}_{k-1}\big).$$

This completes the proof. □

D.2 Proof of Theorem 3.4

Theorem 3.4. Suppose Assumptions 3.1 and 3.2 hold. For a horizon $K \geq 2$, Muon (1) with exact polar decompositions, step size $\eta = 1/K^{(2\alpha-1)/(3\alpha-2)}$, momentum parameter $\beta = 1 - 1/K^{\alpha/(3\alpha-2)}$, and batch size $B > \{2\sqrt{\pi}(1+\sqrt{d_0})\mathcal{C}_\alpha^{1/\alpha}\sigma_1\}^{\alpha/\alpha-1}$ satisfies

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Big(&\frac{f(\boldsymbol{X}_0)-f_\star}{\rho_{\text{ex}}K^{\alpha-1/3\alpha-2}} + \frac{L}{\rho_{\text{ex}}K^{\alpha-1/3\alpha-2}} + \frac{\sigma_0}{\rho_{\text{ex}}B^{\alpha-1/\alpha}K^{\alpha-1/3\alpha-2}} \\ &+ \frac{L}{\rho_{\text{ex}}K^{2\alpha-1/3\alpha-2}} + \frac{\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho_{\text{ex}}K^{2\alpha-2/3\alpha-2}} + \frac{\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho_{\text{ex}}B^{\alpha-1/\alpha}K}\Big), \quad (2)\end{aligned}$$

where $\rho_{\text{ex}} = 1-(1+\sqrt{d_0})2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1/B^{\alpha-1/\alpha} > 0$ and $0 < \mathcal{C}_\alpha \leq 2$ is a scalar depending only on α.

Proof. Set $\hat{\beta} := 1-\beta$. Then $\hat{\beta} = K^{-\alpha/3\alpha-2}$, and Theorem C.5, with $\rho = \rho_{\text{ex}}$, gives

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq{}& \frac{f(\boldsymbol{X}_0)-f_\star}{\eta\rho_{\text{ex}}K} + \frac{Ld_0\eta}{2\rho_{\text{ex}}} + \Big(1+\sqrt{d_0}\Big)\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{\hat{\beta}K\rho_{\text{ex}}} \\ &+ \frac{L\eta\sqrt{d_0}(1+\sqrt{d_0})}{\hat{\beta}\rho_{\text{ex}}} + \Big(1+\sqrt{d_0}\Big)\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{\rho_{\text{ex}}B^{\frac{\alpha-1}{\alpha}}}\Gamma_\alpha\sigma_0\hat{\beta}^{\frac{\alpha-1}{\alpha}} \\ &+ \Big(1+\sqrt{d_0}\Big)\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho_{\text{ex}}B^{\frac{\alpha-1}{\alpha}}K}.\end{aligned}$$

Substituting $\eta = K^{-2\alpha-1/3\alpha-2}$ and $\hat{\beta} = K^{-\alpha/3\alpha-2}$, and using the convention that $\mathcal{O}(\cdot)$ hides constants depending only on α and d_0, yields

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(&\frac{f(\boldsymbol{X}_0)-f_\star}{\rho_{\text{ex}}K^{\alpha-1/3\alpha-2}} + \frac{L}{\rho_{\text{ex}}K^{2\alpha-1/3\alpha-2}} + \frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{\rho_{\text{ex}}K^{2\alpha-2/3\alpha-2}} \\ &+ \frac{L}{\rho_{\text{ex}}K^{\alpha-1/3\alpha-2}} + \frac{\sigma_0}{\rho_{\text{ex}}B^{\frac{\alpha-1}{\alpha}}K^{\alpha-1/3\alpha-2}} + \frac{\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho_{\text{ex}}B^{\frac{\alpha-1}{\alpha}}K}\Bigg).\end{aligned}$$

This completes the proof. □

D.3 Proof of Theorem 3.5

Theorem 3.5. Suppose Assumptions 2.1, 3.1, and 3.2 hold. Assume deterministic constants $\bar{\gamma} \in [0,1)$ and $0 \leq \bar{\nu} < \infty$, independent of the iteration k and horizon K, satisfy $\gamma_k \leq \bar{\gamma}$ and $\nu_k \leq \bar{\nu}$ in every run considered. Then, for any $\eta > 0$ and $\beta \in (0,1)$, Muon (1) with inexact polar decomposition and fixed batch size $B > \{2\sqrt{\pi}\big(1+\sqrt{d_0}(1+\bar{\nu})\big)\mathcal{C}_\alpha^{1/\alpha}\sigma_1/(1-\bar{\gamma})\}^{\alpha/\alpha-1}$ satisfies

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Big(&\frac{f(\boldsymbol{X}_0)-f_\star}{\rho\eta K} + \frac{\beta^2L(1+\bar{\nu})^2\eta}{\rho(1-\beta)} + \frac{(1+\bar{\nu})\sigma_0(1-\beta)^{\alpha-1/\alpha}}{\rho B^{\alpha-1/\alpha}} \\ &+ \frac{L\eta(1+\bar{\nu})^2}{\rho} + \frac{(1+\bar{\nu})\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho(1-\beta)K} + \frac{(1+\bar{\nu})\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho B^{\alpha-1/\alpha}K}\Big),\end{aligned}$$

where $\rho := 1-\bar{\gamma}-\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1/B^{\alpha-1/\alpha} > 0$ and $\mathcal{C}_\alpha \in (0,2]$ depends only on α.

Proof. By Lemma C.6,

$$\eta\,(1-\bar{\gamma})\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \le f(\boldsymbol{X}_0)-f_\star+\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\|\boldsymbol{S}_k\|_F\right] + \frac{Ld_0}{2}\eta^2\,(1+\bar{\nu})^2\,K.$$

Applying Lemma C.7 gives

$$\begin{aligned}
&\eta\,(1-\bar{\gamma})\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \\
&\quad\le f(\boldsymbol{X}_0)-f_\star+\frac{Ld_0}{2}\eta^2\,(1+\bar{\nu})^2\,K+\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta} \\
&\qquad+K\frac{\beta^2}{1-\beta}L\eta^2\sqrt{d_0}(1+\bar{\nu})\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right) \\
&\qquad+K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\alpha-1/\alpha}}\Gamma_\alpha\sigma_0(1-\beta)^{\alpha-1/\alpha}\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right) \\
&\qquad+\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\alpha-1/\alpha}}\sigma_1\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\left(\|\nabla f(\boldsymbol{X}_0)\|_F+\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]\right).
\end{aligned}$$

Rearranging the last term yields

$$\begin{aligned}
&\eta\left(1-\bar{\gamma}-\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1}{B^{\alpha-1/\alpha}}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \\
&\quad\le f(\boldsymbol{X}_0)-f_\star+\frac{Ld_0}{2}\eta^2\,(1+\bar{\nu})^2\,K+\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{1-\beta} \\
&\qquad+K\frac{\beta^2}{1-\beta}L\eta^2\sqrt{d_0}(1+\bar{\nu})\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right) \\
&\qquad+K\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\alpha-1/\alpha}}\Gamma_\alpha\sigma_0(1-\beta)^{\alpha-1/\alpha}\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right) \\
&\qquad+\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{B^{\alpha-1/\alpha}}\sigma_1\eta\left(1+\sqrt{d_0}\,(1+\bar{\nu})\right)\|\nabla f(\boldsymbol{X}_0)\|_F.
\end{aligned}$$

By the assumed lower bound on B, the coefficient in parentheses is $\rho > 0$. Dividing by $\eta\rho K$ and using

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \ge K\min_{0\le k\le K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right]$$

gives

$$\begin{aligned}
\min_{0\le k\le K-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] &\le \frac{f(\boldsymbol{X}_0)-f_\star}{\eta\rho K}+\frac{Ld_0}{2\rho}\eta\,(1+\bar{\nu})^2+\left(1+\sqrt{d_0}(1+\bar{\nu})\right)\frac{\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{(1-\beta)K\rho} \\
&\quad+\frac{\beta^2}{(1-\beta)\rho}L\eta\sqrt{d_0}(1+\bar{\nu})\left(1+\sqrt{d_0}(1+\bar{\nu})\right) \\
&\quad+\left(1+\sqrt{d_0}(1+\bar{\nu})\right)\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}}{\rho B^{\alpha-1/\alpha}}\Gamma_\alpha\sigma_0(1-\beta)^{\alpha-1/\alpha} \\
&\quad+\left(1+\sqrt{d_0}(1+\bar{\nu})\right)\frac{2\sqrt{\pi}\mathcal{C}_\alpha^{1/\alpha}\sigma_1\|\nabla f(\boldsymbol{X}_0)\|_F}{\rho B^{\alpha-1/\alpha}K}.
\end{aligned}$$

The stated bound follows by absorbing constants depending only on α and d_0 into $\mathcal{O}(\cdot)$. Because $\bar{\gamma}$ and $\bar{\nu}$ are uniform in the horizon, a fixed batch size satisfying the theorem's threshold makes $\rho > 0$ independent of K. The displayed initialization term remains explicit; interpreting the bound as a rate in K also requires $\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]$ to stay bounded across the chosen horizon-dependent runs. □

D.4 Proof of Corollary 3.6

Corollary 3.6. Suppose the assumptions and uniform polar bounds of Theorem 3.5 hold with $\sigma_1 = 0$. Set $\rho_0 := 1 - \bar{\gamma} > 0$, independent of K. For a horizon $K \geq 2$, Muon (1) with inexact polar decomposition, step size $\eta = \frac{1}{K^{3/4}}$, momentum parameter $\beta = 1 - \frac{1}{\sqrt{K}}$, and any batch size $B \geq 1$ satisfies

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(\frac{f(\boldsymbol{X}_0)-f_\star}{\rho_0 K^{\frac{1}{4}}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{3}{4}}} + \frac{(1+\bar{\nu})\mathbb{E}[\|\boldsymbol{S}_0\|_F]}{\rho_0\sqrt{K}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{1}{4}}} + \frac{(1+\bar{\nu})\sigma_0}{\rho_0 B^{\frac{\alpha-1}{\alpha}} K^{\frac{\alpha-1}{2\alpha}}}\Bigg).$$

Proof. Since $\sigma_1 = 0$, the batch-size condition in Theorem 3.5 is vacuous for any $B \geq 1$, and $\rho = 1 - \bar{\gamma} = \rho_0 > 0$ is independent of K under the uniform oracle bounds. Applying Theorem 3.5 gives

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(\frac{f(\boldsymbol{X}_0) - f_\star}{\rho_0\eta K} + \frac{L\eta(1+\bar{\nu})^2}{\rho_0} + \frac{(1+\bar{\nu})\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{\rho_0(1-\beta)K} + \frac{\beta^2 L\eta(1+\bar{\nu})^2}{\rho_0(1-\beta)} + \frac{(1+\bar{\nu})\sigma_0(1-\beta)^{\frac{\alpha-1}{\alpha}}}{\rho_0 B^{\frac{\alpha-1}{\alpha}}}\Bigg).$$

Substituting $\eta = K^{-3/4}$ and $1 - \beta = K^{-1/2}$ yields

$$\min_{0\leq k\leq K-1} \mathbb{E}\left[\|\nabla f(\boldsymbol{X}_k)\|_F\right] \leq \mathcal{O}\Bigg(\frac{f(\boldsymbol{X}_0) - f_\star}{\rho_0 K^{\frac{1}{4}}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{3}{4}}} + \frac{(1+\bar{\nu})\mathbb{E}\left[\|\boldsymbol{S}_0\|_F\right]}{\rho_0\sqrt{K}} + \frac{L(1+\bar{\nu})^2}{\rho_0 K^{\frac{1}{4}}} + \frac{(1+\bar{\nu})\sigma_0}{\rho_0 B^{\frac{\alpha-1}{\alpha}} K^{\frac{\alpha-1}{2\alpha}}}\Bigg),$$

where $\beta^2 \leq 1$ is absorbed into the constant. This proves the claim. □

E Experiment Details and Ablation Study

This appendix describes the implementation, optimizer hyperparameters, and training schedules used in the experiments of Section 4.1 and 4.2. We use the `modded-nanogpt` speedrun codebase of [20] and the CIFAR-10 Airbench codebase of [19]. We modify the orthogonalization methods inside Muon to expose the inexact solver and our randomized low-rank projection. All data pipelines, model architecture, and learning-rate schedules are preserved from the original code.

E.1 Randomized Kaczmarz-Style Sparse Sketching

The randomized Kaczmarz-style sparse sketch [9; 46] provides a more computationally efficient alternative to the dense Gaussian sketch $\boldsymbol{\Omega} \sim \mathcal{N}(0,1)^{n\times\ell}$ used in Algorithm 1. The resulting sketching procedure is summarized in Algorithm 2, where Steps 1–3 replace the Gaussian sketch construction in Step 2 of Algorithm 1.

E.2 Pretraining of nanoGPT on FineWeb

Inexact solver and randomized projection hyperparameters. For randomized Muon, after Bayesian hyperparameter tuning, we implement Algorithm 1 with the quintic-theoretical Newton-Schulz polynomial, using $q = 7$ inner iterations together with a randomized projection of target rank $s = 200$, oversampling parameter $p = 10$, and $h = 1$ power iterations. These are then applied to different Muon variants.

Hyperparameters and parameter routing. Following [21; 2], all Muon variants route the embedding and language-model-head layers through AdamW and apply Muon to the remaining

Algorithm 2 Lifted Randomized Kaczmarz-Style Sparse Sketching Polar Decomposition

Require: Matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ with rank r, target rank $1 \le s < r$, oversampling parameter $p \ge 2$ with $\ell = s + p \le \min(m, n)$, power iteration $h \ge 0$, and Newton-Schulz steps $q \ge 0$.

Set $\ell := s + p$ and compute column sampling probabilities

$$\pi_j := \frac{\|\boldsymbol{M}_{:,j}\|_2^2}{\|\boldsymbol{M}\|_F^2}, \qquad j = 1, \dots, n$$

Sample ℓ indices $i_1, \dots, i_\ell$ i.i.d. from $\boldsymbol{\pi}$ with replacement — **Project Down**

Set

$$\boldsymbol{\Omega}_{:,k} := \frac{e_{i_k}}{\sqrt{\ell\, \pi_{i_k}}}, \qquad k = 1, \dots, \ell,$$

where e_j is the j-th canonical basis vector

Form $\boldsymbol{Y} := (\boldsymbol{M}\boldsymbol{M}^\top)^h \boldsymbol{M}\boldsymbol{\Omega}$, and compute orthonormal basis $\boldsymbol{Q} := \mathrm{orth}(\boldsymbol{Y})$

Set $\boldsymbol{B} := \boldsymbol{Q}^\top \boldsymbol{M} \in \mathbb{R}^{\ell\times n}$

Set $\boldsymbol{Z}_0 := \boldsymbol{B}/\|\boldsymbol{B}\|_{\mathrm{op}}$

for $t = 0, 1, \dots, q-1$ **do** — **Iterative Approximation**

  $\boldsymbol{Z}_{t+1} := \varphi(\boldsymbol{Z}_t)$

end for

Return $\mathcal{T}_{h,q}(\boldsymbol{M}; \boldsymbol{\Omega}) := \boldsymbol{Q}\boldsymbol{Z}_q$ — **Project Up**

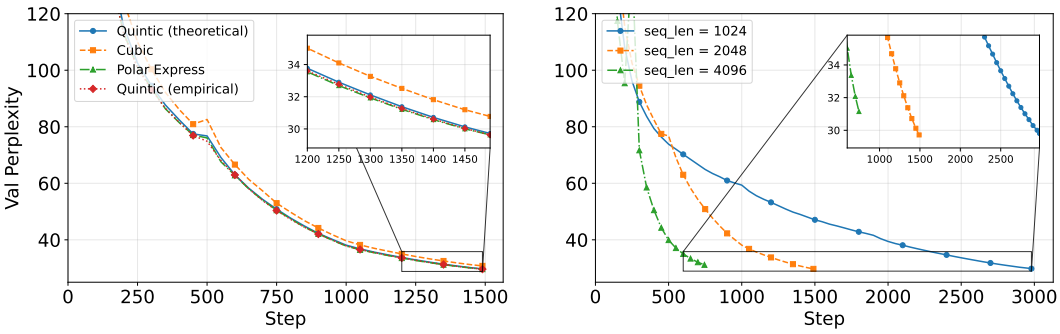
matrix-shaped parameters. The AdamW and SGD-Nesterov baselines instead apply their respective optimizer to every parameter. The Muon learning rate and momentum are tuned via Bayesian optimization, yielding learning rate around 3.25×10^{-2} and momentum around $\beta = 0.9665$; the SGD-Nesterov baseline uses learning rate around 4.18×10^{-4} with Nesterov momentum around 0.86, and the AdamW baseline uses learning rate around 1.80×10^{-3}. All remaining settings follow the `modded-nanogpt` speedrun configuration of [20]: training sequence length 2048 tokens, the three-stage training schedule, and a total of 1490 training iterations. Each configuration is averaged over 5 random seeds, and validation perplexity is computed on a held-out shard of $10{,}485{,}760$ FineWeb tokens. All nanoGPT runs use $4\times$ NVIDIA H100 GPUs, and the wall-clock times reported in Table 2 are end-to-end training times measured on that configuration. For the inexact-solver ablation in Figure 3a, the quintic-empirical solver uses coefficients $(a, b, c) = (3.4445, -4.7750, 2.0315)$ at every step, while the PolarExpress solver uses iteration-dependent coefficients $\{(a_t, b_t, c_t)\}_{t=1}^{9}$ given by

$$(8.1566, -22.4833, 15.8788), \quad (4.0429, -2.8089, 0.5000), \quad (3.8917, -2.7725, 0.5061),$$
$$(3.2858, -2.3681, 0.4645), \quad (2.3005, -1.6112, 0.3833), \quad (1.8631, -1.2042, 0.3422),$$
$$(1.8383, -1.1779, 0.3397), \quad (1.8382, -1.1779, 0.3396), \quad (1.8750, -1.2500, 0.3750).$$

E.3 CNN ON CIFAR-10

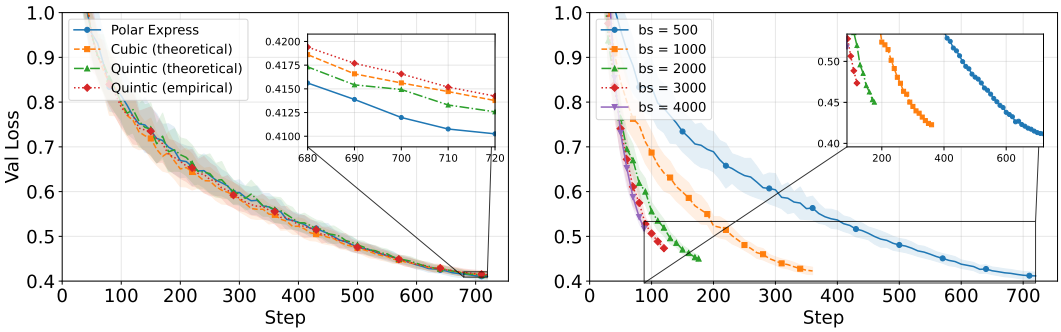
Inexact solver and randomized projection hyperparameters. After hyperparameter tuning, we select the quintic-theoretical Newton-Schulz solver with $q = 7$ inner iterations and a randomized projection of target rank $s = 128$, oversampling parameter $p = 10$, and $h = 2$ power iterations. These are then applied to different Muon variants.

Hyperparameters and parameter routing. Following the parameter-routing convention of the Airbench codebase [19], all Muon variants apply Muon to the 4D convolutional filter weights and route the remaining parameters through SGD with Nesterov momentum, whereas the AdamW and SGD baselines apply their respective optimizer to every parameter. After hyperparameter tuning, we use a batch size of 500 for the main comparison reported in Table 2 and Figure 1. Each configuration is averaged over 50 random seeds. All CIFAR-10 runs use a single NVIDIA A100 GPU, and the wall-clock times reported in Table 2 are end-to-end training times measured on that configuration. For the inexact-solver ablation in Figure 4a, the quintic-empirical solver uses coefficients $(a, b, c) = (3.4445, -4.7750, 2.0315)$ at every step, while the PolarExpress solver uses



(a) Inexact solver (b) Sequence length

Figure 3: nanoGPT ablations for randomized Muon with Nesterov momentum: validation perplexity versus training steps when each hyperparameter is varied in turn and the others are held at the default configuration. Each curve is the mean over 5 random seeds. The insets zoom into the final iterations for a clearer comparison.



(a) Inexact solver (b) Batch size

Figure 4: CIFAR-10 ablations for randomized Muon with Nesterov momentum: validation loss versus training steps when each hyperparameter is varied in turn and the others are held at the default configuration. Each curve is the mean over 50 random seeds. The insets zoom into the final iterations for a clearer comparison.

iteration-dependent coefficients $\{(a_t, b_t, c_t)\}_{t=1}^{9}$ given by

$$(8.2872, -23.5959, 17.3004), \quad (4.1071, -2.9478, 0.5448), \quad (3.9487, -2.9089, 0.5518),$$
$$(3.3184, -2.4885, 0.5100), \quad (2.3007, -1.6689, 0.4188), \quad (1.8913, -1.2680, 0.3768),$$
$$(1.8750, -1.2500, 0.3750), \quad (1.8750, -1.2500, 0.3750), \quad (1.8750, -1.2500, 0.3750).$$

E.4 NANOGPT BENCHMARK ABLATION STUDY.

To assess the sensitivity of randomized Muon with Nesterov's momentum to its hyperparameters on the nanoGPT benchmark, we present further ablation results in Figure 3. As shown in Figure 3a, three of the four inexact solvers (Quintic theoretical, PolarExpress, and Quintic empirical) trace essentially identical convergence curves, while Cubic lags slightly behind throughout training. Figure 3b compares three sequence lengths (1024, 2048, 4096): the two shorter sequences reach similar final perplexity while 4096 ends visibly higher. Because longer sequences process proportionally more tokens per step, they stop in fewer total steps.

E.5 CIFAR-10 BENCHMARK ABLATION STUDY.

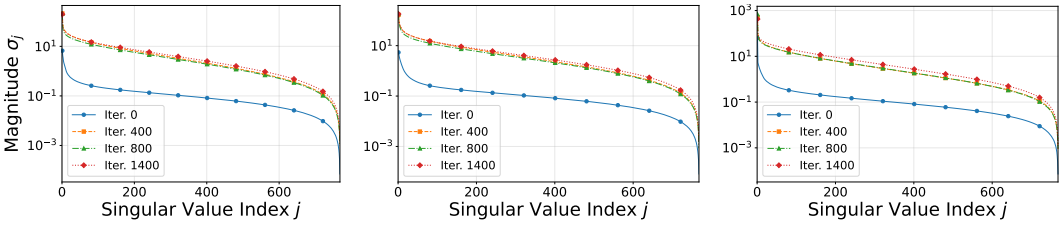
We also investigate the sensitivity of randomized Muon with Nesterov's momentum to its hyperparameters on the CIFAR-10 benchmark; we present further ablation results in Figure 4. As shown in Figure 4a, the four inexact solvers achieve comparable performance on this small-scale CNN:

PolarExpress attains the lowest final validation loss with Quintic (theoretical) close behind, and their convergence trajectories are nearly indistinguishable. Figure 4b shows that smaller batch sizes reach a lower final validation loss but at a slower per-step rate, indicating that the additional update steps afforded by small batches help locate better minima at the cost of slower per-step progress.

E.6 Empirical Study of the Inexact-Polar Quantities on nanoGPT

Theorem 3.5 uses deterministic oracle bounds that hold uniformly across iterations and horizons. To examine alignment in practice, we show the singular spectra of the momentum matrices every 50 training steps on nanoGPT in Figure 5, together with the realized alignment deficit $\widehat{\gamma}_k = 1 - \langle M_k, \mathcal{T}(M_k)\rangle / \|M_k\|_*$ for nonzero M_k in Figure 6a and Table 3. These sampled diagnostics differ from the deterministic conditional-expectation bounds γ_k in Assumption 2.1; a finite observed maximum does not certify a uniform bound. Lemma B.1 establishes such bounds for the stated Gaussian and full-space theoretical solvers under Frobenius normalization. We also plot the spectral bound of Proposition 2.3 along the same run in Figure 6b. All statistics are averaged over the same 5 random seeds used in Section 4.1.

Singular spectra of the momentum matrices. Figure 5 shows the spectra of M_k for the Q, K, and V matrices. This visualization partially supports the low rank nature of the momentum updates, as the singular values decay rapidly.



(a) Q Matrix (b) K Matrix (c) V Matrix

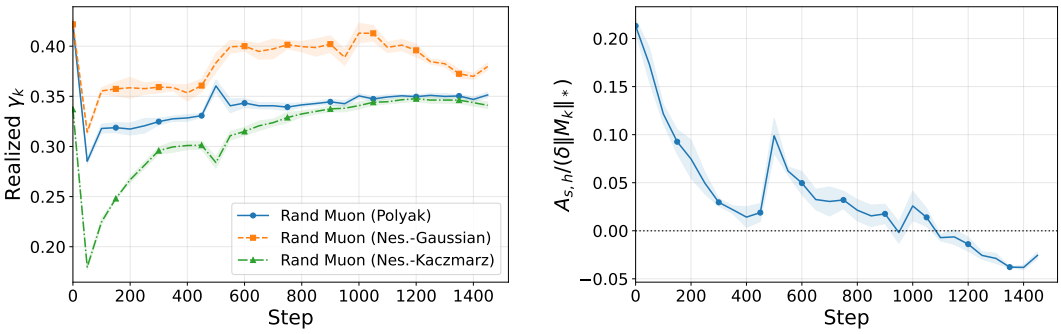
Figure 5: Singular values of the nanoGPT momentum matrices M_k for randomized Muon (Nes.-Gaussian), on a logarithmic vertical axis. Each curve is averaged over the ten attention layers within a seed and then over five seeds. Each attention projection is a 768×768 matrix of full rank 768.

Realized alignment deficits along the trajectory. Figure 6a and Table 3 report $\widehat{\gamma}_k$. After the first few hundred steps, the values are stable, and the reported means of the per-seed maxima are below 0.59 for every randomized variant. This indicates favorable alignment on the sampled iterates. These diagnostics do not determine the theorem's descent coefficient ρ, which also depends on σ_1 and B; the identity $\rho = 1 - \bar{\gamma}$ holds only when $\sigma_1 = 0$.

Table 3: Aggregate realized alignment deficit $\widehat{\gamma}_k$ over the nanoGPT run. Each statistic is first computed within each seed over all sampled training steps and then aggregated across seeds.

Method	$\mathrm{mean}_k\, \widehat{\gamma}_k$	$\max_k \widehat{\gamma}_k$
Rand Muon (Polyak)	0.3409 ± 0.0021	0.5572 ± 0.0093
Rand Muon (Nes.-Gaussian)	0.3824 ± 0.0019	0.5887 ± 0.0102
Rand Muon (Nes.-Kaczmarz)	0.3136 ± 0.0028	0.5831 ± 0.0107

The bound of Proposition 2.3. Figure 6b shows $A_{s,h}/(\delta\|M_k\|_*)$ without the positive part, at the values $s = 200$, $p = 10$ and $h = 1$, which were tuned for validation perplexity rather than to keep this quantity positive. With these hyperparameters, it does become mildly negative late in training. Increasing p, or increasing h when $\varrho_s < 1$, can improve this spectral bound under the conditions in Appendix B.4. Its positivity is not required for the separate uniform Gaussian bound in Lemma B.1.



(a) Realized alignment deficit $\widehat{\gamma}_k$

(b) Bound of Proposition 2.3

Figure 6: (a) Realized alignment deficit $\widehat{\gamma}_k$ over training for the three randomized variants. (b) The quantity $A_{s,h}/(\delta\|M_k\|_*)$ of Proposition 2.3 for randomized Muon (Nes.-Gaussian) at the tuned setting $s = 200$, $p = 10$, $h = 1$.

F LIMITATIONS

The inexact-polar theory depends on the abstract quantities γ_k and ν_k, but we do not provide an adaptive rule for choosing the number of polar iterations q during training. Moreover, our analysis of randomized polar decomposition is primarily developed for Gaussian sketches; extending the theory to sparse or Kaczmarz-style sketches remains an open problem. The current results also do not optimize the choice of the rank s, the oversampling parameter p, or the number of power iterations h, all of which could be selected adaptively in practice.