CONVERGENCE OF MUON UNDER RANDOM RESHUFFLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon, an optimizer that orthogonalizes its momentum matrix, is increasingly used to train large models, often with several passes over a fixed data pool under random reshuffling (RR), which visits every example once per pass. For stochastic gradient descent, the number of gradient evaluations needed for gradient norm at most ϵ scales as ϵ^{-4} under sampling with replacement (IID) and as ϵ^{-3} under reshuffling. Whether this gain carries over to Muon's nonlinear update was open, since RR proofs need updates linear in the sampled gradients and Muon's existing rates assume IID sampling. We give, to our knowledge, the first finite-horizon stationarity guarantee for Muon under RR on smooth nonconvex finite sums, assuming only that each component is smooth and bounded below. The sufficient query budget retains the ϵ^{-3} dependence of reshuffled SGD. The proof tracks the momentum buffer before orthogonalization, where sampling noise cancels over each pass, and a potential-function argument controls the changing spread of the component gradients. In language-model training on a fixed data pool, RR reaches lower training and validation loss than IID after the same number of updates.

1 INTRODUCTION

Muon is increasingly used for large-model training. It replaces the matrix-valued momentum by its polar factor, the matrix with the same singular vectors and every positive singular value set to one, computed approximately by a few Newton–Schulz iterations (Jordan et al., 2024; Liu et al., 2025; Tuddenham et al., 2022). Multi-pass training on a fixed data pool, as in fine-tuning and data-constrained pretraining (Muennighoff et al., 2023), commonly uses *random reshuffling* (RR), in which a fresh permutation processes each objective component (an example or a minibatch) once per pass, or epoch.

RR has a stationarity theory for stochastic gradient descent (SGD) on general smooth nonconvex finite sums, but no comparable Muon guarantee is known. For SGD on a finite sum of N components, standard sufficient query bounds (counts of gradient evaluations) for expected gradient norm at most ϵ improve from order ϵ^{-4} under independent uniform sampling with replacement (IID) to order $\sqrt{N}/\epsilon^3$ under RR. With other problem constants fixed, the latter is smaller below an accuracy threshold of order $N^{-1/2}$ (Mishchenko et al., 2020; Nguyen et al., 2021). Existing Muon analyses under IID sampling assume that each sampled gradient is an unbiased estimate of the full gradient conditional on the past (Shen et al., 2026; Kim & Oh, 2026; Chang et al., 2025). IID sampling satisfies this condition, whereas within-epoch RR sampling need not. This leaves the following question.

Does the reshuffling advantage of SGD carry over to Muon's nonlinear update?

We give, to our knowledge, the first general smooth nonconvex finite-sum stationarity guarantee for Muon under RR. The result bounds the expected nuclear norm (sum of singular values) of the gradient at the output and retains RR-SGD's ϵ^{-3} accuracy dependence. Two obstacles separate this guarantee from existing SGD and IID Muon analyses.

First obstacle: preserving cancellation through nonlinear polar updates. Gradient cancellation does not imply cancellation of Muon's polar directions. At a common evaluation point, deviations of component gradients from the full gradient sum to zero over a full pass. SGD's linear update uses this identity directly. For SGD with momentum, a shifted iterate absorbs the buffer and again moves linearly in the sampled gradients (Qiu et al., 2024). This reduction does not carry over through

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.

the nonlinear polar map (Appendix A). We must instead preserve cancellation inside the buffer and control how its error affects polar descent across passes.

Second obstacle: controlling the changing spread of component gradients. The proof must control two interdependent quantities. The error between the exponentially weighted gradient buffer and the full gradient depends on the spread of component gradients around their average. This spread, or gradient heterogeneity, changes along the run and need not be globally bounded (Appendix C.2). Component smoothness and lower bounds control it by the current objective gap, measured above the average component lower bound. But the descent bound for that gap also contains the buffer error. Our analysis must therefore control both quantities together.

Contributions. We give a stationarity guarantee and two technical contributions that establish it.

- **First general nonconvex stationarity guarantee for Muon under RR.** Theorem 1 bounds the expected nuclear norm of the gradient along Muon (Algorithm 1) on general smooth nonconvex finite sums, for every fixed number of epochs K, assuming only component smoothness and lower bounds, without a uniform variance assumption. The sufficient budget has the same ϵ^{-3} accuracy dependence as RR-SGD, compared with the ϵ^{-4} IID Muon bound of Kim & Oh (2026) for positive gradient variance and initial gap.
- **Preserving full-pass cancellation with carried momentum.** Over a full pass, the sampling noise of the component gradients at a common evaluation point sum to zero. Momentum weights these errors unequally, so its buffer keeps a residual, but the residual depends only on the differences between the weights. When the momentum memory spans at least one pass, these differences are small, and the buffer noise at the end of the pass (the endpoint) is smaller than under IID sampling by a factor of the order of the pass length divided by the memory length. This reduction survives carrying the buffer across passes and still controls polar descent, without momentum resets or extra gradient queries (Section 4.2, Steps 1–3).
- **Joint stability of the objective gap and the buffer error.** The bound on the buffer error depends on the spread of the component gradients, which changes as the iterates move. We control this interaction through a weighted sum of the objective gap and the buffer error, with the weight chosen so that the per-pass decay of the buffer error exactly cancels the buffer-error term charged by the descent inequality. Under the theorem's schedule, the resulting recursion bounds both quantities by the initial objective gap alone, without a uniform variance assumption and without losing the reshuffling gain (Section 4.2, Step 4).

Together, the two estimates give a rate for a stepsize fixed throughout each run, since endpoint cancellation alone leaves the noise scale uncontrolled and stability alone need not preserve the reshuffling gain. In our language-model study, RR attains lower training-loss area under the curve and final losses than IID at matched update budgets (Section 5).

2 Related Work

Random reshuffling theory. For updates without gradient corrections, RR theory covers SGD and Euclidean momentum, and these proofs rely on the update being linear in the sampled gradients, a property that Muon's polar update lacks. Such results exist for convex and nonconvex SGD (Gürbüzbalaban et al., 2021; HaoChen & Sra, 2019; Nagaraj et al., 2019; Mishchenko et al., 2020; Nguyen et al., 2021; Liu, 2026) and for SGD with momentum (Tran et al., 2021; Qiu et al., 2024). SMG (Tran et al., 2021) keeps its momentum term fixed within each epoch and refreshes it at epoch ends under a generalized variance condition. Qiu et al. (2024) analyze momentum that is carried across epochs without reset and with a fixed parameter, using a shifted iterate linear in the sampled gradients.

The closest RR result for polar directions concerns implicit bias rather than general nonconvex stationarity. Li et al. (2026) combine carried momentum with spectral-norm steepest-descent directions, as in Muon, but for separable linear classification, and conclude which separator is selected rather than a stationarity rate on general smooth nonconvex finite sums. Their analysis also uses epochwise zero-sum cancellation. We instead quantify the conditional second moment of the endpoint noise, preserve its gain through the contraction of the carried buffer, and control the remaining trajectory-dependent noise to obtain a general nonconvex stationarity bound.

Table 1: **Closest guarantees under a common query budget.** T counts unit-batch gradient queries, N the number of fixed components, and $r = \min\{m, p\}$ for an $m \times p$ matrix. The product of the last two columns is each row's leading bound on the expected nuclear gradient norm at fixed N, initial gap, variance bound (IID rows), momentum parameter (first row) and operator–nuclear smoothness L (Assumption 1), with lower-order terms suppressed. RR rows assume component lower bounds in place of a variance bound. Appendix F gives exact statements, and Appendix C.1 the rank factors at fixed Frobenius smoothness. † Frobenius result converted with $\|G\|_* \le r^{1/2}\|G\|_F$. ‡ Includes the Muon-MVR2 dimension coefficient and its two queries per update.

Paper	Algorithm	Sampling	Uniform variance bound	r factor	Decay in T
Qiu et al. (2024)	SGD with momentum	RR	Not required	$r^{1/2}$ †	$T^{-1/3}$
Chang et al. (2025)	Muon-MVR2	IID	Required	$r^{3/2}$ ‡	$T^{-1/3}(1+\log T)$
Kim & Oh (2026)	Muon	IID	Required	$r^{1/4}$	$T^{-1/4}$
This work	**Muon**	**RR**	**Not required**	$\boldsymbol{r^{1/6}}$	$\boldsymbol{T^{-1/3}}$

Muon theory under IID sampling. Muon convergence theory under IID sampling relies on gradients that are unbiased conditional on the past, which within-epoch RR sampling does not provide. Muon belongs to the family of methods whose update minimizes a linearized loss over a norm ball, the linear minimization oracle (LMO) family (Bernstein & Newhouse, 2024; Pethick et al., 2025a). Its convergence theory covers orthogonalized directions and their Newton–Schulz (NS) approximations (Li & Hong, 2025; Shen et al., 2026; Sato et al., 2025; Kim & Oh, 2026) as well as inexact orthogonalization (Shulgin et al., 2026a; Do et al., 2026). The underlying momentum-error argument, also used for normalized SGD with momentum (Cutkosky & Mehta, 2020), takes conditional expectations step by step, whereas our RR proof conditions at epoch starts.

State-dependent noise and first-moment stability. First-moment stability controls an expected error norm without squaring it. Component smoothness and lower bounds give an objective-gap bound on finite-sum variance in RR-SGD (Mishchenko et al., 2020, Proposition 2). For SGD with momentum under RR, Qiu et al. (2024) use a Lyapunov sequence built from the loss at the shifted iterate and a quadratic displacement term. Shulgin et al. (2026b, Theorem A.1) bound the expected unsquared momentum error through the average objective gap $f - f^*$ for momentum-LMO methods, including Muon, under an unbiased oracle whose second moment may grow with that gap. This covers fresh unbiased sampling on our finite sums, but not within-epoch RR sampling. Our epochwise potential couples the objective gap with the nuclear-norm buffer error. Its expected value controls heterogeneity without losing endpoint cancellation (Eq. (7) and Appendix F).

Gradient corrections and sign methods. Corrections for normalized or sign updates use extra gradient queries, whereas our momentum buffer, an exponential moving average (EMA) of the sampled component gradients, is carried across epochs and uses one component gradient per update, without correction. Momentum-free RR with elementwise sign updates can fail even on a smooth strongly convex finite sum (Qin et al., 2023, Proposition 3.5). SignRVR (Qin et al., 2023) and variance-reduced RR for SGD (Malinovsky et al., 2023) use one full gradient per epoch as a reference, and variance-reduced Muon methods correct the momentum by evaluating each sampled component gradient at two points (Chang et al., 2025; Qian et al., 2025). Table 1 lists the closest guarantees, and Appendix F compares their metrics, assumptions and query costs and discusses other results.

3 Preliminaries

Notation. Write $\mathbb{N} = \{1, 2, \ldots\}$ and $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}$. For $m \times p$ matrices, $r = \min\{m, p\}$ is the ambient rank bound, not an observed rank. The norms $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_F$, and $\|\cdot\|_*$ are operator (spectral), Frobenius, and nuclear norms. With tr denoting trace, the Frobenius inner product is $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$.

For a rank-s matrix W, write its compact singular value decomposition (SVD) as $W = U_s \Sigma_s V_s^\top$ over the positive singular values. The canonical partial polar is $\mathrm{Polar}(W) = U_s V_s^\top$, with $\mathrm{Polar}(0) = 0$.

Expectation $\mathbb{E}$ averages over algorithmic and output randomness. Unless parameters are explicitly held fixed, $\mathcal{O}(\cdot)$ and $\Theta(\cdot)$ hide only universal numerical constants, and logarithmic factors are explicit.

3.1 Problem setting: Nonconvex finite-sum optimization

We study nonconvex finite-sum optimization of a matrix-valued parameter $W \in \mathbb{R}^{m\times p}$,

$$\min_{W\in\mathbb{R}^{m\times p}} f(W) := \frac{1}{N}\sum_{i=1}^{N} f_i(W),$$

where $N \geq 2$ counts fixed, continuously differentiable component losses $f_i : \mathbb{R}^{m\times p} \to \mathbb{R}$, and $f^* = \inf_W f(W) > -\infty$. As in existing Muon theory, we analyze one matrix rather than layer-wise parameters (Li & Hong, 2025; Sato et al., 2025).

Definition 1 (ϵ-stationary output)**.** *For a given tolerance $\epsilon > 0$, we call a possibly random $W \in \mathbb{R}^{m\times p}$ an ϵ-stationary output (in the nuclear norm) if $\mathbb{E}\|\nabla f(W)\|_* \leq \epsilon$.*

Nuclear stationarity matches the operator-norm update geometry and implies Frobenius stationarity because $\|G\|_F \leq \|G\|_*$ for every matrix G (Shen et al., 2026).

3.2 Muon under random reshuffling

Muon optimizer. Muon (Jordan et al., 2024) forms a matrix-valued momentum buffer and takes a step along its polar factor. This is a spectral-norm steepest-descent direction (Bernstein & Newhouse, 2024), since for any buffer M, $\mathrm{Polar}(M)$ maximizes $\langle M, O\rangle_F$ over $\|O\|_{\mathrm{op}} \leq 1$ (Proposition 1). Muon uses orthogonalization, unlike coordinatewise Adam (Kingma & Ba, 2015) or the curvature-based Shampoo and SOAP (Gupta et al., 2018; Vyas et al., 2025). For the analysis, we use the exact polar factor and the classical momentum recursion of Algorithm 1, as in Shen et al. (2026). The public implementation adds optional Nesterov acceleration and a shape-dependent stepsize multiplier, which we neither analyze nor use (Appendix G.1).

Random reshuffling. For a run of K epochs, at the start of epoch $k \in [K]$, a permutation $\pi^k = (\pi_1^k, \ldots, \pi_N^k)$ of $[N]$ is drawn uniformly and independently of the past, and the epoch's N updates use the components in this order. Each component is thus used exactly once per epoch, whereas IID sampling draws each index independently and uniformly from $[N]$.

Fixed components and query cost. A component may itself be a minibatch, but its membership stays fixed throughout the run. An epoch permutes this entire component list, rather than a newly selected subset. Each update uses one component-gradient query, and no full-gradient reference or extra gradient correction is computed.

Momentum buffer and update direction. Algorithm 1 carries the classical momentum buffer M across epochs with coefficient $\beta \in [0, 1)$. Multiplying M by $1 - \beta$ gives an EMA of the gradients. Both conventions yield the same update because $\mathrm{Polar}((1-\beta)M) = \mathrm{Polar}(M)$, so this normalization is an analysis device, not a change in the stepsize. Theorem 1 specifies a stepsize η and momentum β held fixed within each run, and Appendix E.3 treats approximate directions.

Algorithm 1 Muon under random reshuffling

Require: total epochs K, number of components N, momentum $\beta \in [0, 1)$, initial weight $\tilde{W}_0 \in \mathbb{R}^{m\times p}$, stepsize $\eta \geq 0$.
Initialize $W_0^1 := \tilde{W}_0$ and the momentum buffer $M_0^1 := \mathbf{0}$
for $k = 1$ to K **do**
  Sample $\pi^k = (\pi_1^k, \ldots, \pi_N^k)$ uniformly from the permutations of $[N]$
  for $i = 1$ to N **do**
    $G_i^k := \nabla f_{\pi_i^k}(W_{i-1}^k)$
    $M_i^k := \beta M_{i-1}^k + G_i^k$
    $O_i^k := \mathrm{Polar}(M_i^k)$
    $W_i^k := W_{i-1}^k - \eta\, O_i^k$
  end for
  Carry over $W_0^{k+1} := W_N^k$ and $M_0^{k+1} := M_N^k$ without reset
end for

4 MAIN THEORETICAL RESULTS

Our RR guarantee uses only the two assumptions below, without a uniform variance assumption.

Assumption 1 (Lipschitz smoothness)**.** *Each component loss f_i is L-smooth for a common $L > 0$ with respect to the norm pair $(\|\cdot\|_{\mathrm{op}}, \|\cdot\|_*)$, that is, for all $W, V \in \mathbb{R}^{m\times p}$ and all $i \in [N]$,*

$$\|\nabla f_i(W) - \nabla f_i(V)\|_* \le L\|W - V\|_{\mathrm{op}}.$$

Discussion of Assumption 1. Operator–nuclear smoothness matches polar-update geometry (Shen et al., 2026). It gives the average f the same smoothness and each component Frobenius L-smoothness. Conversely, Frobenius L_F-smoothness implies it with $L = rL_F$. Proposition 2 proves both directions and the necessity of r in general, so rank comparisons fix L rather than L_F.

Assumption 2 (Lower boundedness)**.** *Each component loss f_i is lower bounded, that is, there exists $f_i^* > -\infty$ such that $f_i(W) \ge f_i^*$ for all $W \in \mathbb{R}^{m\times p}$.*

Discussion of Assumption 2. Component lower bounds give a state-dependent self-bound in place of a uniform variance assumption. Assumptions 1–2 imply $\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(W) - \nabla f(W)\|_F^2 \le 2L\big(f(W) - \frac{1}{N}\sum_{i=1}^N f_i^*\big)$ for every W (Lemma 4, Appendix C, following Proposition 2 of Mishchenko et al., 2020). The right-hand side need not be globally bounded (Appendix C.2), and Eq. (7) controls it in expectation at the visited epoch starts.

4.1 CONVERGENCE OF MUON UNDER RANDOM RESHUFFLING

Section 4.2 sketches the proof of the main guarantee and Appendix E gives it in full.

Theorem 1 (Convergence of Muon under RR)**.** *Suppose Assumptions 1–2 hold with $L > 0$ and $N \ge 2$. Fix $K \ge 1$, a deterministic initialization $W_0^1 = \tilde{W}_0$, and $M_0^1 = 0$, and sample each epoch's permutation independently and uniformly. Let*

$$H := f(\tilde{W}_0) - \frac{1}{N}\sum_{j=1}^N f_j^* \ge 0, \qquad \lambda := \min\{1, (N/(rK))^{1/3}\}, \qquad \beta := 1 - \lambda/N,$$

with $r = \min\{m, p\}$. Run Algorithm 1 with momentum β and stepsize

$$\eta := \frac{1}{100N}\sqrt{\frac{H}{L}}\min\{\lambda, \sqrt{\lambda/K}\}. \tag{1}$$

The epoch-start and inner-iterate averages,

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\nabla f(W_0^k)\|_* \quad \text{and} \quad \frac{1}{KN}\sum_{k=1}^K\sum_{i=1}^N \mathbb{E}\|\nabla f(W_{i-1}^k)\|_*,$$

are both bounded by

$$200\sqrt{LH}\left(K^{-1/2} + r^{1/6}K^{-1/3}N^{-1/6}\right). \tag{2}$$

Consequently, for k_{out} and i_{out} uniform on $[K]$ and $[N]$, independent of each other and of the run, the same bound holds for $\mathbb{E}\|\nabla f(W_{\mathrm{out}})\|_*$ under either output rule $W_{\mathrm{out}} = W_0^{k_{\mathrm{out}}}$ or $W_{\mathrm{out}} = W_{i_{\mathrm{out}}-1}^{k_{\mathrm{out}}}$.

Discussion of Theorem 1. The bound has the same exponents of K and N as RR-SGD (Mishchenko et al., 2020, Theorem 4) and SGD with momentum under RR (Qiu et al., 2024), in unsquared gradient norms. The factor $r^{1/6}$ is at fixed operator–nuclear L. Against SGD with momentum at its own Frobenius constant L_F, it remains the smaller factor whenever $L \le r^{2/3}L_F$ (Appendix C.1). Proposition 4 extends the guarantee to the Jordan quintic, the Newton–Schulz polynomial of the public Muon implementation (Jordan et al., 2024), at any fixed depth in exact arithmetic and without a lower bound on the nonzero singular values of the buffer. Its proof lower bounds the descent inner product directly rather than the polar approximation error, at the cost of a constant stepsize rescaling and a weaker rank factor.

Momentum and stepsize. The momentum $\beta = 1 - \lambda/N$ is horizon-tuned (set from K), like the momentum in the IID Muon bound of Kim & Oh (2026), and makes the momentum memory span at least one pass (Section 4.2). A known positive deterministic $\widehat{H} \ge H$ may replace H in the schedule and bound, and for nonnegative component losses $\widehat{H} = f(\tilde{W}_0)$ suffices. Multiplying the stepsize by 4 or 100 retains the constant 200 (Corollary 2), and constants are not optimized.

4.2 PROOF SKETCH OF THEOREM 1

Roadmap. Steps 1–3 resolve the first obstacle (second contribution) by preserving full-pass cancellation through the carried EMA and polar descent. They give a stationarity bound that still depends on the trajectory's noise scale. Step 4 resolves the second obstacle (third contribution), since a joint potential bounds that scale using only initial constants. Step 5 balances the terms to obtain Eq. (2). Polar extremality, the weighted without-replacement identity and the component self-bound are standard.

For the sketch, assume $H > 0$. The momentum condition $N(1-\beta) = \lambda \le 1$ makes the momentum memory scale $(1-\beta)^{-1} = N/\lambda \ge N$ span at least one pass. Write $S_i^k = (1-\beta)M_i^k$ for the EMA buffer, $\bar{g}_k = \nabla f(W_0^k)$ for the epoch-start gradient, and $\rho_k = \|\bar{g}_k - S_0^k\|_*$ for the epoch-start buffer error. Expectation $\mathbb{E}_k$ is conditional on the history before epoch k.

Step 1, bound the descent error linearly. For a current buffer M, set $S = (1-\beta)M$ and $O = \mathrm{Polar}(S) = \mathrm{Polar}(M)$, so $\|O\|_{\mathrm{op}} \le 1$. Let g be the full gradient at the current iterate. Then polar extremality $\langle S, O\rangle_F = \|S\|_*$ and the duality $\langle A, B\rangle_F \le \|A\|_* \|B\|_{\mathrm{op}}$ yield $\langle g, O\rangle_F \ge \|g\|_* - 2\|g - S\|_*$, a standard normalized/LMO device (Cutkosky & Mehta, 2020; Bernstein & Newhouse, 2024; Pethick et al., 2025a; Lau et al., 2025). Thus (Lemma 6)

$$f(W_i^k) \le f(W_{i-1}^k) - \eta\|\nabla f(W_{i-1}^k)\|_* + 2\eta\|\nabla f(W_{i-1}^k) - S_i^k\|_* + \frac{L\eta^2}{2}. \tag{3}$$

Because the error enters Eq. (3) linearly through $\|\nabla f(W_{i-1}^k) - S_i^k\|_*$, the proof tracks the buffer in gradient space instead of linearizing the polar update by a shifted iterate. The first-moment bound on this term comes from a second-moment estimate (Step 2).

Step 2, freeze the gradients and isolate the reshuffling noise. At the start of epoch k, define the centered component gradients $\xi_j = \nabla f_j(W_0^k) - \bar{g}_k$ and their spread (heterogeneity) $\sigma_k^2 = N^{-1}\sum_j \|\xi_j\|_F^2$. These gradients are fixed at the epoch start only for the analysis, while the algorithm evaluates gradients at its moving iterates. With $\alpha_{i,s} = (1-\beta)\beta^{i-s}$, unrolling the EMA gives (Eq. (13))

$$\begin{aligned}\nabla f(W_{i-1}^k) - S_i^k = \underbrace{\nabla f(W_{i-1}^k) - \bar{g}_k}_{\text{full-gradient drift}} + \underbrace{\beta^i(\bar{g}_k - S_0^k)}_{\text{initial buffer error}} - \underbrace{\sum_{s=1}^{i}\alpha_{i,s}\xi_{\pi_s^k}}_{\text{frozen RR noise}} \\ + \underbrace{\sum_{s=1}^{i}\alpha_{i,s}\big(\nabla f_{\pi_s^k}(W_0^k) - \nabla f_{\pi_s^k}(W_{s-1}^k)\big)}_{\text{buffer drift}}.\end{aligned} \tag{4}$$

Only the frozen-noise term requires a sampling argument. The initial buffer error has norm $\beta^i\rho_k$, and each drift has nuclear norm at most $L\eta(i-1)$ (Lemma 7). Within-epoch samples need not be conditionally unbiased. Conditioning at the epoch start instead gives the weighted RR identity (Mishchenko et al., 2020; Qiu et al., 2024)

$$\mathbb{E}_k\Big\|\sum_{s=1}^{i}\alpha_{i,s}\xi_{\pi_s^k}\Big\|_F^2 = \frac{N\sum_s \alpha_{i,s}^2 - (\sum_s \alpha_{i,s})^2}{N-1}\,\sigma_k^2. \tag{5}$$

At the end of a full pass, define the mean weight $\bar{\alpha} = N^{-1}\sum_{s=1}^N \alpha_{N,s}$. Since $\sum_{s=1}^N \xi_{\pi_s^k} = 0$,

$$\sum_{s=1}^{N}\alpha_{N,s}\xi_{\pi_s^k} = \sum_{s=1}^{N}(\alpha_{N,s} - \bar{\alpha})\xi_{\pi_s^k}.$$

Only deviations from the mean EMA weight remain. Each deviation is at most $(1-\beta)(1-\beta^{N-1}) \le \lambda^2/N$, whereas for β fixed as N grows the deviations would be of order $1-\beta$. Let θ_i be the coefficient in Eq. (5). For the centered weights, this gives

$$\theta_N = \frac{N}{N-1}\sum_{s=1}^{N}(\alpha_{N,s} - \bar{\alpha})^2 \le \frac{2\lambda^4}{N}.$$

Corollary 1 gives the matching order $\theta_N = \Theta(\lambda^4/N)$ for $N \geq 2$ and $0 < \lambda \leq 1$, versus $\Theta(\lambda^2/N)$ for IID. For the same frozen vectors and weights, RR reduces the root-mean-square Frobenius noise by a factor of order λ. The corollary also gives $\sum_{i=1}^{N} \theta_i \leq 2\lambda^2$. Conversion to nuclear norm costs $\sqrt{r}$, and Step 3 carries the gain across epochs.

Step 3, carry the endpoint buffer error across epochs by a contraction. The same decomposition at $\nabla f(W_N^k) - S_N^k$ gives (Lemma 8)

$$\mathbb{E}_k \rho_{k+1} \leq (1 - \lambda/2)\rho_k + \sqrt{2r/N}\, \lambda^2 \sigma_k + 2L\eta N.$$

Write $\bar{\sigma}_K = K^{-1} \sum_k \mathbb{E}\sigma_k$. Summing the contraction costs a factor of order $1/\lambda$, leaving one power of λ from the endpoint noise's λ^2. Thus the noise contribution is of order $\lambda\sqrt{r/N}\,\bar{\sigma}_K$, versus $\sqrt{r/N}\,\bar{\sigma}_K$ under IID (Remark 1), alongside drift $L\eta N/\lambda$ and the decaying initial buffer error.

The within-epoch errors also satisfy (Lemma 8)

$$\sum_{i=1}^{N} \mathbb{E}_k \|\nabla f(W_{i-1}^k) - S_i^k\|_* \leq N\rho_k + \sqrt{2rN}\lambda\sigma_k + L\eta N^2.$$

The first term contributes $2\eta N \rho_k$ to epochwise descent through Eq. (3), which Step 4 cancels. Summing the descent and contraction inequalities gives, for fixed $\eta > 0$ and $0 < \lambda \leq 1$ (Proposition 3, using $f(\tilde{W}_0) - f^* \leq H$),

$$\frac{1}{KN} \sum_{k,i} \mathbb{E}\|\nabla f(W_{i-1}^k)\|_* \leq \frac{H}{\eta KN} + \frac{4\mathbb{E}\rho_1}{\lambda K} + 6\sqrt{\tfrac{2r}{N}}\lambda\bar{\sigma}_K + \frac{10L\eta N}{\lambda} + \frac{L\eta}{2}. \tag{6}$$

Step 4, jointly control the objective gap and buffer error. To close Eq. (6), we must bound the trajectory-dependent noise $\bar{\sigma}_K$. Smoothness and component lower bounds control σ_k^2 by the current objective gap (Lemma 4). Define the joint potential

$$\Psi_k := f(W_0^k) - \frac{1}{N}\sum_{j=1}^{N} f_j^* + \frac{4\eta N}{\lambda}\rho_k.$$

The weight $4\eta N/\lambda$ makes the buffer-error terms from descent and contraction cancel exactly, since

$$\underbrace{2\eta N \rho_k}_{\text{descent error}} + \frac{4\eta N}{\lambda} \underbrace{\left(-\frac{\lambda}{2}\rho_k\right)}_{\text{buffer contraction}} = 0.$$

The remaining noise feeds back into the same potential through $\sigma_k^2 \leq 2L\Psi_k$. Writing $u_k := \mathbb{E}\Psi_k$, taking expectations and applying Jensen's inequality give

$$u_{k+1} \leq u_k + 12\eta\sqrt{rLN}\lambda\sqrt{u_k} + \frac{11L\eta^2N^2}{\lambda}.$$

This recurrence needs no independently assumed noise bound (Lemma 9). Summing to the epoch with the largest expected potential yields a quadratic inequality in the square root of that maximum. Lemma 10 solves it under the schedule in Eq. (1) and applies the self-bound to obtain

$$\max_{1\leq k\leq K+1} \mathbb{E}\Psi_k \leq 2H, \qquad \max_{1\leq k\leq K} \mathbb{E}\sigma_k^2 \leq 4LH. \tag{7}$$

Thus $\bar{\sigma}_K \leq 2\sqrt{LH}$ replaces the trajectory-dependent quantity in Eq. (6) by initial constants, without a uniform variance assumption.

Step 5, balance the remaining noise against drift and buffer memory. Longer momentum memory reduces frozen noise but increases the cost of tracking a moving gradient and carrying past buffer error. For $K\lambda \geq 1$, the schedule bounds every term of Eq. (6) by a universal multiple of $\sqrt{LH}(K\lambda)^{-1/2}$, and $\lambda = (N/(rK))^{1/3}$ balances $\lambda\sqrt{r/N}$ against $(K\lambda)^{-1/2}$ unless capped at one, when the former is no larger. The factor $r^{1/6}$ thus comes from the Frobenius-to-nuclear conversion of the noise, since polar descent is already in nuclear norm. Appendix E handles $K\lambda < 1$.

4.3 QUERY BUDGETS UNDER RESHUFFLING AND IID SAMPLING

With N, r and L fixed, and with positive initial gaps and a positive IID variance bound as defined below, the sufficient budget for an ϵ-stationary output is $\mathcal{O}(\epsilon^{-3})$ under RR and $\mathcal{O}(\epsilon^{-4})$ under the existing IID Muon bound. The RR result uses component lower bounds, whereas the IID comparison uses an objective lower bound and a uniform variance bound. On a common objective satisfying both sets of assumptions, we compare the bounds at the same number T of component-gradient queries and the same average expected nuclear gradient norm over pre-update iterates. Each row below uses the trajectory of its indicated sampler. For RR, $T = KN$ with pre-update iterates indexed as $W_{(k-1)N+i-1} := W_{i-1}^k$, and Theorem 1 gives the first bound below. For IID, let $D = f(\tilde{W}_0) - f^*$ be the initial gap and let σ^2 be a uniform Frobenius variance bound, a constant unlike the epoch-dependent spread σ_k of Step 2. Proposition 5 proves the second bound with the leading-order rate of Theorem 4 of Kim & Oh (2026).

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(W_{t-1})\|_* &\le 200\sqrt{LH}\left(\sqrt{\frac{N}{T}} + \frac{(rN)^{1/6}}{T^{1/3}}\right) && \text{under RR,} \\ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(W_{t-1})\|_* &\le 9\left(\sqrt{\frac{LD}{T}} + \frac{\sigma\sqrt{r}}{\sqrt{T}} + \frac{(r\sigma^2 LD)^{1/4}}{T^{1/4}}\right) && \text{under IID.} \end{aligned} \tag{8}$$

Solving each bound for an ϵ-stationary output gives the sufficient budgets (Appendix F.2)

$$\begin{aligned} T_{\mathrm{RR}}(\epsilon) &= \mathcal{O}\big(N + NLH\epsilon^{-2} + \sqrt{rN}(LH)^{3/2}\epsilon^{-3}\big), \\ T_{\mathrm{IID}}(\epsilon) &= \mathcal{O}\big(1 + (LD + r\sigma^2)\epsilon^{-2} + r\sigma^2 LD\epsilon^{-4}\big). \end{aligned}$$

These sufficient upper bounds compare guarantees, not an RR–IID lower-bound separation. Because the RR budget grows with N, it is the smaller of the two only below an accuracy threshold of order $\sqrt{r/N}$ (Appendix F.2), so a sharper accuracy exponent need not give the smaller bound at a fixed finite budget. Section 5 compares the samplers at matched budgets in language-model training.

5 LANGUAGE-MODEL EXPERIMENTS

On a fixed language-model training pool, with matched optimizer settings and final update budget, Muon under RR has lower final training, validation and test cross-entropy (CE) than under IID in every seed pair. It also has lower training-loss area under the curve over the recorded checkpoints.

Data and sampling. We train a 124M-parameter decoder from scratch on 32,768 packed FineWeb-Edu sequences (Penedo et al., 2024) with 512 next-token targets each, with validation and test splits of 4,096 sequences made at document level before packing. A component is a fixed minibatch of 32 sequences, giving 1,024 components, which RR permutes once per pass and IID samples uniformly with replacement. Both start from the same weights within each seed pair and use 8,192 updates (134.2M prediction targets).

Optimizer and evaluation. Training-loss area under the curve (AUC), computed by the trapezoidal rule over the nine common evaluation checkpoints and divided by the update budget of 8,192, is the predeclared primary metric. The learning rate (stepsize η) 0.005 and momentum 0.95 minimize, on two pilot seeds, this AUC averaged with equal weight over IID and over the RR conditions of the pilot, and the same pair minimizes each sampler's own pilot AUC. Across all nine pilot settings, RR has lower training AUC than IID in both pilot seeds (Table 3). Five disjoint confirmation seeds then run with five quintic Newton–Schulz iterations, fixed momentum and AdamW on the non-Muon parameters, as in practice, apart from the Nesterov and shape-scaling options, which are disabled (Appendices G.1–G.2).

Table 2: **RR versus IID at the matched final budget.** Mean $\pm$ sample standard deviation (SD) over five paired seeds, lower is better. AUC includes initialization, and test uses each final model. Table 4 retains all confirmation conditions.

Sampling	Training AUC	Final train CE	Final validation CE	Final test CE
RR	4.0333 ± 0.0016	2.5215 ± 0.0058	4.6985 ± 0.0054	4.6772 ± 0.0059
IID	4.0716 ± 0.0059	2.6455 ± 0.0181	5.1587 ± 0.0471	5.1352 ± 0.0469

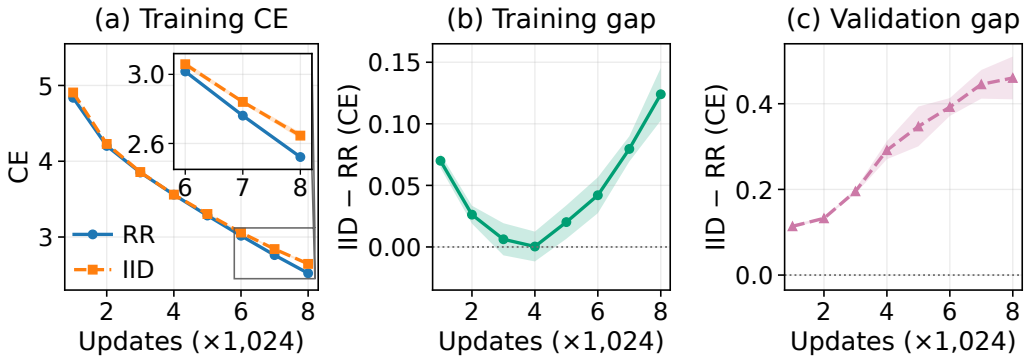


Figure 1: **Matched RR and IID for the 124M language model.** Five-seed means with sample SD shading. (a) Full-pool training CE, with an inset for updates 6,144–8,192. SD bands are narrower than the markers. (b,c) Paired training and validation gaps, IID minus RR, with SD across pairs. Positive values favor RR. Panels omit the shared initial CE (≈ 10.99). Figure 3 restores the full range, and Figure 4 shows the same data by seed.

Training optimization at matched budgets. RR has lower training AUC and final training CE in all five seed pairs (Figure 1, Table 2). The paired AUC difference (IID minus RR) is 0.038, with an approximate 97.5% paired-percentile bootstrap interval $[0.034, 0.043]$, Bonferroni-adjusted for the two predeclared primary contrasts. The final training CE difference is 0.124. At each noninitial RR checkpoint, every component has been used once during the preceding 1,024 updates. IID has no such coverage constraint. The mean training gap is nearly zero at update 4,096 and grows afterwards (Figure 1(b)). Appendix G.3 gives the complete metric and timing results.

Held-out performance at the final budget. On the held-out splits, which have no visit structure, RR has lower final validation and test CE in all five seed pairs, with a mean test CE difference (IID minus RR) of 0.458. Appendix G.4 reports validation-checkpoint diagnostics for interpreting these final-budget comparisons.

Coverage and repeated-list controls. IID changes both sampling dependence and visit counts. The controls in Appendix G.2 shuffle a c-times replicated component list at each cycle. Each cycle has $1{,}024c$ updates, and every component is visited eight times over the full budget. Both controls have lower mean AUC and final losses than IID (Table 4). The other predeclared primary contrast, RR $c = 8$ minus RR $c = 1$, has AUC difference 0.029 with 97.5% interval $[0.025, 0.032]$ (Appendix G.3). Thus matching total visit counts does not eliminate the observed difference between reshuffling schedules. This comparison does not separate causal contributions of coverage and temporal dependence.

Momentum 0.9995 satisfies the momentum condition $N(1-\beta) \le 1$ of Theorem 1 with $N(1-\beta) = 0.512$; the selected momentum 0.95 gives 51.2 and lies outside this condition (Appendices D.2 and I). The separate fixed-N synthetic check uses the theorem's momentum. At 100 times the stepsize in Eq. (1), covered by Corollary 2, RR has about 25% lower mean endpoint buffer error and 5% lower average nuclear gradient norm than matched IID in both families (Appendix H).

6 Conclusion

We establish, to our knowledge, the first general smooth nonconvex finite-sum stationarity guarantee for Muon under RR. Its sufficient query budget has the ϵ^{-3} accuracy dependence of RR-SGD, against ϵ^{-4} for the existing IID Muon bound with positive gradient variance and initial gap. The analysis preserves endpoint cancellation through carried momentum, then couples the objective gap and the buffer error to control trajectory-dependent noise without a uniform variance assumption. In our language-model study, RR attains lower training AUC and final losses than IID at matched budgets. Appendix I discusses the remaining theory–practice gap.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, literature checking, mathematical derivations and proof auditing, and development and auditing of the experiment and analysis code. The authors are responsible for the mathematical claims, reported results, and final submission.

REFERENCES

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33: 17526–17535, 2020.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: 16th Annual Workshop on Optimization for Machine Learning (NeurIPS 2024 Workshop)*, 2024. arXiv:2409.20325.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *International Conference on Machine Learning*, pp. 3855–3912. PMLR, 2023.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of Muon and beyond. *arXiv preprint arXiv:2509.15816v5*, 2025.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. In *OPT 2025: 17th Annual Workshop on Optimization for Machine Learning (NeurIPS 2025 Workshop)*, 2025. arXiv:2506.15054.

Hanseul Cho and Chulhee Yun. SGDA with shuffling: Faster convergence for nonconvex-PŁ minimax optimization. In *International Conference on Learning Representations*, 2023.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pp. 2260–2268. PMLR, 2020.

Aniket Das, Bernhard Schölkopf, and Michael Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. *Advances in Neural Information Processing Systems*, 35:6749–6762, 2022.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On MUON optimization: From non-convergence to an error analysis with Polar Express and the Newton–Schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021.

Jeff Z. HaoChen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, pp. 2624–2633. PMLR, 2019.

Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex Hölder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Anastasia Koloskova, Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *International Conference on Machine Learning*, pp. 25058–25086. PMLR, 2024.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. In *International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=OT9cxeWbEO`. Extended version: arXiv:2602.11557v2.

Mingyi Li and Taira Tsuchiya. Muon with finite Newton–Schulz: The smoothing benefit in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2608.26288v1*, 2026.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Zijian Liu. Random reshuffling dominates stochastic gradient descent. In *Conference on Learning Theory*, pp. 4859–4882. PMLR, 2026.

Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1347–1357. PMLR, 2023.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pp. 4703–4711. PMLR, 2019.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of Muon optimizer for nonconvex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning*, pp. 49069–49104. PMLR, 2025a.

Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Antonio Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping & non-Euclidean (L_0, L_1)-smoothness. *Advances in Neural Information Processing Systems*, 38:14417–14441, 2025b.

Xun Qian and Peter Richtárik. Convergence analysis of Muon-type methods with inexact LMO in the degenerate case. *arXiv preprint arXiv:2606.21581*, 2026. URL https://arxiv.org/abs/2606.21581.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtárik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Zhen Qin, Zhishuai Liu, and Pan Xu. Convergence of sign-based random reshuffling algorithms for nonconvex optimization. *arXiv preprint arXiv:2310.15976v4*, 2023.

Junwen Qiu, Bohao Ma, and Andre Milzarek. Random reshuffling with momentum: Complexity bounds and last-iterate convergence. *arXiv preprint arXiv:2404.18452v3*, 2024. Revised June 27, 2026.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, 2020.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From Muon to Gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IelAHU5MVz.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory*, pp. 3250–3284. PMLR, 2020.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv preprint arXiv:2507.01598v5*, 2025.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *International Conference on Artificial Intelligence and Statistics*, pp. 5167–5175. PMLR, 2026a.

Egor Shulgin, Mohamed Awad, Peter Richtárik, and Eduard Gorbunov. General analysis of LMO-based optimizers: Beyond bounded variance. In *International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=s0Q8nVsrMu.

Trang H. Tran, Lam M. Nguyen, and Quoc Tran-Dinh. SMG: A shuffling gradient-based method with momentum. In *International Conference on Machine Learning*, pp. 10379–10389. PMLR, 2021.

Trang H. Tran, Katya Scheinberg, and Lam M. Nguyen. Nesterov accelerated shuffling gradient method for convex optimization. In *International Conference on Machine Learning*, pp. 21703–21732. PMLR, 2022.

Trang H. Tran, Quoc Tran-Dinh, and Lam M. Nguyen. Shuffling momentum gradient algorithm for convex optimization. *Vietnam Journal of Mathematics*, 53:773–801, 2025. doi: 10.1007/s10013-024-00699-7.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in Neural Information Processing Systems*, 32:3727–3740, 2019b.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. In *International Conference on Learning Representations*, 2025.

APPENDIX OVERVIEW

The dependency table follows the main proof sketch. Frozen-noise cancellation and buffer tracking establish Contribution 2, and joint stability then controls the remaining noise for Contribution 3. All ingredients needed for Theorem 1 are stated below, with citations identifying standard results. Here $\tau > 0$ is the stepsize prefactor introduced in Appendix A.

Part of the argument	Statements used
Geometry and self-bounding	Lemma 1, Propositions 1–2, Lemmas 2–4
Frozen RR cancellation	Lemma 5 and Corollary 1, applied conditionally on the epoch-start history
Trajectory and buffer error	Lemmas 6, 7, and 8, and the explicit fixed-stepsize bound of Proposition 3
Stability and stationarity	Lemmas 9 and 10, and Corollary 2, including both ranges of $K\lambda$
Theorem and larger prefactors	Specialize the general-prefactor results to $\tau = 1/100, 1/25, 1$ in the proof of Theorem 1

Appendix E.3 gives the separate fixed-depth Jordan extension, and Appendix F gives exact literature and query-budget comparisons. Appendices G and H contain the language-model protocol and the fixed-N tracking check, and Appendix I states the scope limits. The optional finite-horizon transfer for Taylor-type Newton–Schulz iterations is retained as a separate supplementary note, not used by the main theorem.

A NOTATION AND SETUP

Throughout the theory, components are fixed and Polar denotes the canonical partial polar factor of Section 3. Here I_d is the $d \times d$ identity and diag forms a diagonal matrix. The rescaled buffer $S_i^k = (1-\beta)M_i^k$ is an analysis device only, since positive scaling leaves the polar direction unchanged. For deterministic initialization, let $\mathcal{F}_{k-1} := \sigma(\pi^1, \dots, \pi^{k-1})$, with $\mathcal{F}_0$ the trivial sigma-field, and set $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$. The epoch-start state and its frozen component gradients are $\mathcal{F}_{k-1}$-measurable, while π^k is independent of this history and uniform. Thus the conditional calculations freeze the epoch-start history before any element of the new permutation is revealed. At an epoch boundary $(W_0^{k+1}, M_0^{k+1}) = (W_N^k, M_N^k)$, and no buffer reset occurs. This also defines $(W_0^{K+1}, M_0^{K+1}) = (W_N^K, M_N^K)$ and $S_0^{K+1} = (1-\beta)M_0^{K+1}$ as terminal notation for telescoping, without running an additional epoch. We use τ for the proof stepsize prefactor, and the main theorem uses $\tau = 1/100$. As in Section 4.2, the epoch-start buffer error, the heterogeneity of the component gradients at an epoch start, and its average over epochs are

$$\rho_k := \|\nabla f(W_0^k) - S_0^k\|_*, \qquad \sigma_k^2 := \frac{1}{N}\sum_{j=1}^{N} \|\nabla f_j(W_0^k) - \nabla f(W_0^k)\|_F^2, \qquad \bar{\sigma}_K := \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\sigma_k.$$

Why the Euclidean shifted iterate does not apply to Muon. Suppress the epoch index and let G_i be the current component gradient. For SGD with momentum, $M_i = \beta M_{i-1} + G_i$ and $W_i = W_{i-1} - \eta M_i$, the shifted iterate $Y_i = W_i + \frac{\beta}{1-\beta}(W_i - W_{i-1})$ satisfies $Y_i = Y_{i-1} - \frac{\eta}{1-\beta}G_i$ (Qiu et al., 2024). The polar map is scale invariant but not additive, since $\mathrm{Polar}(\mathrm{diag}(1,0)) + \mathrm{Polar}(I_2) = \mathrm{diag}(2,1)$, whereas $\mathrm{Polar}(\mathrm{diag}(1,0) + I_2) = I_2$. Hence $-\eta\sum_i \mathrm{Polar}(S_i^k)$ is not affine in the gradients.

B BASIC FACTS ABOUT MATRIX NORMS

Definition 2 (Matrix norms). *Write $\sigma_j(X)$ for the jth singular value of X, with $r = \min\{m, p\}$ and zero singular values included. The nuclear, Frobenius, and operator norms are $\|X\|_* = \sum_{j=1}^r \sigma_j(X)$, $\|X\|_F = (\sum_{j=1}^r \sigma_j(X)^2)^{1/2}$, and $\|X\|_{\mathrm{op}} = \max_{1\le j\le r} \sigma_j(X)$, respectively.*

Lemma 1 (Hölder's inequality). *For any $A, B \in \mathbb{R}^{m\times p}$,*

$$|\langle A, B\rangle_F| \le \|A\|_* \|B\|_{\mathrm{op}}.$$

Proof. This is the Schatten trace inequality for the pair $(1, \infty)$ (Bhatia, 1997, Ch. IV). Equivalently, apply $|u^\top Bv| \le \|B\|_{\mathrm{op}}$ to each pair of unit singular vectors in an SVD of A, multiply by the corresponding singular value, and sum. □

Proposition 1 (Polar extremality and norm comparisons). *For $X \in \mathbb{R}^{m\times p}$, set $r = \min\{m, p\}$ and $P = \mathrm{Polar}(X)$. Then*

$$\|X\|_F = \sqrt{\mathrm{tr}(X^\top X)} = \Big(\sum_{i,j} X_{ij}^2\Big)^{1/2}, \qquad \langle X, P\rangle_F = \|X\|_*,$$

$$\|X\|_{\mathrm{op}} \le \|X\|_F \le \|X\|_* \le \sqrt{r}\,\|X\|_F, \qquad \|X\|_F \le \sqrt{r}\,\|X\|_{\mathrm{op}},$$

$$\|P\|_{\mathrm{op}} \le 1, \qquad \mathrm{Polar}(cX) = P \quad (c > 0).$$

In particular, $\|X\|_* \le r\|X\|_{\mathrm{op}}$, and

$$\max_{\|O\|_{\mathrm{op}}\le 1} \langle X, O\rangle_F = \|X\|_*.$$

Proof. The norm identities and comparisons follow from the singular values and scalar Cauchy–Schwarz (Bhatia, 1997, Ch. IV). For a compact SVD $X = U_s\Sigma_s V_s^\top$ over its $s > 0$ positive singular values, $P = U_s V_s^\top$ has operator norm one and $\langle X, P\rangle_F = \mathrm{tr}(\Sigma_s) = \|X\|_*$. Joint rotations within repeated positive singular-value subspaces leave P unchanged, and zero singular-value subspaces do not enter its definition. Thus the partial polar is well defined for rank-deficient matrices, and for $X = 0$ we use $P = 0$. Positive scaling changes only Σ_s. Finally, Lemma 1 bounds the maximum by $\|X\|_*$, and the feasible choice $O = P$ attains it. □

C Smoothness and Component-Gradient Spread

C.1 Smoothness and gradient-gap bounds

Proposition 2 (Smoothness of the average and conversion of norms). *Under Assumption 1, the average $f = N^{-1}\sum_i f_i$ is continuously differentiable and L-smooth in the same operator–nuclear norm pair. Operator–nuclear L-smoothness of any differentiable function implies Frobenius L-smoothness. Conversely, Frobenius L_F-smoothness implies operator–nuclear rL_F-smoothness.*

Proof. Differentiation of the finite sum and the triangle inequality give

$$\|\nabla f(W) - \nabla f(V)\|_* \le \frac{1}{N}\sum_i \|\nabla f_i(W) - \nabla f_i(V)\|_* \le L\|W - V\|_{\mathrm{op}}.$$

For any differentiable h, Proposition 1 gives the implications

$$\|\nabla h(W) - \nabla h(V)\|_F \le \|\nabla h(W) - \nabla h(V)\|_* \le L\|W - V\|_{\mathrm{op}} \le L\|W - V\|_F,$$

$$\|\nabla h(W) - \nabla h(V)\|_* \le \sqrt{r}\|\nabla h(W) - \nabla h(V)\|_F \le \sqrt{r}L_F\|W - V\|_F \le rL_F\|W - V\|_{\mathrm{op}}.$$

The factor r cannot generally be removed, since for $h(W) = \frac{L_F}{2}\|W\|_F^2$ and a difference matrix with all r singular values equal to one, the ratio of the nuclear gradient difference to the operator parameter difference is rL_F. □

Interpreting the dimension factors. The powers of r in Table 1 are upper-bound coefficients at fixed operator–nuclear L, not an intrinsic separation. At fixed Frobenius smoothness L_F the two Frobenius-based rows keep $r^{1/2}$ and $r^{3/2}$, the IID Muon row becomes $r^{1/2}$ because $L = rL_F$ enters its noise term $(r\sigma^2 LD)^{1/4}$, and substituting $L = rL_F$ in Eq. (2) makes the coefficient of $K^{-1/3}N^{-1/6}$ equal to $200r^{2/3}\sqrt{L_F H}$. For any function with Frobenius constant L_F, Proposition 2 gives $L_F \le L \le rL_F$, and the RR Muon factor $r^{1/6}\sqrt{L}$ is below the factor $r^{1/2}\sqrt{L_F}$ of SGD with momentum whenever $L \le r^{2/3}L_F$, that is, unless L lies within a factor $r^{1/3}$ of its largest possible value rL_F. The first row's final $r^{1/2}$ norm conversion may be loose, and we do not establish a matching lower bound.

Lemma 2 (Descent under operator–nuclear smoothness). *Let $f : \mathbb{R}^{m\times p} \to \mathbb{R}$ be L-smooth with respect to $(\|\cdot\|_{\mathrm{op}}, \|\cdot\|_*)$. Then, for all $X, Y \in \mathbb{R}^{m\times p}$,*

$$f(Y) \le f(X) + \langle \nabla f(X), Y - X\rangle_F + \frac{L}{2}\|Y - X\|_{\mathrm{op}}^2.$$

Proof. This is the standard descent lemma for smoothness in a norm pair (see, e.g., Bernstein & Newhouse, 2024; Pethick et al., 2025a), and the integral argument is included here. Put $\Delta = Y - X$. The fundamental theorem of calculus and the chain rule yield

$$\begin{aligned} f(Y) - f(X) - \langle \nabla f(X), \Delta\rangle_F &= \int_0^1 \langle \nabla f(X + t\Delta) - \nabla f(X), \Delta\rangle_F \, dt \\ &\le \int_0^1 \|\nabla f(X + t\Delta) - \nabla f(X)\|_* \|\Delta\|_{\mathrm{op}} \, dt \\ &\le \int_0^1 Lt\|\Delta\|_{\mathrm{op}}^2 \, dt = \frac{L}{2}\|\Delta\|_{\mathrm{op}}^2, \end{aligned}$$

where the first inequality is Lemma 1. This proves the claim. □

Lemma 3 (Nuclear gradient-gap). *Let $h : \mathbb{R}^{m\times p} \to \mathbb{R}$ be L-smooth in the operator–nuclear norm pair with $L > 0$, and let $\ell \in \mathbb{R}$ satisfy $h(Y) \ge \ell$ for every Y. Then, for every X,*

$$\|\nabla h(X)\|_*^2 \le 2L(h(X) - \ell).$$

In particular, the inequality holds for $h = f$, $\ell = f^*$ and for each $h = f_i$, $\ell = f_i^*$. The lower bound need not be attained or equal the infimum.

Proof. Set $g = \nabla h(X)$, $P = \mathrm{Polar}(g)$, and $t = \|g\|_*/L$. Proposition 1 gives $\langle g, P\rangle_F = \|g\|_*$ and $\|P\|_{\mathrm{op}} \le 1$. Applying Lemma 2 to h at $X - tP$ gives

$$\ell \le h(X - tP) \le h(X) - t\|g\|_* + \frac{Lt^2}{2}\|P\|_{\mathrm{op}}^2 \le h(X) - \frac{\|g\|_*^2}{2L}.$$

Rearrangement proves the result, including $g = 0$ since then $t = 0$. The preceding proposition verifies the required smoothness for the average f. □

Lemma 4 (Self-bounding component heterogeneity). *Under Assumptions 1–2, for any $W \in \mathbb{R}^{m\times p}$,*

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(W) - \nabla f(W)\|_F^2 \le 2L\left(f(W) - \frac{1}{N}\sum_{i=1}^N f_i^*\right).$$

Attribution. This is the operator–nuclear specialization of the smooth, lower-bounded component argument in Mishchenko et al. (2020, Proposition 2), proved below with the norms and constants used here. Its state-dependent variance control replaces a uniform bounded-variance assumption. Related expected-smoothness and growth conditions are studied by Schmidt & Le Roux (2013); Vaswani et al. (2019a;b); Gower et al. (2021); Khaled & Richtárik (2023).

Proof. At the fixed point W, write $g_i = \nabla f_i(W)$ and $\bar{g} = N^{-1}\sum_i g_i = \nabla f(W)$. Expanding the Frobenius squares gives the exact variance decomposition

$$\begin{aligned} \frac{1}{N}\sum_i \|g_i - \bar{g}\|_F^2 &= \frac{1}{N}\sum_i \big(\|g_i\|_F^2 - 2\langle g_i, \bar{g}\rangle_F + \|\bar{g}\|_F^2\big) \\ &= \frac{1}{N}\sum_i \|g_i\|_F^2 - \|\bar{g}\|_F^2 \le \frac{1}{N}\sum_i \|g_i\|_*^2 \\ &\le \frac{2L}{N}\sum_i (f_i(W) - f_i^*) = 2L\left(f(W) - \frac{1}{N}\sum_i f_i^*\right), \end{aligned}$$

where the final inequality applies Lemma 3 separately to each component. No component minimizer or global bound on the component gradients is required. □

C.2 SMOOTH NONCONVEX EXAMPLES AND UNBOUNDED GRADIENT SPREAD

Let $J = 16$ and $W \in \mathbb{R}^{16\times 8}$. For each component, let $A_j \in \mathbb{R}^{16\times 8}$ be its center and a_j its curvature coefficient. The homogeneous case has $a_j = 1$, and the heterogeneous case uses 16 equally spaced values in $[0.5, 1.5]$. Define

$$\phi_j(W) = \frac{a_j}{2}\|W - A_j\|_F^2 + 2\sum_{u,v}\big(1 - \cos(W_{uv} - A_{j,uv})\big), \qquad f(W) = \frac{1}{J}\sum_{j=1}^{J}\phi_j(W). \tag{9}$$

Both families share the centers A_j. Appendix H gives the generation and initialization details.

The case names refer only to the quadratic coefficients a_j, not to identical component gradients or identical full component Hessians. Because both cases have mean coefficient one and share all centers, their averaged Hessians are identical. Their full-gradient difference is the constant matrix $-J^{-1}\sum_j (a_j - 1)A_j$, where a_j denotes the heterogeneous coefficient. This pair isolates curvature heterogeneity in the components rather than comparing unrelated full-objective geometries.

Global constants. Each component has the exact minimum $\phi_j^* = 0$ at A_j, since both terms are nonnegative. Thus $H = f(W_0)$ is known before training. Its gradient is $a_j(W - A_j) + 2\sin(W - A_j)$, with the sine applied entrywise, and its Frobenius Lipschitz constant is at most $a_j + 2$. Proposition 2 therefore gives valid operator–nuclear constants $L = 8(\max_j a_j + 2)$, namely 24 and 28. These constants are global, not estimates along the run. The full-objective Hessian certifies nonconvexity. Let $\bar{a} = J^{-1}\sum_j a_j = 1$ and let i denote the imaginary unit. The Hessian is diagonal in entrywise coordinates, and the smallest possible diagonal value at entry (u, v) is $\bar{a} - 2|J^{-1}\sum_j \exp(-\mathrm{i}A_{j,uv})|$. For the fixed centers this value is negative, so both averages are nonconvex, and the supplementary regeneration check records the amplitudes.

For homogeneous curvature, centered component gradients are bounded because the linear W terms cancel and the remaining sinusoidal terms are bounded. For heterogeneous curvature, write $B_j(W)$ for the bounded remainder from the fixed centers and sinusoidal terms. Then

$$\nabla\phi_j(W) - \nabla f(W) = (a_j - \bar{a})W + B_j(W).$$

Let $\mathrm{Var}(a_j) = J^{-1}\sum_j (a_j - \bar{a})^2$. Along any unit-Frobenius direction Δ,

$$\frac{1}{J}\sum_j \|\nabla\phi_j(t\Delta) - \nabla f(t\Delta)\|_F^2 = \mathrm{Var}(a_j)t^2 + O(|t| + 1) \longrightarrow \infty.$$

The heterogeneous family therefore satisfies component smoothness and lower bounds but admits no uniform variance constant.

D RESHUFFLING NOISE AND ENDPOINT CANCELLATION

D.1 WEIGHTED NOISE AND THE FULL-PASS GAIN

Lemma 5 (Weighted without-replacement second moment)**.** *Let $N \geq 2$, let $g_1, \ldots, g_N \in \mathbb{R}^{m\times p}$ satisfy $N^{-1}\sum_{j=1}^{N} g_j = 0$, and let π be a uniform random permutation of $[N]$. For any deterministic real weights $\alpha_1, \ldots, \alpha_i$, $1 \leq i \leq N$,*

$$\mathbb{E}_\pi \left\| \sum_{s=1}^{i} \alpha_s g_{\pi_s} \right\|_F^2 = \frac{N\sum_{s=1}^{i}\alpha_s^2 - (\sum_{s=1}^{i}\alpha_s)^2}{N-1} \cdot \frac{1}{N}\sum_{j=1}^{N}\|g_j\|_F^2.$$

The same identity holds conditionally on a sigma-field $\mathcal{F}$ when the vectors and weights are $\mathcal{F}$-measurable and the permutation is independent of $\mathcal{F}$, with finite required second moments. For EMA weights $\alpha_s = (1-\beta)\beta^{i-s}$, $\beta \in [0, 1)$,

$$0 \leq \frac{N\sum_{s=1}^{i}\alpha_s^2 - (\sum_{s=1}^{i}\alpha_s)^2}{N-1} \leq \frac{2(1-\beta)}{1+\beta}.$$

This is a weighted extension of the sampling-without-replacement identity in Mishchenko et al. (2020, Lemma 1). We prove both forms here without claiming either as a new contribution.

Proof. Write $V = N^{-1}\sum_j \|g_j\|_F^2$. Uniformity gives $\mathbb{E}_\pi\|g_{\pi_s}\|_F^2 = V$. For $s \neq t$, each ordered pair of distinct indices has probability $1/[N(N-1)]$. Therefore

$$\mathbb{E}_\pi\langle g_{\pi_s}, g_{\pi_t}\rangle_F = \frac{\|\sum_j g_j\|_F^2 - \sum_j \|g_j\|_F^2}{N(N-1)} = -\frac{V}{N-1}.$$

Expanding the squared norm and using $2\sum_{s<t}\alpha_s\alpha_t = (\sum_s \alpha_s)^2 - \sum_s \alpha_s^2$ yields

$$\mathbb{E}_\pi\left\|\sum_{s=1}^{i}\alpha_s g_{\pi_s}\right\|_F^2 = V\left(\sum_s \alpha_s^2 - \frac{2}{N-1}\sum_{s<t}\alpha_s\alpha_t\right) = V\,\frac{N\sum_s \alpha_s^2 - (\sum_s \alpha_s)^2}{N-1}.$$

Conditionally on $\mathcal{F}$, the vectors and weights are fixed while the permutation remains uniform, so the same finite covariance expansion proves the conditional claim. In our application $\mathcal{F}$ is the complete history before the new epoch, so the gradients are frozen at that epoch's starting point. No conditional-unbiasedness assertion is needed after a prefix of the current permutation has been exposed.

The coefficient is nonnegative because $(\sum_{s=1}^{i}\alpha_s)^2 \le i\sum_{s=1}^{i}\alpha_s^2 \le N\sum_{s=1}^{i}\alpha_s^2$. For the EMA weights, geometric sums give

$$\Gamma_i := \sum_{s=1}^{i}\alpha_s = 1-\beta^i, \qquad Q_i := \sum_{s=1}^{i}\alpha_s^2 = \frac{1-\beta}{1+\beta}(1-\beta^{2i}).$$

Thus $(NQ_i - \Gamma_i^2)/(N-1) \le NQ_i/(N-1) \le 2(1-\beta)/(1+\beta)$, using $N/(N-1) \le 2$. These identities include $\beta = 0$, with the usual empty-power convention $\beta^0 = 1$. □

Corollary 1 (Slow-EMA consequences of Lemma 5)**.** *Let $N \ge 2$, $0 < \lambda \le 1$, $\beta = 1 - \lambda/N$, and $\alpha_{i,s} = (1-\beta)\beta^{i-s}$ for $1 \le s \le i \le N$. Define*

$$\theta_i := \frac{N\sum_{s=1}^{i}\alpha_{i,s}^2 - (\sum_{s=1}^{i}\alpha_{i,s})^2}{N-1}.$$

Then, uniformly over N and λ,

$$\sum_{i=1}^{N}\theta_i \le 2\lambda^2, \qquad \frac{e^{-2}}{12}\frac{\lambda^4}{N} \le \theta_N \le \frac{1}{8}\frac{\lambda^4}{N} \le 2\frac{\lambda^4}{N}.$$

Consequently, for centered $g_1, \dots, g_N$ and $\sigma^2 = N^{-1}\sum_j \|g_j\|_F^2$,

$$\sum_{i=1}^{N}\mathbb{E}_\pi\left\|(1-\beta)\sum_{s=1}^{i}\beta^{i-s}g_{\pi_s}\right\|_F^2 \le 2\lambda^2\sigma^2, \qquad \mathbb{E}_\pi\left\|(1-\beta)\sum_{s=1}^{N}\beta^{N-s}g_{\pi_s}\right\|_F^2 \le 2\frac{\lambda^4}{N}\sigma^2.$$

Proof. Write $\delta := 1-\beta = \lambda/N$. Since every EMA weight is at most δ, we have

$$\theta_i \le \frac{N}{N-1}\sum_{s=1}^{i}\alpha_{i,s}^2 \le 2i\delta^2.$$

Summing and using $N+1 \le 2N$ gives $\sum_i \theta_i \le \delta^2 N(N+1) = \frac{\lambda^2(N+1)}{N} \le 2\lambda^2$.

At the endpoint, reindex the weights as $w_t := \delta\beta^t$ for $0 \le t \le N-1$. Expanding the right side of the identity $N\sum_t w_t^2 - (\sum_t w_t)^2 = \sum_{0\le s<t\le N-1}(w_s - w_t)^2$ gives

$$\theta_N = \frac{1}{N-1}\sum_{0\le s<t\le N-1}(w_s - w_t)^2.$$

For $s < t$, the difference $w_s - w_t$ satisfies

$$w_s - w_t = \delta^2\sum_{j=s}^{t-1}\beta^j, \qquad \delta^2(t-s)\beta^{N-2} \le w_s - w_t \le \delta^2(t-s).$$

All exponents in this sum lie between 0 and $N-2$, so these bounds also cover $N=2$. Moreover, the sum of squared index differences evaluates to

$$\sum_{0\le s<t\le N-1}(t-s)^2 = N\sum_{t=0}^{N-1}t^2 - \left(\sum_{t=0}^{N-1}t\right)^2 = \frac{N^2(N^2-1)}{12}.$$

Substituting these estimates yields

$$\beta^{2N-4}\frac{(N+1)\lambda^4}{12N^2} \le \theta_N \le \frac{(N+1)\lambda^4}{12N^2}.$$

For the lower bound, $\beta \ge 1-1/N$ and $\log(1-x) = -\int_0^x (1-t)^{-1}\,dt \ge -x/(1-x)$ give

$$(N-2)\log\beta \ge (N-2)\log(1-1/N) \ge -\frac{N-2}{N-1} \ge -1.$$

Hence $\beta^{2N-4} \ge e^{-2}$. Finally, the relation $1 \le (N+1)/N \le 3/2$ holds for $N \ge 2$, proving the displayed universal constants. In particular, $\theta_2 = \lambda^4/16$ exactly. The second-moment consequences follow by multiplying the coefficients by σ^2 in Lemma 5. The weaker endpoint constant 2 is retained in the subsequent tracking proof. □

Same frozen statistic under IID sampling. Keep the same centered vectors and weights, but draw $j_1,\dots,j_N$ independently and uniformly from $[N]$. For $s \ne t$, independence and centering give $\mathbb{E}\langle g_{j_s}, g_{j_t}\rangle_F = \langle \mathbb{E}g_{j_s}, \mathbb{E}g_{j_t}\rangle_F = 0$. Therefore

$$\mathbb{E}\left\|\sum_{s=1}^{i}\alpha_{i,s}g_{j_s}\right\|_F^2 = Q_i\sigma^2, \qquad Q_i = \frac{1-\beta}{1+\beta}(1-\beta^{2i}).$$

In particular,

$$\mathbb{E}\left\|\sum_{s=1}^{N}\alpha_{N,s}g_{j_s}\right\|_F^2 = \frac{1-\beta}{1+\beta}(1-\beta^{2N})\sigma^2. \tag{10}$$

Since $Q_i \le i(\lambda/N)^2$, $\sum_{i=1}^N Q_i \le \lambda^2(N+1)/(2N) \le \lambda^2$. At the endpoint,

$$\frac{1-e^{-2}}{2}\,\frac{\lambda^2}{N} \le Q_N \le \frac{\lambda^2}{N}.$$

The upper bound uses N weights each at most λ/N. For the lower bound, $(1-\beta)/(1+\beta) \ge \lambda/(2N)$, and $1-x \le e^{-x}$ gives $\beta^{2N} \le e^{-2\lambda}$, so $1-\beta^{2N} \ge 1-e^{-2\lambda} \ge (1-e^{-2})\lambda$, where the last step is concavity of $\lambda \mapsto 1-e^{-2\lambda}$ on $[0,1]$. Thus Corollary 1 and Eq. (10) give endpoint coefficients $\Theta(\lambda^4/N)$ for RR and $\Theta(\lambda^2/N)$ for IID, with constants uniform over $N \ge 2$ and $0 < \lambda \le 1$. This comparison concerns the same frozen vectors and weights, and if $\sigma = 0$, both second moments are zero. Figure 2 plots the endpoint weights and both coefficients at $N = 1{,}024$, the number of components in the language-model study.

Remark 1 (What this endpoint-tracking closure gives under IID)**.** *The first-moment closure below uses the prefix coefficients, whose sum has order at most λ^2 for either sampler, and the endpoint coefficient. Writing $\Gamma_N := 1-\beta^N$ for the total EMA weight of a pass, Bernoulli's inequality and $\beta^N \le e^{-\lambda} \le 1-\lambda/2$ give $\lambda/2 \le \Gamma_N \le \lambda$, and summing a recursion $u_{k+1} \le \beta^N u_k + c\,\mathbb{E}\sigma_k + d$ with $u_k \ge 0$ gives*

$$\frac{1}{K}\sum_{k=1}^{K}u_k \le \frac{u_1}{\Gamma_N K} + \frac{c}{\Gamma_N}\bar{\sigma}_K + \frac{d}{\Gamma_N}.$$

The RR endpoint coefficient after tracking is therefore $\sqrt{r\theta_N}/\Gamma_N = \Theta(\lambda\sqrt{r/N})$, whereas the IID coefficient $\sqrt{rQ_N}/\Gamma_N = \Theta(\sqrt{r/N})$ carries no factor λ, so replacing $\bar{\sigma}_K$ by a fixed positive proxy leaves a nondecaying IID endpoint term at fixed N, r. This describes the present closure, and stepwise IID analyses obtain vanishing bounds under their own assumptions (Cutkosky & Mehta, 2020; Shen et al., 2026).

For fixed $\beta \in [0,1)$ and $N \to \infty$, both Q_N and $\theta_N = (NQ_N - \Gamma_N^2)/(N-1)$ tend to $(1-\beta)/(1+\beta)$ while $\Gamma_N \to 1$, so $\sqrt{r\theta_N}/\Gamma_N \to \sqrt{r(1-\beta)/(1+\beta)} > 0$, and this coefficient alone does not yield a fixed-momentum guarantee with a fixed positive heterogeneity proxy.

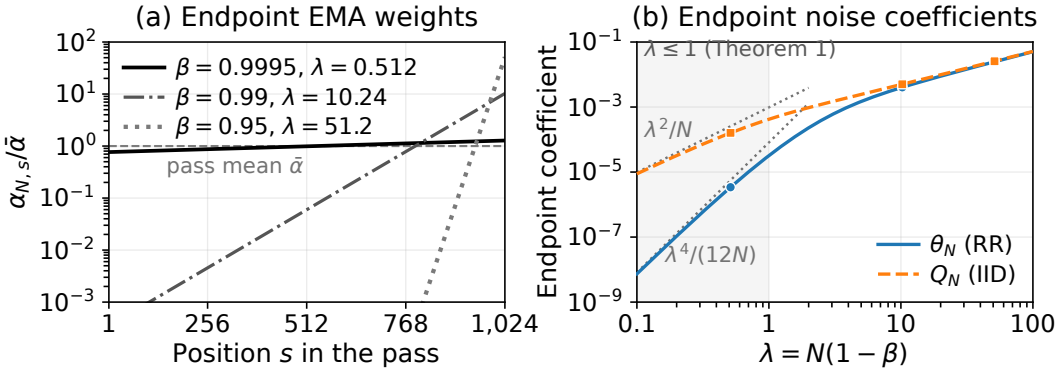


Figure 2: **Endpoint EMA weights and frozen-noise coefficients at $N = 1{,}024$.** (a) The weights $\alpha_{N,s} = (1-\beta)\beta^{N-s}$ that the buffer assigns at the end of a pass to the sample at position s, divided by their pass mean $\bar{\alpha} = (1-\beta^N)/N$, for the three pilot momenta of Appendix G.2. At $\beta = 0.9995$, where $\lambda = N(1-\beta) = 0.512$ satisfies the momentum condition $\lambda \le 1$ of Theorem 1, every weight lies between 0.76 and 1.28 times the mean, so the residual $\sum_s(\alpha_{N,s} - \bar{\alpha})\xi_{\pi_s}$ left by the zero-sum identity of Section 4.2 is small. At $\beta = 0.95$ the last 100 samples carry 99.4% of the total weight, and the identity removes only 3.7% of the squared coefficient (Appendix D.2). (b) The squared endpoint coefficients θ_N under RR (Corollary 1) and Q_N under IID (Eq. (10)) as functions of λ, with markers at the three pilot momenta and dotted lines for the leading small-λ terms $\lambda^4/(12N)$ and λ^2/N. For $\lambda \le 1$ the RR coefficient is smaller by a factor of order λ^2, with $\theta_N/Q_N = 0.021$ at $\lambda = 0.512$ and 0.076 at $\lambda = 1$, whereas at the fixed momenta $\beta = 0.99$ and 0.95 the ratio is 0.81 and 0.96. All values are computed from the closed forms in Lemma 5 and involve no training run.

D.2 What remains of frozen cancellation at fixed momentum

For any fixed $0 < \beta < 1$ and $N \ge 2$, retain the centered frozen gradients of Lemma 5. Let $\Gamma_N = 1 - \beta^N$ and let $Q_N = (1-\beta)(1-\beta^{2N})/(1+\beta)$ be their IID squared-Frobenius coefficient, as in Eq. (10). The RR coefficient is $\theta_N = (NQ_N - \Gamma_N^2)/(N-1)$, so

$$\frac{\theta_N}{Q_N} = \frac{N - \dfrac{1+\beta}{1-\beta}\dfrac{1-\beta^N}{1+\beta^N}}{N-1}. \tag{11}$$

The formula follows by substituting the geometric sums in the weighted identity and does not assume $N(1-\beta) \le 1$. Indeed, for the nonnegative weights $w_s = (1-\beta)\beta^{N-s}$, $Q_N = \sum_s w_s^2 \le (\sum_s w_s)^2 = \Gamma_N^2 \le NQ_N$, which gives $0 \le \theta_N/Q_N \le 1$. If all centered gradients vanish, both noise moments vanish and the coefficient comparison is interpreted algebraically, not as a ratio of nonzero observed variances.

The LM setting $N = 1{,}024$, $\beta = 0.95$ gives $\theta_N/Q_N = 0.962854$, a 3.71% reduction of this *squared Frobenius frozen-noise coefficient*, not of training loss or wall-clock time. For fixed β and increasing N, Eq. (11) tends to one by the limits in Remark 1. Thus cancellation persists at finite N, but this identity alone does not remove moving-trajectory drift or close a vanishing fixed-momentum rate. It quantifies why the language-model experiment complements rather than directly verifies the high-memory analysis.

E Proof of Theorem 1 and Approximate-Direction Extensions

E.1 Descent, tracking, and the potential recurrence

We first obtain an explicit bound for an arbitrary fixed stepsize and then close its heterogeneity term using a first-moment potential function. The auxiliary estimates below assume $N \ge 2$, $L > 0$, $\lambda \in (0, 1]$, $\beta = 1 - \lambda/N$, and $\eta > 0$ constant within the entire K-epoch run. The degenerate $H = 0$ case is handled separately in the theorem proof below. Each epoch uses a uniform permutation

independent of all past epochs. All updates use the exact polar direction. Within an epoch, suppress its superscript and write

$$M_i = \beta M_{i-1} + \nabla f_{\pi_i}(W_{i-1}), \qquad S_i = (1-\beta)M_i,$$
$$O_i = \mathrm{Polar}(S_i) = \mathrm{Polar}(M_i), \qquad W_i = W_{i-1} - \eta O_i.$$

Thus $\|O_i\|_{\mathrm{op}} \le 1$ and

$$\|W_t - W_0\|_{\mathrm{op}} \le \eta t \quad (0 \le t \le N), \qquad \beta^N \le e^{-\lambda} \le 1 - \lambda/2. \tag{12}$$

The path bound follows from summing the updates and the triangle inequality. For the contraction, $\log(1-x) \le -x$ gives $\beta^N \le e^{-\lambda}$, and convexity gives $e^{-\lambda} \le (1-\lambda) + \lambda e^{-1} \le 1 - \lambda/2$ for $0 \le \lambda \le 1$. For fixed K and deterministic initialization, the path bound, smoothness, and the finite reshuffling sample space ensure that all expectations below are finite. Set $g_{i-1} := \nabla f(W_{i-1})$, and let ρ_k and σ_k be as in Appendix A. Here ρ_k is the random buffer-error norm, and the tracking argument controls its expectation $\mathbb{E}\rho_k$. The notation $\mathbb{E}_k$ is the conditional expectation given $\mathcal{F}_{k-1}$ defined in Appendix A, and the terminal quantities with index $K+1$ use its epoch-boundary convention.

Lemma 6 (Linear-form descent)**.** *For every update,*

$$f(W_i) \le f(W_{i-1}) - \eta\|g_{i-1}\|_* + 2\eta\|g_{i-1} - S_i\|_* + \frac{L\eta^2}{2}.$$

Proof. Polar extremality (Proposition 1) and trace duality give

$$\begin{aligned}\langle g_{i-1}, O_i\rangle_F &= \langle S_i, O_i\rangle_F + \langle g_{i-1} - S_i, O_i\rangle_F \\ &\ge \|S_i\|_* - \|g_{i-1} - S_i\|_* \\ &\ge \|g_{i-1}\|_* - 2\|g_{i-1} - S_i\|_*.\end{aligned}$$

Apply Lemma 2 and $\|O_i\|_{\mathrm{op}} \le 1$. □

Frozen-gradient decomposition. This restates Eq. (4) in the epoch-local notation. Let $\xi_j := \nabla f_j(W_0) - \nabla f(W_0)$, and let

$$\hat{S}_i := \beta^i S_0 + (1-\beta)\sum_{s=1}^{i} \beta^{i-s}\nabla f_{\pi_s}(W_0)$$

be the buffer that the same permutation would produce from the frozen epoch-start gradients. Since $(1-\beta)\sum_{s=1}^{i}\beta^{i-s}\nabla f_{\pi_s}(W_0) = (1-\beta^i)\nabla f(W_0) + (1-\beta)\sum_{s=1}^{i}\beta^{i-s}\xi_{\pi_s}$,

$$g_{i-1} - S_i = \underbrace{g_{i-1} - \nabla f(W_0)}_{(\mathrm{I})_i} + \underbrace{\beta^i(\nabla f(W_0) - S_0)}_{(\mathrm{II})_i} \underbrace{- (1-\beta)\sum_{s=1}^{i}\beta^{i-s}\xi_{\pi_s}}_{(\mathrm{III})_i} + \underbrace{\hat{S}_i - S_i}_{(\mathrm{IV})_i}. \tag{13}$$

These are, in order, the full-gradient drift, initial buffer error, frozen RR noise, and buffer drift of Eq. (4) in the proof sketch.

Lemma 7 (Slow-EMA prefix and endpoint bounds)**.** *For every epoch,*

$$\sum_{i=1}^{N}\mathbb{E}_k\|(\mathrm{III})_i\|_* \le \sqrt{2rN}\lambda\sigma_k, \qquad \mathbb{E}_k\|(\mathrm{III})_N\|_* \le \sqrt{\frac{2r}{N}}\lambda^2\sigma_k. \tag{14}$$

Moreover, for $i \in [N]$,

$$\|(\mathrm{I})_i\|_* \le L\eta(i-1), \qquad \|(\mathrm{II})_i\|_* = \beta^i\rho_k, \qquad \|(\mathrm{IV})_i\|_* \le L\eta(i-1).$$

Proof. In the notation of Corollary 1, that corollary gives $\sum_i \theta_i \le 2\lambda^2$ and $\theta_N \le 2\lambda^4/N$. Conditional on the past, the ξ_j are fixed and centered, so the weighted RR identity gives $\mathbb{E}_k\|(\mathrm{III})_i\|_F^2 = \theta_i\sigma_k^2$. Using $\|X\|_* \le \sqrt{r}\|X\|_F$, Jensen's inequality, and Cauchy–Schwarz,

$$\sum_{i=1}^{N}\mathbb{E}_k\|(\mathrm{III})_i\|_* \le \sqrt{r}\,\sigma_k\sum_{i=1}^{N}\sqrt{\theta_i} \le \sqrt{rN\sum_i\theta_i}\,\sigma_k \le \sqrt{2rN}\lambda\sigma_k.$$

The endpoint estimate follows by the same norm conversion and $\theta_N \le 2\lambda^4/N$. For $(\mathrm{IV})_i$, unrolling $S_i = \beta S_{i-1} + (1-\beta)\nabla f_{\pi_i}(W_{i-1})$ from the epoch start and subtracting the definition of $\hat{S}_i$ gives

$$\hat{S}_i - S_i = (1-\beta)\sum_{s=1}^{i} \beta^{i-s}\big[\nabla f_{\pi_s}(W_0) - \nabla f_{\pi_s}(W_{s-1})\big].$$

By Assumption 1 and Eq. (12), each bracket has nuclear norm at most $L\eta(s-1) \le L\eta(i-1)$, and the weights are nonnegative with sum at most one, which proves the bound on $(\mathrm{IV})_i$. Smoothness of f (Proposition 2) and the same path bound control $(\mathrm{I})_i$, and the formula for $(\mathrm{II})_i$ is the definition of ρ_k. □

Lemma 8 (First-moment tracking and epoch mismatch)**.** *With momentum carried across epochs,*

$$\mathbb{E}_k \rho_{k+1} \le (1-\lambda/2)\rho_k + \sqrt{\frac{2r}{N}}\lambda^2\sigma_k + 2L\eta N, \tag{15}$$

$$\sum_{i=1}^{N} \mathbb{E}_k \|g_{i-1} - S_i\|_* \le N\rho_k + \sqrt{2rN}\lambda\sigma_k + L\eta N^2. \tag{16}$$

Proof. Since $(W_0^{k+1}, S_0^{k+1}) = (W_N^k, S_N^k)$, the endpoint decomposition is

$$\nabla f(W_N) - S_N = [\nabla f(W_N) - \nabla f(W_0)] + \beta^N[\nabla f(W_0) - S_0] + (\mathrm{III})_N + (\mathrm{IV})_N.$$

Taking norms and conditional expectations, and using Eq. (12) together with the endpoint bound in Eq. (14) and the drift bounds of Lemma 7,

$$\mathbb{E}_k \rho_{k+1} \le L\eta N + \beta^N \rho_k + \mathbb{E}_k\|(\mathrm{III})_N\|_* + L\eta(N-1) \le (1-\lambda/2)\rho_k + \sqrt{\frac{2r}{N}}\lambda^2\sigma_k + 2L\eta N.$$

For the second claim, sum the norms of the four terms of Eq. (13) over i and take $\mathbb{E}_k$. Lemma 7 gives

$$\sum_{i=1}^{N} \mathbb{E}_k \|g_{i-1} - S_i\|_* \le 2L\eta \sum_{i=1}^{N}(i-1) + \rho_k \sum_{i=1}^{N}\beta^i + \sqrt{2rN}\lambda\sigma_k,$$

and $\sum_i \beta^i \le N$ and $2\sum_i (i-1) \le N^2$ complete the bound. □

Proposition 3 (Explicit bound for an arbitrary fixed stepsize)**.** *Let $D := f(\tilde{W}_0) - f^*$ and $\bar{\rho}_1 := \mathbb{E}\rho_1$, with $\bar{\sigma}_K$ as in Appendix A. For any $\eta > 0$ and $\lambda \in (0, 1]$, the run satisfies*

$$\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\|\nabla f(W_{i-1}^k)\|_* \le \left[\frac{D}{\eta KN} + \frac{4\bar{\rho}_1}{\lambda K} + 6\sqrt{\frac{2r}{N}}\lambda\bar{\sigma}_K + \frac{10L\eta N}{\lambda} + \frac{L\eta}{2}\right]. \tag{17}$$

Proof. Taking total expectation in Eq. (15), rearranging, and summing over k gives

$$\frac{\lambda}{2}\sum_{k=1}^{K}\mathbb{E}\rho_k \le \bar{\rho}_1 - \mathbb{E}\rho_{K+1} + \sqrt{\frac{2r}{N}}\lambda^2\sum_{k=1}^{K}\mathbb{E}\sigma_k + 2L\eta NK.$$

Since $\rho_{K+1} \ge 0$,

$$\sum_{k=1}^{K}\mathbb{E}\rho_k \le \frac{2\bar{\rho}_1}{\lambda} + 2\sqrt{\frac{2r}{N}}\lambda\sum_{k=1}^{K}\mathbb{E}\sigma_k + \frac{4L\eta NK}{\lambda}.$$

Applying Eq. (16) and then this bound yields

$$\begin{aligned}\mathcal{E} &:= \sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\|g_{i-1}^k - S_i^k\|_* \\ &\le N\sum_{k=1}^{K}\mathbb{E}\rho_k + \sqrt{2rN}\lambda\sum_{k=1}^{K}\mathbb{E}\sigma_k + L\eta N^2K \\ &\le \frac{2N\bar{\rho}_1}{\lambda} + 3\sqrt{2rN}\lambda\sum_{k=1}^{K}\mathbb{E}\sigma_k + \left(\frac{4}{\lambda}+1\right)L\eta N^2K \\ &= \frac{2N\bar{\rho}_1}{\lambda} + 3\sqrt{2rN}\lambda\sum_{k=1}^{K}\mathbb{E}\sigma_k + \frac{4+\lambda}{\lambda}L\eta N^2K \\ &\le \frac{2N\bar{\rho}_1}{\lambda} + 3\sqrt{2rN}\lambda\sum_{k=1}^{K}\mathbb{E}\sigma_k + \frac{5L\eta N^2K}{\lambda},\end{aligned}$$

where the final step uses $0 < \lambda \le 1$. Telescoping Lemma 6 over all KN updates gives

$$\eta\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\|\nabla f(W_{i-1}^k)\|_* \le D + 2\eta\mathcal{E} + \frac{L\eta^2KN}{2}.$$

Indeed, the function-value sum is $f(\tilde{W}_0) - \mathbb{E}f(W_N^K) \le D$. Substituting the bound on $\mathcal{E}$ and dividing by ηKN proves the claim. □

Closing the heterogeneity term. Proposition 3 still contains the trajectory-dependent average $\bar{\sigma}_K$ in Eq. (17). The following argument bounds it along the run without imposing any independent variance condition. The potential recurrence below is used only to control $\bar{\sigma}_K$, and the stationarity bound of Corollary 2 is then read off from Eq. (17), as in the proof sketch of Section 4.2.

Lemma 9 (First-moment potential recurrence)**.** *Define the nonnegative potential*

$$\Psi_k := f(W_0^k) - \frac{1}{N}\sum_{j=1}^{N}f_j^* + \frac{4\eta N}{\lambda}\rho_k.$$

Then

$$\begin{aligned}\mathbb{E}_k\Psi_{k+1} \le \Psi_k &- \eta\sum_{i=1}^{N}\mathbb{E}_k\|\nabla f(W_{i-1}^k)\|_* \\ &+ 6\eta\sqrt{2rN}\lambda\sigma_k + \frac{11L\eta^2N^2}{\lambda}.\end{aligned} \tag{18}$$

Proof. Taking $\mathbb{E}_k$ in Lemma 6, summing over the epoch, and applying Eq. (16) gives

$$\begin{aligned}\mathbb{E}_k\left[f(W_N^k) - \frac{1}{N}\sum_{j=1}^{N}f_j^*\right] \le f(W_0^k) - \frac{1}{N}\sum_{j=1}^{N}f_j^* - \eta\sum_{i=1}^{N}\mathbb{E}_k\|\nabla f(W_{i-1}^k)\|_* \\ + 2\eta N\rho_k + 2\eta\sqrt{2rN}\lambda\sigma_k + 2L\eta^2N^2 + \frac{L\eta^2N}{2}.\end{aligned}$$

Multiplying Eq. (15) by $4\eta N/\lambda$ gives

$$\frac{4\eta N}{\lambda}\mathbb{E}_k\rho_{k+1} \le \frac{4\eta N}{\lambda}\rho_k - 2\eta N\rho_k + 4\eta\sqrt{2rN}\lambda\sigma_k + \frac{8L\eta^2N^2}{\lambda}.$$

Adding the two inequalities, the terms $\pm 2\eta N\rho_k$ cancel, the noise coefficients sum to 6, and the drift terms satisfy

$$2L\eta^2N^2 + \frac{8L\eta^2N^2}{\lambda} + \frac{L\eta^2N}{2} = \frac{L\eta^2N^2}{\lambda}\left(2\lambda + 8 + \frac{\lambda}{2N}\right) \le \frac{11L\eta^2N^2}{\lambda}$$

for $\lambda \le 1$ and $N \ge 2$, since $2 + 8 + 1/4 = 10.25$. This is Eq. (18).

Lemma 4 gives

$$\sigma_k^2 \le 2L\left(f(W_0^k) - \frac{1}{N}\sum_{j=1}^{N} f_j^*\right) \le 2L\Psi_k.$$

Writing $u_k := \mathbb{E}\Psi_k$, Jensen's inequality gives $\mathbb{E}\sigma_k \le \mathbb{E}\sqrt{2L\Psi_k} \le \sqrt{2Lu_k}$. Taking total expectations in Eq. (18) and dropping the nonpositive descent term therefore gives

$$u_{k+1} \le u_k + a\sqrt{u_k} + b, \qquad a := 12\eta\sqrt{rLN}\lambda, \qquad b := \frac{11L\eta^2N^2}{\lambda}. \tag{19}$$

□

E.2 Unified stability and stationarity for positive stepsize prefactors

We now close the scalar recursion and derive the stationarity bound once for every positive deterministic prefactor. The stepsize of Theorem 1 and the larger prefactors used in the experiments then follow as special cases, without changing the algorithm or its horizon dependence. In particular, the $\tau = 1/100$ specialization below and $\sigma_k^2 \le 2L\Psi_k$ prove both bounds in Eq. (7).

Lemma 10 (Stability for arbitrary positive stepsize prefactors)**.** *Under the assumptions and initialization of Theorem 1, keep its λ, β and let $H > 0$. For deterministic $\tau > 0$, set*

$$v := \min\{\lambda, \sqrt{\lambda/K}\}, \qquad \eta_\tau := \frac{\tau}{N}\sqrt{\frac{H}{L}}\,v, \qquad R_\tau := 6\tau + \sqrt{1 + 4\sqrt{2}\tau + 47\tau^2}.$$

Use the potential Ψ_k of Lemma 9, with $\eta = \eta_\tau$. Then

$$\max_{1\le k\le K+1} \mathbb{E}\Psi_k \le R_\tau^2 H, \qquad \bar\sigma_K \le \sqrt{2}R_\tau\sqrt{LH}. \tag{20}$$

In particular, $\tau = 1/100$ gives the schedule of Theorem 1 and satisfies

$$\max_{1\le k\le K+1} \mathbb{E}\Psi_k \le 2H, \qquad \bar\sigma_K \le 2\sqrt{LH}. \tag{21}$$

Proof. The inequalities (18) and (19) hold for arbitrary fixed $\eta > 0$. Write $u_k = \mathbb{E}\Psi_k$ and let a, b be the coefficients in Eq. (19). For completeness, the parameters in this lemma are

$$\lambda := \min\{1, (N/(rK))^{1/3}\}, \qquad v := \min\{\lambda, \sqrt{\lambda/K}\}, \qquad \eta_\tau := \frac{\tau}{N}\sqrt{\frac{H}{L}}\,v.$$

The two elementary schedule inequalities are

$$\frac{Kv^2}{\lambda} \le 1, \qquad \sqrt{\frac{r}{N}}\lambda Kv \le \sqrt{\frac{rK}{N}}\lambda^{3/2} \le 1.$$

The last inequality follows from $rK\lambda^3/N \le 1$, which holds both when $\lambda = 1$ (then $rK \le N$) and when the cap is inactive (then $rK\lambda^3/N = 1$). Also $\rho_1 = \|\nabla f(\tilde W_0)\|_* \le \sqrt{2LH}$ by Lemma 3, applied to f with lower bound $N^{-1}\sum_j f_j^*$, and $v/\lambda \le 1$. Consequently,

$$\begin{aligned} u_1 = \Psi_1 &= f(\tilde W_0) - \frac{1}{N}\sum_{j=1}^{N} f_j^* + \frac{4\eta_\tau N}{\lambda}\rho_1 = H + \frac{4\eta_\tau N}{\lambda}\rho_1 \\ &\le H + 4\frac{\tau}{N}\sqrt{\frac{H}{L}}\,\frac{v}{\lambda}N\sqrt{2LH} \\ &\le H + 4\sqrt{2}\tau H = (1 + 4\sqrt{2}\tau)H, \end{aligned}$$

and

$$Ka = 12\tau\sqrt{H}\sqrt{\frac{r}{N}}\lambda Kv \le 12\tau\sqrt{H},$$

$$Kb = 11\tau^2 H\frac{Kv^2}{\lambda} \le 11\tau^2 H.$$

In particular,

$$u_1 \le (1+4\sqrt{2}\tau)H, \qquad Ka \le 12\tau\sqrt{H}, \qquad Kb \le 11\tau^2 H. \tag{22}$$

All expectations are finite by Eq. (12). Let $U := \max_{1\le j\le K+1} u_j$, a maximum of expectations, and let j_* attain it. Summing Eq. (19) over $k < j_*$ and using $u_k \le U$ gives

$$U - u_1 \le a(j_*-1)\sqrt{U} + b(j_*-1) \le Ka\sqrt{U} + Kb,$$

also valid for $j_* = 1$ (an empty sum). The budget estimates in Eq. (22) then give

$$U \le H(1+4\sqrt{2}\tau+11\tau^2) + 12\tau\sqrt{H}\sqrt{U}.$$

Set $x = \sqrt{U/H}$. Dividing by $H > 0$ and moving the term $12\tau x$ to the left gives

$$x^2 - 12\tau x \le 1 + 4\sqrt{2}\tau + 11\tau^2.$$

Adding $36\tau^2$ to both sides completes the square, so that

$$(x-6\tau)^2 \le 1 + 4\sqrt{2}\tau + 47\tau^2, \qquad x \le 6\tau + \sqrt{1+4\sqrt{2}\tau+47\tau^2} = R_\tau.$$

Thus $U \le R_\tau^2 H$, proving the first inequality in Eq. (20) without a smallness condition on τ. This uniform bound and Jensen's inequality give, at every epoch,

$$\mathbb{E}\sigma_k \le \sqrt{2L\mathbb{E}\Psi_k} \le \sqrt{2LR_\tau^2 H} = \sqrt{2}R_\tau\sqrt{LH}.$$

The trajectory-dependent average noise $\bar{\sigma}_K$ is defined as the arithmetic mean of $\mathbb{E}\sigma_k$ over all K epochs. Averaging this uniform upper bound proves the second inequality in Eq. (20).

For the special case $\tau = 1/100$, the definition gives $R_{1/100} < 11/10$, so $R_{1/100}^2 < 121/100 < 2$. Therefore $u_k \le 2H$ for every $1 \le k \le K+1$, and the same Jensen calculation gives

$$\mathbb{E}\sigma_k \le \sqrt{2L\mathbb{E}\Psi_k} \le \sqrt{2L(2H)} = 2\sqrt{LH}.$$

Averaging this last bound proves Eq. (21). This recovers the smaller-prefactor stability conclusion from the same maximal-expectation argument. □

Corollary 2 (Stationarity for positive stepsize prefactors)**.** *Under the hypotheses of Lemma 10, define*

$$C_\tau := \max\{\tau^{-1} + 4\sqrt{2} + 12R_\tau + 11\tau,\ \sqrt{2}R_\tau + 2\tau\}.$$

Both stationarity averages of Theorem 1 are bounded by

$$C_\tau\sqrt{LH}(K\lambda)^{-1/2} \le C_\tau\sqrt{LH}\left(K^{-1/2} + r^{1/6}K^{-1/3}N^{-1/6}\right). \tag{23}$$

In particular, $C_\tau < 200$ for $\tau \in \{1/100, 1/25, 1\}$. Thus multiplying the stepsize in Eq. (1) *by* 1, 4, *or* 100 *preserves its displayed final constant* 200*. A known deterministic $\widehat{H} \ge H > 0$ can replace H in both the schedule and bound. When $H = 0$, the zero-gradient argument in the theorem proof below applies for every stepsize.*

Proof. First suppose $K\lambda \ge 1$, so $v = \sqrt{\lambda/K}$ and $\eta_\tau = \tau\sqrt{H/L}\sqrt{\lambda/K}/N$. Write $d := (K\lambda)^{-1/2} \le 1$, so that $\sqrt{\lambda/K} = \lambda d \le d$ and $1/(K\lambda) = d^2 \le d$. Proposition 3 applies with $\eta = \eta_\tau$, and its five terms are bounded using $D \le H$, the initial buffer error $\bar{\rho}_1 = \|\nabla f(\tilde{W}_0)\|_* \le \sqrt{2LH}$, the stability estimate $\bar{\sigma}_K \le \sqrt{2}R_\tau\sqrt{LH}$ of Eq. (20), and $\sqrt{r/N}\lambda \le d$ from the choice of λ (the second schedule inequality of Lemma 10). In order,

$$\frac{D}{\eta_\tau KN} \le \frac{H}{\tau\sqrt{H/L}\sqrt{\lambda/K}\,K} = \tau^{-1}\sqrt{LH}\,d, \qquad \frac{4\bar{\rho}_1}{\lambda K} \le \frac{4\sqrt{2LH}}{K\lambda} \le 4\sqrt{2}\sqrt{LH}\,d,$$

$$6\sqrt{\frac{2r}{N}}\lambda\bar{\sigma}_K \le 12R_\tau\sqrt{LH}\sqrt{\frac{r}{N}}\lambda \le 12R_\tau\sqrt{LH}\,d, \qquad \frac{L\eta_\tau}{2} = \frac{\tau\sqrt{LH}\sqrt{\lambda/K}}{2N} \le \frac{\tau}{4}\sqrt{LH}\,d,$$

$$\frac{10L\eta_\tau N}{\lambda} = \frac{10\tau\sqrt{LH}\sqrt{\lambda/K}}{\lambda} = 10\tau\sqrt{LH}\,d.$$

It follows that the inner-iterate average is at most

$$\sqrt{LH}\big(\tau^{-1} + 4\sqrt{2} + 12R_\tau + \tfrac{41}{4}\tau\big)(K\lambda)^{-1/2}. \tag{24}$$

To obtain the epoch-start average, smoothness gives

$$\|\nabla f(W_0^k)\|_* \le \|\nabla f(W_{i-1}^k)\|_* + L\eta_\tau(i-1).$$

Averaging over i and using $L\eta_\tau N/2 = \tau\sqrt{LH}\,\sqrt{\lambda/K}/2 \le \tau\sqrt{LH}\,d/2$ yields

$$\|\nabla f(W_0^k)\|_* \le \frac{1}{N}\sum_i \|\nabla f(W_{i-1}^k)\|_* + \frac{\tau}{2}\sqrt{LH}(K\lambda)^{-1/2}.$$

Taking expectations, averaging over k, and combining with Eq. (24), both averages are at most

$$\sqrt{LH}\big(\tau^{-1} + 4\sqrt{2} + 12R_\tau + 11\tau\big)(K\lambda)^{-1/2},$$

the first entry in the maximum defining C_τ, which therefore bounds both averages in this regime.

For $K\lambda < 1$, $\eta_\tau = \tau\sqrt{H/L}\lambda/N$. Lemma 3 with lower bound $N^{-1}\sum_j f_j^*$ and Eq. (20) imply

$$\begin{aligned}\mathbb{E}\|\nabla f(W_0^k)\|_* &\le \mathbb{E}\sqrt{2L\left(f(W_0^k) - \frac{1}{N}\sum_{j=1}^N f_j^*\right)} \\ &\le \sqrt{2L\mathbb{E}\Psi_k} \le \sqrt{2}R_\tau\sqrt{LH}.\end{aligned}$$

Each inner pre-update iterate has expected gradient norm at most $\sqrt{2}R_\tau\sqrt{LH} + L\eta_\tau N \le (\sqrt{2}R_\tau + 2\tau)\sqrt{LH}$. Since $(K\lambda)^{-1/2} > 1$, the second term defining C_τ bounds both averages.

Finally, Eq. (23) follows for every $K \ge 1$ from

$$(K\lambda)^{-1/2} = \max\{K^{-1/2}, r^{1/6}K^{-1/3}N^{-1/6}\} \le K^{-1/2} + r^{1/6}K^{-1/3}N^{-1/6}.$$

For $\tau = 1$, $R_1 = 6 + \sqrt{48 + 4\sqrt{2}} < 134/10$ and $\sqrt{2} < 3/2$, so

$$12 + 4\sqrt{2} + 12R_1 < 178.8 < 200, \qquad \sqrt{2}R_1 + 2 < 200.$$

For $\tau = 1/100$ and $1/25$, respectively, the definition gives $R_\tau < 1.1$ and $R_\tau < 1.4$. Substitution in C_τ gives $C_{1/100} < 120$ and $C_{1/25} < 49$, both below 200. The same budget estimates justify replacing H by a known positive $\widehat{H} \ge H$, as detailed in the theorem proof below. □

The guarantee at $\tau = 1$ uses $R_1^2 H \approx 177.56H$ and $C_1 \approx 177.56$, not the $2H$ stability estimate for $\tau = 1/100$. For an asymptotic family, τ is held fixed (or the displayed C_τ must be controlled uniformly), whereas the pointwise finite-horizon statement allows every deterministic $\tau > 0$. At fixed N, r, L, H, τ, the bound is $O(K^{-1/3})$.

Proof of Theorem 1. First consider $H = 0$. The nonnegative component gaps satisfy $N^{-1}\sum_j (f_j(\tilde{W}_0) - f_j^*) = 0$, so each gap vanishes. Lemma 3, applied to each component with lower bound f_j^*, gives $\nabla f_j(\tilde{W}_0) = 0$ for every j. Starting from $M_0^1 = 0$, induction over updates using $\mathrm{Polar}(0) = 0$ gives $M_i^k = 0$ and $W_i^k = \tilde{W}_0$ throughout. Thus both stationarity averages and output norms are zero for any $\eta \ge 0$, including the zero stepsize in Eq. (1) or a positive stepsize based on $\widehat{H} > 0$.

Now suppose $H > 0$. Choosing $\tau = 1/100$ in Lemma 10 and Corollary 2 gives exactly the stepsize in Eq. (1). The lemma supplies the $2H$ stability estimate in Eq. (21), while the corollary covers both $K\lambda \ge 1$ and $K\lambda < 1$ and has $C_{1/100} < 120 < 200$. Hence both stationarity averages satisfy Eq. (2). For the output rules in the statement, independence and uniformity give

$$\mathbb{E}\|\nabla f(W_0^{k_{\mathrm{out}}})\|_* = \frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\nabla f(W_0^k)\|_*,$$

$$\mathbb{E}\|\nabla f(W_{i_{\mathrm{out}}-1}^{k_{\mathrm{out}}})\|_* = \frac{1}{KN}\sum_{k=1}^K\sum_{i=1}^N \mathbb{E}\|\nabla f(W_{i-1}^k)\|_*.$$

The two average bounds therefore give precisely the stated output guarantees.

A known positive upper bound. If only a deterministic $\widehat{H} > 0$ with $\widehat{H} \geq H$ is known, use $\widehat{H}$ in the stepsize instead. For $H > 0$, set $\eta_\tau = (\tau/N)\sqrt{\widehat{H}/L}\, v$ and retain the potential Ψ_k with this stepsize. The schedule inequalities are unchanged, and the initial bound uses

$$\Psi_1 = H + \frac{4\eta_\tau N}{\lambda}\rho_1, \qquad \rho_1 \leq \sqrt{2LH} \leq \sqrt{2L\widehat{H}}.$$

Consequently, all right-hand sides of the budget estimates in Eq. (22) hold with H replaced by $\widehat{H}$. The maximal-expectation argument then gives $\max_k \mathbb{E}\Psi_k \leq R_\tau^2 \widehat{H}$, and the five terms of Proposition 3 are bounded as in the proof of Corollary 2 with $D \leq H \leq \widehat{H}$ and $\bar{\rho}_1 \leq \sqrt{2L\widehat{H}}$, which gives Eq. (23) with $\sqrt{L\widehat{H}}$. At $\tau = 1/100$, this gives $\max_k \mathbb{E}\Psi_k \leq 2\widehat{H}$ and the theorem's displayed bound with $\sqrt{L\widehat{H}}$. For $H = 0$, the earlier argument applies to any stepsize, including this replacement. □

E.3 A FIXED-DEPTH GUARANTEE FOR THE JORDAN QUINTIC WITHOUT A SPECTRAL GAP

The quintic iteration of Jordan et al. (2024) starts from $X_0 = M/\max\{1, \|M\|_F\}$, and its qth iterate defines $\mathcal{J}_q(M) := X_q$. Here a *spectral gap* means a uniform positive lower bound on the nonzero singular values of X_0, not separation between its largest singular values. Lemma 11 and Proposition 4 give exact-arithmetic guarantees without this bound, and neither is used in Theorem 1.

Lemma 11 (Bounded alignment of the Jordan quintic)**.** *In exact arithmetic, initialize* $X_0 = M/\max\{1, \|M\|_F\}$ *for a matrix* $M \in \mathbb{R}^{m\times p}$*, and iterate*

$$X_{j+1} = \big(a_\varphi I_m + b_\varphi X_j X_j^\top + c_\varphi (X_j X_j^\top)^2\big) X_j \quad (j \geq 0),$$

with $(a_\varphi, b_\varphi, c_\varphi) = (3.4445, -4.7750, 2.0315)$*. Let* $C_\varphi = 1.21$*. For every integer* $q \geq 0$*,*

$$\|X_q\|_{\mathrm{op}} \leq C_\varphi, \qquad \nu_q^{\mathcal{J}}(M) := \|X_q - \mathrm{Polar}(M)\|_{\mathrm{op}} < 1, \qquad \langle M, X_q\rangle_F \geq (1 - \nu_q^{\mathcal{J}}(M))\|M\|_*.$$

Proof. In a fixed singular-vector basis, each nonzero singular value evolves by $\varphi(s) = a_\varphi s + b_\varphi s^3 + c_\varphi s^5$. The quadratic factor $a_\varphi + b_\varphi t + c_\varphi t^2$ in $t = s^2$ has positive leading coefficient and discriminant $-5.189382 < 0$, so positive singular values stay positive and zeros stay zero. The derivative $\varphi'(s) = 3.4445 - 14.3250s^2 + 10.1575s^4$ is quadratic in s^2 with signs $+, -, -, +$ at $s = 0.55, 0.56, 1, C_\varphi$, so its positive roots are $s_- \in (0.55, 0.56)$ and $s_+ \in (1, C_\varphi)$, and the maximum of φ on $[0, C_\varphi]$ is attained at s_- or at C_φ. Since $\varphi''(s) = s(-28.65 + 40.63s^2) < 0$ on $[0.55, 0.56]$, concavity gives the exact rational bounds

$$\varphi(s_-) \leq \varphi(0.55) + 0.01\varphi'(0.55) = 1.202683218125 < C_\varphi,$$
$$\varphi(C_\varphi) = 0.97782903269315 < C_\varphi,$$

so $\varphi([0, C_\varphi]) \subseteq [0, C_\varphi)$. All singular values of X_0 lie in $[0, 1]$, so every positive one remains in $(0, C_\varphi) \subset (0, 2)$, which gives $\nu_q^{\mathcal{J}}(M) < 1$ against the canonical partial polar factor in the same basis, with both matrices zero when $M = 0$. The alignment bound follows from Lemma 1 and $\langle M, \mathrm{Polar}(M)\rangle_F = \|M\|_*$. Transposing the routine for $m > p$ preserves all conclusions. □

From alignment to a rate. Since $\varphi(1/2) = 1.188859375 > 1$ and $\varphi(1) = 0.701$, the map can leave the unit spectral ball and does not converge to the polar factor as q grows, so a transfer based on convergence to a unit-norm polar direction does not apply. The lemma supplies the positive alignment that a descent step needs, with C_φ in place of 1 in the path bound, but the margin $1 - \nu_q^{\mathcal{J}}(M)$ is not uniform over inputs and need not be bounded away from zero. The proposition below therefore lower bounds the inner product that descent uses instead of the polar error. This yields a convergence guarantee for the same Jordan polynomial and raw buffer normalization, at the cost of a weaker rank factor and a constant stepsize rescaling, and it does not change Theorem 1.

Proposition 4 (Fixed-depth Jordan Muon under RR)**.** *Under the assumptions, initialization, and momentum of Theorem 1, fix any integer* $q \geq 1$ *and use the map* $\mathcal{J}_q$ *of Lemma 11 in exact arithmetic. Keep the classical buffer recursion* $M_i^k = \beta M_{i-1}^k + G_i^k$*, but replace the parameter update by*

$$W_i^k = W_{i-1}^k - \frac{\eta}{C_\varphi}\mathcal{J}_q(M_i^k), \qquad C_\varphi = 1.21,$$

where η is the stepsize in Eq. (1)*, or its factor-*4 *or factor-*100 *counterpart. Then*

$$\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\|\nabla f(W_{i-1}^{k})\|_{*} \le 400C_{\varphi}\sqrt{rLH}\left(K^{-1/2}+r^{1/6}K^{-1/3}N^{-1/6}\right) + \frac{\sqrt{r}\,(1-\beta)}{4}. \tag{25}$$

The same bound holds for the uniformly sampled inner output of Theorem 1. No lower bound on positive singular values, buffer norms, or alignment margins is assumed. At fixed N, r, L, H, this bound is $O(K^{-1/3})$ for every fixed q, including $q = 5$.

Proof. **1. A depth-independent scalar lower bound.** Write $p(t) = a_{\varphi} + b_{\varphi}t + c_{\varphi}t^2$ and $\varphi(s) = s\,p(s^2)$, with the coefficients of Lemma 11. Exact rational arithmetic gives

$$\min_{t\in\mathbb{R}} p(t) = a_{\varphi} - \frac{b_{\varphi}^2}{4c_{\varphi}} > \frac{5}{8}, \qquad p(16/25) = 1.2206024 > 1.$$

Also $p'(t) < 0$ on $[0, 16/25]$. Hence $\varphi(s) \ge s$ for $0 \le s \le 4/5$, whereas $\varphi(s) \ge (5/8)s \ge 1/2$ for $4/5 \le s \le C_{\varphi}$. Combined with the invariant interval $\varphi([0, C_{\varphi}]) \subseteq [0, C_{\varphi}]$ proved in Lemma 11, these inequalities imply, by induction,

$$\varphi^{\circ q}(s) \ge \min\{s, 1/2\} \ge s/2 \quad (0 \le s \le 1,\ q \ge 0).$$

Indeed, values below $1/2$ cannot decrease, and values at least $1/2$ cannot subsequently fall below $1/2$. This lower bound does not assert convergence of the scalar iterates to one.

2. A scale-aware descent quantity. Put $\delta = 1 - \beta > 0$ and $S = \delta M$. Define, only within this proof,

$$h_{\delta}(x) = \frac{x^2}{\max\{\delta, x\}} \quad (x \ge 0), \qquad \omega = \frac{1}{2C_{\varphi}}.$$

Because $M/\max\{1, \|M\|_F\} = S/\max\{\delta, \|S\|_F\}$, the singular-value lower bound gives, for $O = \mathcal{J}_q(M)/C_{\varphi}$,

$$\|O\|_{\mathrm{op}} \le 1, \qquad \langle S, O\rangle_F \ge \frac{\|S\|_F^2}{2C_{\varphi}\max\{\delta, \|S\|_F\}} = \omega h_{\delta}(\|S\|_F).$$

The function h_{δ} is nonnegative and 2-Lipschitz, since its derivative is $2x/\delta$ below δ and one above δ. Thus, for $g = \nabla f(W)$ and $e = \|g - S\|_*$, norm comparison and Hölder's inequality imply

$$\begin{aligned}\langle g, O\rangle_F &\ge \omega h_{\delta}(\|S\|_F) - e \\ &\ge \omega h_{\delta}(\|g\|_F) - (1+2\omega)e \ge \omega h_{\delta}(\|g\|_F) - 2e,\end{aligned}$$

since $1 + 2\omega = 1 + 1/C_{\varphi} < 2$. The smoothness lemma now gives

$$f(W - \eta O) \le f(W) - \omega\eta h_{\delta}(\|g\|_F) + 2\eta e + L\eta^2/2. \tag{26}$$

3. The RR tracking and stability estimates still close. The rescaled direction has norm at most one, so the path bound of Eq. (12), the EMA recursion, and consequently Lemmas 7–8 hold along this new trajectory. Repeating the telescoping calculation of Proposition 3 with Eq. (26) bounds $\omega(KN)^{-1}\sum_{k,i}\mathbb{E}h_{\delta}(\|\nabla f(W_{i-1}^{k})\|_F)$ by the same five terms on the right-hand side of Eq. (17), with $\bar{\sigma}_K$ and the buffer errors computed along the Jordan trajectory. This is a reapplication of the calculation, not an identification of Jordan and exact-polar iterates. In the potential calculation the new descent term is still nonpositive and is dropped. The $\pm 2\eta N\rho_k$ cancellation, scalar recurrence, and stability estimate of Lemma 10 therefore remain unchanged. For $K\lambda \ge 1$, the five estimates in the proof of Corollary 2 give

$$\frac{1}{KN}\sum_{k,i}\mathbb{E}h_{\delta}(\|\nabla f(W_{i-1}^{k})\|_F) \le 2C_{\varphi}\,200\sqrt{LH}\left(K^{-1/2}+r^{1/6}K^{-1/3}N^{-1/6}\right).$$

4. Convert back to stationarity. For $0 \le x \le \delta$, the largest value of $x - h_{\delta}(x) = x - x^2/\delta$ is $\delta/4$, and above δ this difference is zero. Consequently $\|g\|_* \le \sqrt{r}\|g\|_F \le \sqrt{r}\{h_{\delta}(\|g\|_F) + \delta/4\}$.

Averaging proves Eq. (25) when $K\lambda \geq 1$. If $K\lambda < 1$, the gradient-gap and path-bound argument in Corollary 2 uses only the unchanged stability estimate, so its constant-200 bound already holds and is no larger than the displayed bound. If $H = 0$, every initial component gradient vanishes and $\mathcal{J}_q(0) = 0$, so all iterates remain at the initial point. Finally $\delta = \lambda/N = O(K^{-1/3})$ at fixed N, r once $K \geq N/r$, and independent uniform output selection gives the stated output rule. □

What this adds, and what it does not. Proposition 4 closes the fixed-depth Jordan case without a horizon-dependent approximation factor, by replacing closeness to the polar map with a quadratic-to-linear descent quantity whose constant does not depend on q. The raw buffer input matters because the cutoff 1 becomes $1 - \beta$ in EMA units, which vanishes under the horizon tuning. The 124M runs normalize by $\|M\|_F$ (Appendix G.1), and with this cutoff-free normalization the same proof applies with $h(x) = x$, so Eq. (25) holds without the additive term $\sqrt{r}(1 - \beta)/4$. The result keeps the tuned momentum, a rescaled stepsize, fixed components and exact arithmetic, does not retain the sharp polar rank factor, and is not a guarantee for the fixed-β, mixed-precision, multi-matrix language-model runs. The exact-rational certificate and numerical matrix checks are in `theory/JORDAN_RR_CERTIFICATE.json`.

F Closest Related Guarantees and Oracle Differences

F.1 Related algorithms and oracle costs

Complementary RR and matrix-optimization results. Beyond Section 2, RR theory includes refined SGD bounds (Safran & Shamir, 2020; Rajput et al., 2020; Ahn et al., 2020; Koloskova et al., 2024; Liu, 2026), convex and accelerated shuffling (Tran et al., 2025; 2022), and minimax methods (Das et al., 2022; Cho & Yun, 2023). For Muon, Nagashima & Iiduka (2026) study learning-rate and batch-size schedules, with the same unit-batch nuclear-norm $T^{-1/4}$ order as the IID row of Table 1. Other work covers trust-region and spectral-norm-constrained interpretations (Kovalev, 2025; Chen et al., 2025), heavy-tailed noise or Hölder smoothness (Choudhury et al., 2026; Iiduka, 2026), layer-wise LMO updates (Riabinin et al., 2026), and generalized gradient clipping (Pethick et al., 2025b). Qian & Richtárik (2026) treat inexact Muon-type LMOs with vanishing positive singular values under additional assumptions and layer-wise (L^0, L^1)-smoothness. Proposition 4 instead assumes component operator–nuclear smoothness for the fixed-depth Jordan result, with weaker rank dependence.

Shuffling momentum and implicit bias. SMG (Tran et al., 2021, Algorithm 1 and Assumption 1(c)) keeps an epoch-start momentum anchor fixed within a pass and assumes that component variance grows at most with the squared full-gradient norm. Our heterogeneity control instead follows from the current objective gap. Li et al. (2026, arXiv v2, Theorem 4.4 and Lemma C.12) already use RR carried momentum, spectral-norm directions and epoch zero-sum cancellation. Their approximate max-margin result requires separable multiclass linear classification and loss-specific margin conditions, whereas Theorem 1 gives general nonconvex stationarity.

Assumptions behind Table 1 and the Muon-MVR2 coefficient. Table 1 uses Assumption 1, which implies component Frobenius smoothness with the same constant, together with a lower-bounded objective, and its RR rows also require component lower bounds. Its rates average expected *unsquared* gradient norms, and constants belong to each comparator. Muon-MVR2 uses component Frobenius smoothness, a lower-bounded objective, an unbiased uniformly bounded-variance oracle and an anytime schedule, and it evaluates the same sample at the current and previous parameters, so at most two component-gradient evaluations serve each update, whereas Algorithm 1 uses one component gradient per update without correction or epoch anchor.

For Muon-MVR2, retain the coefficient in Chang et al. (2025, arXiv v5, Theorem 4.2), where $\mathcal{U}$ counts updates. In that result's notation, $X_t \in \mathbb{R}^{m \times n}$ with $m \geq n$, $L > 0$ is the component

Frobenius-smoothness constant, σ^2 bounds the variance of each unbiased sample gradient, and f^* is an objective lower bound. The coefficient is

$$\begin{aligned} G_{\mathcal{U}} &= f(X_1) - f^* + 2L^{-1}\sigma^2 + 2Ln \\ &\quad + \left(16Ln + 4L^{-1}\sigma^2 + \frac{Ln}{2}\right)(1 + \log \mathcal{U}) \\ &\quad + \sqrt{4\sigma^2 + (32L^2n + 8\sigma^2)(1 + \log \mathcal{U})}. \end{aligned}$$

With $n = r$ and $\mathcal{U} \geq 1$, the coefficient satisfies

$$G_{\mathcal{U}} \leq \left[f(X_1) - f^* + 6L^{-1}\sigma^2 + \frac{37}{2}Lr + \sqrt{32L^2r + 12\sigma^2}\right](1 + \log \mathcal{U}).$$

Thus $G_{\mathcal{U}} = \mathcal{O}(r(1 + \log \mathcal{U}))$ at fixed $L > 0, \sigma$ and initial gap. Taking $\mathcal{U} = \lfloor T/2 \rfloor \geq 1$ gives $\mathcal{U} = \Theta(T)$ and respects T queries because the method uses at most $2\mathcal{U}$ evaluations. After the $\sqrt{r}$ Frobenius-to-nuclear conversion (Proposition 1), Table 1 reports $\mathcal{O}(r^{3/2}T^{-1/3}(1 + \log T))$. This is a simplified stated upper bound, not a claim of optimal dimension dependence, and the logarithm belongs to the cited anytime schedule.

Relation to the momentum-LMO analysis of Shulgin et al. (2026b). Their Theorem A.1 combines smoothness in a primal–dual norm pair, an unbiased oracle and a state-dependent bound on the sampled gradient's second moment (their Eqs. (2)–(4)), with first-moment tracking and a self-bounding treatment of momentum's history dependence, and their Appendix A.3 closes the average first-moment momentum error through the average expected objective gap $f - f^*$ before solving a scalar square-root inequality. Our component assumptions imply that condition for a fresh uniform component gradient, with their coefficients $A = L$, $B = 0$ and $C = 2L(f^* - N^{-1}\sum_j f_j^*)$ when their lower bound is f^*, and their two norm-compatibility constants may be taken as $\sqrt{r}$ and 1 for the nuclear–Frobenius pair, so unbounded component variance still permits an IID guarantee.

Their without-replacement sampling is confined to a fresh minibatch (their Eqs. (9)–(10), Corollary A.5). Zero conditional noise means remove cross-time covariance in their Eq. (44), whereas RR need not have this property. We instead condition at epoch starts and retain the covariance of frozen gradients over the full pass.

What is new here is the joint potential whose coefficient cancels the carried nuclear-norm buffer error without losing the extra λ factor from endpoint cancellation (Lemma 9), giving the maximum of expected potentials in Eq. (7). Neither the component self-bound (Mishchenko et al., 2020, Proposition 2) nor scalar self-bounding is claimed as new. For fixed nondegenerate stochastic constants, their Appendix A.3, Eq. (57) and Theorem A.3 give $T^{-1/4}$ and a sufficient ϵ^{-4} budget, compared with our $T^{-1/3}$ bound at fixed N, r (Section 4.3).

The finite-NS spectral smoothing of Li & Tsuchiya (2026) uses an online-to-nonconvex conversion with unbiased queries between iterates, a different recursion and oracle from Algorithm 1. Remark 1 identifies the limitation of our endpoint closure under IID.

F.2 Self-contained comparison statements

For the comparisons in Section 4.3 and Table 1, Proposition 5 proves the IID rate with an explicit constant. Appendix F summarizes the other cited guarantees, their hypotheses, and their oracle costs. Each comparator has its own constants and iterates, even where symbols coincide.

Proposition 5 (IID exact-polar comparator)**.** *Let $m, p, b, T \geq 1$ be integers, $L > 0$, and W_0 deterministic. Let $f : \mathbb{R}^{m\times p} \to \mathbb{R}$ satisfy operator–nuclear L-smoothness and $f^* := \inf f > -\infty$, and let $D = f(W_0) - f^* > 0$ and $r = \min\{m, p\}$. At step t, let G_t average $b \geq 1$ fresh IID gradients at W_{t-1}, each unbiased with Frobenius variance at most σ^2, uniformly in its query point. Assume the fresh sampling draws are independent across steps. Initialize $M_0 = 0$ and use*

$$M_t = \beta M_{t-1} + G_t, \qquad W_t = W_{t-1} - \eta\,\mathrm{Polar}(M_t).$$

For $\sigma > 0$, set

$$\delta := 1 - \beta = \min\left\{1, \frac{\sqrt{LDb}}{\sigma\sqrt{rT}}\right\}, \qquad \eta = \sqrt{\frac{\delta D}{LT}}.$$

Then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(W_{t-1})\|_* \le 9\left[\sqrt{\frac{LD}{T}}+\frac{\sigma\sqrt{r}}{\sqrt{bT}}+\left(\frac{r\sigma^2 LD}{bT}\right)^{1/4}\right]. \tag{27}$$

This bound has the leading order of Kim & Oh (2026, Theorem 4), and the constant 9 and the $\sqrt{r}$ factor in the lower-order noise term follow from the stepwise first-moment argument below (Cutkosky & Mehta, 2020) rather than being quoted from that theorem. For $\sigma = 0$, take $\delta = 1$ and the same stepsize.

Proof. Write $g_t = \nabla f(W_{t-1})$, $\xi_t = G_t - g_t$, and $S_t = \delta M_t$. Positive scaling gives $\mathrm{Polar}(M_t) = \mathrm{Polar}(S_t)$. If $\mathcal{F}_{t-1}$ is the history preceding the fresh batch, then

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\|\xi_t\|_F^2 \mid \mathcal{F}_{t-1}] \le \sigma^2/b.$$

Unrolling $S_t = \beta S_{t-1} + \delta(g_t + \xi_t)$ gives

$$S_t - g_t = -\beta^t g_1 + \sum_{j=2}^{t}\beta^{t-j+1}(g_{j-1} - g_j) + \delta\sum_{j=1}^{t}\beta^{t-j}\xi_j.$$

The middle sum has nuclear norm at most $L\eta/\delta$, because each polar step has operator norm at most η. For $j < \ell$, conditional centering gives $\mathbb{E}\langle\xi_j, \xi_\ell\rangle_F = \mathbb{E}\langle\xi_j, \mathbb{E}[\xi_\ell \mid \mathcal{F}_{\ell-1}]\rangle_F = 0$. Thus the final sum has expected squared Frobenius norm at most

$$\frac{\delta^2\sigma^2}{b}\sum_{j=1}^{t}\beta^{2(t-j)} \le \frac{\delta\sigma^2}{b(1+\beta)} \le \frac{\delta\sigma^2}{b}.$$

By the norm comparisons and Cauchy–Schwarz,

$$\mathbb{E}\|S_t - g_t\|_* \le \beta^t\|g_1\|_* + \frac{L\eta}{\delta} + \sigma\sqrt{\frac{r\delta}{b}}.$$

Polar extremality, the triangle inequality, and trace duality give $\langle g_t, \mathrm{Polar}(S_t)\rangle_F \ge \|g_t\|_* - 2\|g_t - S_t\|_*$. The descent lemma and telescoping give

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le \frac{D}{\eta T} + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t - S_t\|_* + \frac{L\eta}{2}.$$

Substituting the mismatch bound, $\sum_{t=1}^{T}\beta^t \le 1/\delta$, $\|g_1\|_* \le \sqrt{2LD}$, and $2L\eta/\delta + L\eta/2 \le 5L\eta/(2\delta)$ for $\delta \le 1$ therefore gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|g_t\|_* \le \frac{D}{\eta T} + \frac{5L\eta}{2\delta} + \frac{2\sqrt{2LD}}{\delta T} + 2\sigma\sqrt{\frac{r\delta}{b}}.$$

Set $c = \sqrt{LD/T}$ and $h = (r\sigma^2 LD/(bT))^{1/4}$. If $\delta = 1$, its defining condition gives $\sigma\sqrt{r/b} \le c$, so the right side is at most $(11/2 + 2\sqrt{2})c < 9c$. If $\delta < 1$, the choice of δ gives $\sqrt{LD/(\delta T)} = \sigma\sqrt{r\delta/b} = h$, so its first, second, and fourth terms sum to $(1 + 5/2 + 2)h = (11/2)h$, and its third term is $2\sqrt{2}\sigma\sqrt{r}/\sqrt{bT}$. Since $11/2 < 9$ and $2\sqrt{2} < 9$, this proves Eq. (27) without enlarging the $\sqrt{r}$ factor. When $\sigma = 0$, the same calculation with $\delta = 1$ gives the stated deterministic bound. □

Like the RR schedule of Theorem 1, this momentum is horizon-tuned, since $1 - \beta$ scales as $T^{-1/2}$ when $L, D, \sigma > 0$ and r, b are fixed.

Matched-query bounds and sufficient query budgets. For an independent uniform inner-iterate output, solving the two bounds in Eq. (8) for $\mathbb{E}\|\nabla f(W_{\mathrm{out}})\|_* \le \epsilon$ gives the sufficient budgets

$$T_{\mathrm{RR}}(\epsilon) = \mathcal{O}\left(N + \frac{NLH}{\epsilon^2} + \frac{\sqrt{rN}(LH)^{3/2}}{\epsilon^3}\right), \tag{28}$$

$$T_{\mathrm{IID}}(\epsilon) = \mathcal{O}\left(1 + \frac{LD + r\sigma^2}{\epsilon^2} + \frac{r\sigma^2 LD}{\epsilon^4}\right). \tag{29}$$

At fixed N, r, L and positive H, D, σ, these are $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-4})$ sufficient budgets. RR also permits uniform epoch-start output, and its integer choice is derived below. To derive the matched-query rates for $b = 1$, flatten the RR pre-update iterates as $W^{\mathrm{RR}}_{(k-1)N+i-1} := W^k_{i-1}$, for $k \in [K]$, $i \in [N]$. Then $T = KN$ is both the number of updates and the number of component-gradient queries, and

$$K^{-1/2} = \sqrt{N/T}, \qquad r^{1/6}K^{-1/3}N^{-1/6} = (rN)^{1/6}T^{-1/3}.$$

Substitution in Eq. (2) gives the RR row of Eq. (8), and Eq. (27) with $b = 1$ gives the IID row. A direct comparison requires the same function and initialization to satisfy both statements. Proposition 2 allows a common valid L, but IID additionally requires uniform variance. The gap H is distinct from D and σ, and each result tunes its own parameters.

For $D > 0$, $\sigma \geq 0$, and $\epsilon > 0$, choose the IID horizon

$$T = \left\lceil \max\left\{1, \frac{27^2 LD}{\epsilon^2}, \frac{27^2 r\sigma^2}{\epsilon^2}, \frac{27^4 r\sigma^2 LD}{\epsilon^4}\right\}\right\rceil,$$

and then choose δ, η for this horizon as in Proposition 5 with $b = 1$. The three terms in Eq. (27), each including its multiplier 9, are at most $\epsilon/3$ by the second, third, and fourth entries of this maximum, respectively. For an independent uniform $U \in [T]$, conditioning on the run gives

$$\mathbb{E}\|\nabla f(W_{U-1})\|_* = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(W_{t-1})\|_* \leq \epsilon.$$

Since $\lceil x \rceil \leq x + 1$ and a maximum of nonnegative terms is at most their sum, this choice is bounded by a universal constant times $1 + (LD + r\sigma^2)\epsilon^{-2} + r\sigma^2 LD\epsilon^{-4}$, which proves Eq. (29). The RR integer choice below completes whole epochs, so its ceiling contributes N rather than 1. The IID budget need not be rounded to an epoch boundary, although Eq. (8) compares the bounds at a common $T = KN$. If $\sigma = 0$, the IID accuracy budget has order $1 + LD/\epsilon^2$. If $D = 0$, Lemma 3 applied to f shows that W_0 is already stationary and can be returned without a query. If $H = 0$, the zero-gap branch of Theorem 1 applies to RR. These are sufficient, horizon-tuned gradient-query counts, excluding direction computation and the initial loss evaluation for H.

Derivation of the RR budget in Eq. (28)**.** For $H > 0$ and $\epsilon > 0$, it suffices to choose

$$K = \left\lceil \max\left\{1, \left(\frac{400\sqrt{LH}}{\epsilon}\right)^2, \left(\frac{400\sqrt{LH}\, r^{1/6}N^{-1/6}}{\epsilon}\right)^3\right\}\right\rceil.$$

Each term of Eq. (2) is then at most $\epsilon/2$. One component-gradient query per update gives $T = NK$ and the displayed order, and the ceiling contributes the N term. An available $\widehat{H} \geq H$ replaces H throughout. When $H = 0$, the initial iterate is already stationary and no updates are needed.

Scope of lower-bound comparisons. The ϵ^{-4} rate is minimax-optimal for general bounded-variance stochastic oracles (Arjevani et al., 2023), which does not establish an RR–IID lower-bound separation on a common fixed finite sum. The shuffling lower bounds of Cha et al. (2023) concern convex objectives and do not prove tightness of our nonconvex Muon guarantee. With other positive problem constants fixed, comparing the leading terms of the two sufficient budgets gives a crossover scale proportional to $\sqrt{r/N}$, not a guarantee of empirical dominance at every budget.

G Language-Model Protocol and Additional Results

This appendix specifies the main experiment and its statistical interpretation. Appendix D.2 quantifies the frozen-noise comparison at its momentum setting, and Appendix I states the broader scope limits.

G.1 Model, data, and optimizer

The model is a decoder-only Transformer with 123,944,448 trainable parameters, namely 12 layers, width 768, 12 attention heads and feed-forward width 3,072, with tied token embedding and output weights. It uses learned positional embeddings, bias-free linear layers, scale-only layer normalization,

GELU, context length 512 and the GPT-2 tokenizer (vocabulary 50,257). All runs start from random initialization without pretrained weights. Dropout is zero and activation checkpointing is enabled. The 48 hidden weight matrices use Muon, and embeddings and one-dimensional parameters use AdamW.

The source is FineWeb-Edu `sample-10BT` (Penedo et al., 2024), not the entire 10B-token sample. The preparation manifest fixes the dataset revision and tokenizer fingerprint. Text-hash partitioning assigns documents to train, validation or test before packing. Exact text duplicates and repeated source identifiers are removed, whereas near-duplicate removal is not claimed. Within each split, tokenized documents are concatenated with end-of-text markers and packed into disjoint rows of 513 tokens, providing 512 next-token targets each. The fixed training pool contains 32,768 rows (16,777,216 targets), and validation and test each contain 4,096 rows (2,097,152 targets). The same pool is used for all seeds and samplers. All reported test values are from final-budget checkpoints, evaluated after confirmation, without selecting on test loss.

A fixed grouping seed, 20260923, partitions training rows into 1,024 minibatch components of 32 sequences. Each run uses microbatches of four sequences and eight gradient-accumulation steps, hence batch 32. Eight A100-SXM4-40GB GPUs execute independent single-GPU runs, not distributed training of one model. GPU assignments rotate across conditions. Each run processes 8,192 updates, or 134,217,728 targets. A reporting pass spans 1,024 updates with no momentum or learning-rate reset.

The confirmation uses constant Muon learning rate 0.005 and momentum 0.95, without warmup, clipping, weight decay, Nesterov acceleration or shape-dependent scaling. The momentum query is Frobenius-normalized before five iterations of the polynomial with coefficients $(3.4445, -4.7750, 2.0315)$, and the source records its zero-input safeguard. Parameter and momentum storage and the direction calculation are FP32, forward/backward activations use bfloat16 and TF32 is disabled. The common AdamW side optimizer has learning rate 3×10^{-4}, coefficients $(0.9, 0.95)$ and zero weight decay. PyTorch is 2.8.0 with CUDA runtime 12.6. In particular, $N = 1{,}024$ and $\beta = 0.95$ give $N(1-\beta) = 51.2$, outside the momentum condition $N(1-\beta) \leq 1$ in Theorem 1. Finite-depth directions do not certify a uniform approximation bound along the trajectory. Appendix D.2 gives the exact frozen-noise relation at fixed momentum.

G.2 Pilot selection, paired seeds, and complete-cycle controls

The pilot evaluates learning rates $\{0.0025, 0.005, 0.01\}$ and momenta $\{0.95, 0.99, 0.9995\}$ on seeds 101 and 102 for four conditions, namely RR, RR with replication factors $c = 4$ and $c = 8$, and IID, which gives 72 complete runs. For a replication factor c, each fixed component is repeated c times, and each cycle independently reshuffles the entire replicated list. Its actual cycle has $1,024c$ updates and is not a concatenation of c independent short permutations. Over the fixed budget, RR with $c = 1, 4, 8$ completes $8, 2, 1$ cycles, respectively, and every training row is used exactly eight times. IID uses the same number of draws from the same fixed component set but need not visit each row eight times. A single IID control is used, not three copies selected against different RR conditions.

Within each candidate, we first average training AUC over the two pilot seeds separately for each condition, then give weight one half to IID and one half to the average of the three RR conditions. This prevents the three RR conditions from receiving three quarters of the selection weight. The score is minimized, never the RR–IID margin maximized. Table 3 lists every candidate. All 72 runs completed, and any numerical failure would have disqualified the common candidate. The selected pair $(0.005, 0.95)$ also minimizes the individual pilot AUC of each of the four methods on this grid. The common-setting and individually selected comparisons therefore use the same confirmation runs, not two independent confirmations.

Five disjoint confirmation seeds, 201–205, each run all four conditions from the same initial weights, giving 20 complete runs. These seeds do not change the prepared pool or the fixed minibatch membership, only initialization and sampling randomness. The pilot's two seeds are not added to the confirmation sample size. All settings, failures and non-favorable comparisons are retained. In particular, in the pilot's $(0.01, 0.9995)$ condition, RR $c = 1$ loses to IID at the final training checkpoint for seed 101, and RR $c = 8$ has the lowest two-seed final mean. This is not removed from the records.

G.3 METRICS, UNCERTAINTY, AND TIMING

Every method evaluates the entire fixed training pool with equal token weight and the entire validation split at updates $0, 1024, \ldots, 8192$. The primary metric is trapezoidal training CE AUC divided by 8,192, and it includes the common initialization, which is omitted only from the panels of the main figure and of Figure 4. Per-update sampled-minibatch losses are retained but are not substituted for the common training objective. At each noninitial RR $c = 1$ checkpoint, every component has been used once in the preceding 1,024 updates, whereas IID has no such coverage constraint. Held-out evaluation has no corresponding visit structure.

The two predeclared primary contrasts compare RR $c = 1$ with IID and with RR $c = 8$, and all differences below are reported as the comparator minus RR $c = 1$, so positive values favor RR $c = 1$. We resample five paired seed differences, with 20,000 bootstrap replicates and fixed bootstrap seed 3811. Each primary contrast has a 97.5% percentile interval (Bonferroni allocation for the two contrasts), and secondary contrasts use pointwise 95% intervals. With five pairs these are approximate small-sample uncertainty descriptions rather than exact coverage guarantees. Error bands in figures are sample standard deviations rather than confidence intervals, and descriptive means across settings are not independent seeds. Table 4 reports every confirmation condition, and Figure 3 restores the full loss range including initialization.

RR $c = 1$ has lower AUC, final training, validation and test CE than IID in all five pairs. The corresponding mean differences (IID minus RR $c = 1$) are 0.03833, 0.12400, 0.46024 and 0.45794. The final test difference has pointwise 95% bootstrap interval $[0.42238, 0.49883]$, and test is secondary and never used to tune the run. The other primary AUC contrast, RR $c = 8$ minus RR $c = 1$, is 0.02899, with 97.5% interval $[0.02493, 0.03200]$. However, the RR $c = 8$ minus RR $c = 4$ AUC interval $[-0.00201, 0.00873]$ includes zero. The four mean AUCs and final losses in Table 4 increase from $c = 1$ to $c = 4$, $c = 8$ and IID, but the $c = 4$ and $c = 8$ AUCs are not statistically resolved. Because each RR condition uses every row eight times, their final losses differ despite equal total visit counts. These are finite-budget control results, not a monotonic-performance guarantee from the N-dependence of Eq. (8).

A target CE of 4.29395, fixed before confirmation at 80% of the selected pilot's family-balanced initial-to-final loss improvement, is first met by all 20 runs at the same observed update 2,048, without interpolation between evaluation points, and the recorded training-only times to that update (about 1,000 seconds) and to the final budget (about 4,030 seconds) agree across samplers within

Table 3: **All pilot candidates.** Two seeds per condition, and AUC uses all nine training-pool checkpoints. The family-balanced score determines the common setting and is not the mean of four equally weighted conditions. The selected learning rate and momentum are **0.005 and 0.95**.

Learning rate	Momentum	RR $c = 1$	RR $c = 4$	RR $c = 8$	IID	Selection score
0.0025	0.95	4.05334	4.07139	4.07395	4.08319	4.07471
0.0025	0.99	4.17990	4.20774	4.21399	4.22290	4.21172
0.0025	0.9995	4.66256	4.73064	4.74359	4.75912	4.73569
0.005	0.95	4.03230	4.06175	4.06361	4.07080	4.06167
0.005	0.99	4.22786	4.26822	4.28067	4.29132	4.27512
0.005	0.9995	5.23249	5.34381	5.36714	5.37291	5.34370
0.01	0.95	4.13697	4.16652	4.17819	4.17711	4.16884
0.01	0.99	4.38711	4.44504	4.45532	4.46503	4.44709
0.01	0.9995	6.95231	6.98334	6.98132	7.00266	6.98749

Table 4: **All four confirmation conditions at the fixed final budget.** Mean $\pm$ sample SD, lower is better. RR with c copies reshuffles the c-times replicated component list, so its cycle has $1{,}024c$ updates. AUC includes initialization, and test uses each final model.

Sampling	Training AUC	Final train CE	Final validation CE	Final test CE
RR ($c = 1$)	4.0333 ± 0.0016	2.5215 ± 0.0058	4.6985 ± 0.0054	4.6772 ± 0.0059
RR, $c = 4$	4.0584 ± 0.0023	2.6026 ± 0.0120	4.9061 ± 0.0276	4.8821 ± 0.0283
RR, $c = 8$	4.0623 ± 0.0053	2.6203 ± 0.0151	5.0186 ± 0.0083	4.9932 ± 0.0082
IID	4.0716 ± 0.0059	2.6455 ± 0.0181	5.1587 ± 0.0471	5.1352 ± 0.0469

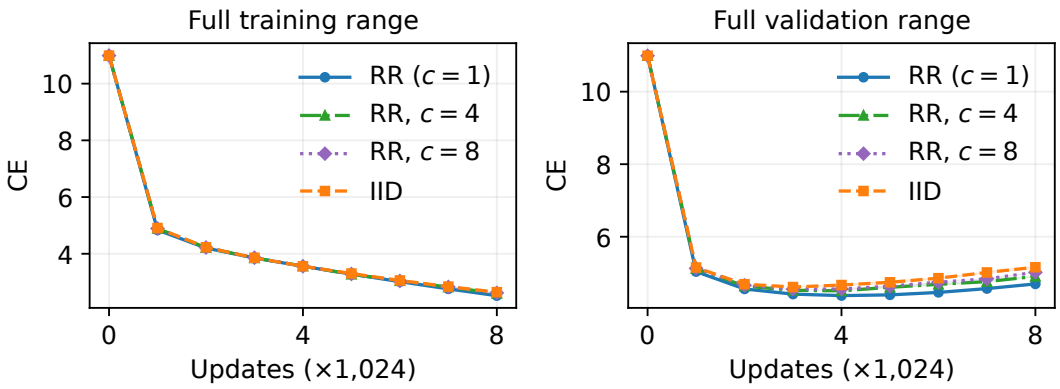


Figure 3: **Full-range confirmation curves for all four conditions.** Markers show five-seed means and shading shows sample SD at all nine checkpoints, including initialization, and straight segments connect observations. The common initial decrease is much larger than the between-sampler differences, shown more clearly by the paired-gap panels in Figures 1 and 4. These panels retain the full loss range, not a magnified effect size.

seed variation for runs executing concurrently on assigned GPUs. AUC and final CE improvements therefore do not imply a measured time-to-target improvement.

G.4 Recorded validation minima and late-stage degradation

We select each run's lowest validation CE on the recorded grid $0, 1024, \ldots, 8192$, breaking ties toward the earlier update, then average across seeds. Table 5 pairs this minimum with training CE at that update and the subsequent validation increase. This is a descriptive diagnostic, not new training, tuning or test-set selection.

Table 5: **Best recorded validation checkpoints and subsequent increase.** Mean ± sample SD across five seeds. Parentheses in the update column count seeds, not epochs. Training CE is measured at the same selected update as minimum validation CE, and these updates need not match across methods. The last column averages each run's final-minus-minimum validation CE. This is a descriptive selection on the recorded grid rather than an early-stopped test evaluation, since intermediate-checkpoint test CE is unavailable.

Sampling	Selected update (seed count)	Minimum validation CE	Training CE at that update	Final validation CE	Final minus minimum val CE
RR ($c = 1$)	4,096 (5)	4.3765 ± 0.0068	3.5589 ± 0.0039	4.6985 ± 0.0054	0.3220 ± 0.0039
RR, $c = 4$	4,096 (5) 3,072 (3)	4.5084 ± 0.0065	3.5665 ± 0.0081	4.9061 ± 0.0276	0.3977 ± 0.0264
RR, $c = 8$	4,096 (2)	4.5494 ± 0.0071	3.7444 ± 0.1654	5.0186 ± 0.0083	0.4692 ± 0.0138
IID	3,072 (5)	4.6131 ± 0.0083	3.8574 ± 0.0130	5.1587 ± 0.0471	0.5457 ± 0.0495

RR $c = 1$ has a lower recorded validation minimum in all five pairs, but reaches it later than IID. Validation subsequently rises while training CE falls, so final-budget gaps include different amounts of late-stage overfitting. The common-target timing result is reported separately in Appendix G.3.

Figure 4 preserves all per-seed reversals. RR has lower training CE in four of five pairs at update 3,072, three at 4,096, and four at 5,120, and at 4,096 the mean paired gap is only 0.00035 (SD 0.01212). Validation gaps are positive in every pair at all eight post-initialization checkpoints, but are not monotone for every seed. The gap panels in Figure 1 use the sample SD of paired differences, and its absolute-loss panel uses the sample SD within each sampler. The interpretation of these matched-budget gaps follows Appendix G.3.

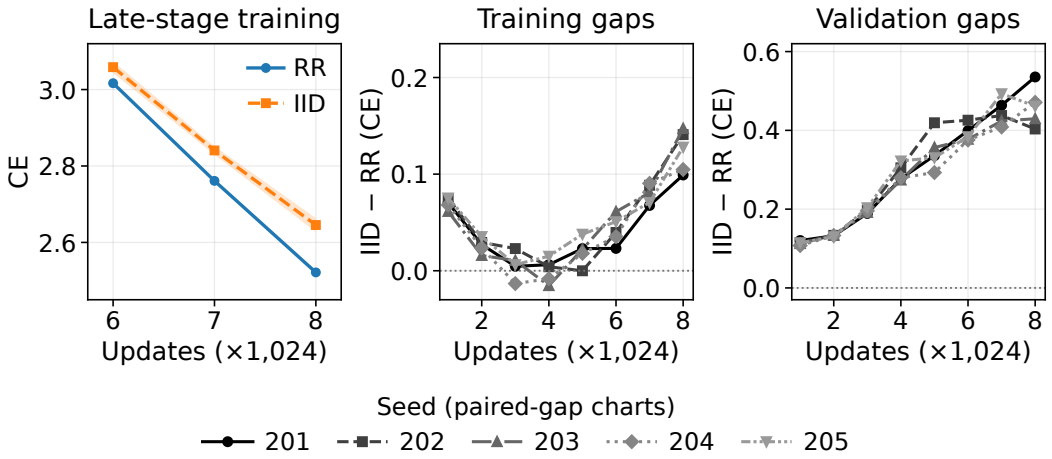


Figure 4: **Late-stage losses and seedwise paired differences.** The same confirmation records as Figure 1 are shown in detail. The left panel shows late-stage training CE as mean ± sample SD. The middle and right panels show all five paired IID-minus-RR training and validation gaps at the eight post-initialization checkpoints. Positive gaps favor RR, and negative gaps are retained. No smoothing, curve normalization or new measurements are used. The near-zero mean training gap at update 4,096 is not uniform per-seed dominance, whereas every recorded validation gap is positive.

H Synthetic Check of the Buffer-Tracking Mechanism

Purpose and setup. This fixed-N check complements the language-model study by measuring the endpoint buffer error of Lemma 8. We use Eq. (9) with $W \in \mathbb{R}^{16\times 8}$ and four copies of each of the $J = 16$ components, giving $N = 64$. Both families share the centers A_j, with Gaussian entries of standard deviation 0.3, centered across components. The objective seed is fixed within each family. The run seed generates standard-normal Z_{uv} and initial entries $2 + 0.5Z_{uv}$. Paired samplers share W_0, and the families may have different initial gaps. The heterogeneous family has unbounded global gradient spread (Appendix C.2), so its IID control is not covered by a bounded-variance guarantee.

Let (η_0, β_0) be the prescription of Theorem 1 at $K = 128$ for each paired initialization. The 80 runs cover both families, ten paired seeds 201–210, both samplers, and stepsizes η_0 and $100\eta_0$, with β_0 and no buffer resets. IID uses the same settings, not its own tuning. Updates use float64 SVD, without clipping or auxiliary optimizers. Corollary 2 covers the larger step with stationarity constant 200 and stability bound $R_1^2 H$, not $2H$.

Tracking and stationarity. For $T = KN = 8{,}192$, we report the all-pre-update average $\widehat{\mathcal{G}}_T = T^{-1}\sum_{t=1}^{T} \|\nabla f(W_{t-1})\|_*$ and the mean endpoint buffer error $\|\nabla f(W_N^k) - S_N^k\|_*$ over the same 128 reporting windows, where $S_i^k = (1-\beta_0)M_i^k$. Table 6 shows lower endpoint buffer error under RR at both stepsizes. At $100\eta_0$, the buffer error improves by 24.4% and 25.1% and average stationarity by 5.4% and 5.0%. At the literal theorem step η_0, both samplers' average and terminal gradient norms stay near 46, and RR reduces buffer error by 27–41% but average stationarity by only 0.0046–0.0061%. Smaller buffer error therefore need not yield a comparable optimization gain, and these observations do not establish a causal decomposition.

Frozen-noise benchmark. For gradients at a common frozen point, the squared-Frobenius noise coefficients are Q_N under IID and θ_N under RR (Appendix D). Here $\lambda = (1/16)^{1/3}$ and $Q_N/\theta_N = 75.74$. The measured ratios at each sampler's own epoch-start iterates range from 75.53 to 76.44, and the supplementary archive retains all diagnostic definitions and 80 records.

Table 6: **Buffer tracking and stationarity at fixed $N = 64$.** Mean $\pm$ sample SD across ten seeds. Each run's endpoint error is averaged over 128 epochs before averaging across seeds. Reduction is $100(1-\text{mean RR error}/\text{mean IID error})$, not a stationarity gain or an average of per-seed percentages. Terminal gradient norms are retained in the supplementary summary.

		Average stationarity $\widehat{\mathcal{G}}_T$		Endpoint buffer error		Error reduction vs IID (%)
Problem	Step	RR	IID	RR	IID	
Homogeneous	η_0	46.36 ± 0.92	46.36 ± 0.92	0.81 ± 0.02	1.11 ± 0.04	27.0
Heterogeneous	η_0	46.46 ± 0.91	46.47 ± 0.91	0.84 ± 0.02	1.43 ± 0.08	41.1
Homogeneous	$100\eta_0$	18.90 ± 0.30	19.97 ± 0.31	3.93 ± 0.12	5.20 ± 0.13	24.4
Heterogeneous	$100\eta_0$	20.05 ± 0.32	21.10 ± 0.31	3.76 ± 0.13	5.02 ± 0.11	25.1

I LIMITATIONS

Theory and practical updates. Theorem 1 gives a finite-horizon stationarity guarantee for a single matrix with fixed components and exact polar updates. Its momentum and stepsize are selected once for the prescribed horizon and held constant throughout the run. The momentum condition $N(1-\beta) \leq 1$ makes the momentum memory span at least one pass, and it is part of this parameter prescription, not a claimed necessary condition for convergence. Proposition 4 extends the analysis to any fixed Jordan depth in exact arithmetic without a spectral-gap assumption, with a constant stepsize rescaling and a looser rank-dependent bound. The language-model study evaluates shared, pilot-selected settings as a practical comparison rather than a direct test of this prescription. Extending the guarantees to momentum selected independently of the horizon and to mixed-precision, layer-wise training remains future work.