ON THE CONVERGENCE AND DIVERGENCE OF MUON: THE ROLE OF MOMENTUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has emerged as an effective optimizer for training large language models. Previous studies have established momentum-dependent convergence guarantees for Muon. Some obtain convergence by choosing β approaching one as the training horizon grows. Other analyses give guarantees for arbitrary fixed $\beta \in (0, 1)$, but often require growing batch sizes or additional conditions such as noise that vanishes with the gradient norm. These results do not establish both a convergence threshold independent of the training horizon and small-momentum divergence on the same fixed problem. To our knowledge, we provide the first joint analysis of these two behaviors for Muon within a fixed smooth finite-sum setting. We establish sufficient high-momentum thresholds β_+^{perm} and $\beta_+^{\mathrm{i.i.d.}}$ for best-iterate stationarity bounds under the affine growth condition and diminishing stepsizes: pathwise for arbitrary epochwise permutations and in expectation for i.i.d. uniform sampling. Under the strong growth condition, these bounds tend to zero as the horizon grows. We also establish a sufficient low-momentum threshold β_- and a smooth, strongly convex quadratic for which, from suitable initializations, Muon diverges to infinity for all $\beta \in [0, \beta_-]$: on every permutation path and on an event of arbitrarily high probability under i.i.d. sampling. Within each sampling scheme, the high-momentum guarantees apply to the same quadratic with the initialization and stepsize schedule unchanged. Together, these results show that changing momentum alone can separate triple divergence from vanishing stationarity bounds on the same fixed problem.

1 INTRODUCTION

Muon has emerged as a popular and effective optimizer for training large language models (Jordan et al., 2024; Liu et al., 2025). Recent studies report improvements in pretraining efficiency across several model scales and training settings (Wen et al., 2026; Semenov et al., 2025). These empirical successes motivate a closer examination of the conditions under which Muon converges and the choices that can cause it to fail. In particular, momentum is a central part of Muon, making its effect on convergence important for understanding the algorithm.

Previous studies have established momentum-dependent convergence guarantees for Muon. Some obtain vanishing error by choosing the momentum coefficient β closer to one as the prescribed training horizon increases (Shen et al., 2026; Sato et al., 2025). Other analyses give bounds for arbitrary fixed $\beta \in (0, 1)$, although these bounds can retain a nonzero stochastic error term when the batch size is fixed (Pethick et al., 2025; Kovalev, 2025). Convergence guarantees with vanishing error can instead rely on increasing the batch size or on additional conditions under which gradient noise decreases with the gradient norm (Li & Hong, 2025; Nagashima & Iiduka, 2026; Zhang & Lin, 2026). These results explain how momentum and batch size affect convergence guarantees. They do not, however, establish both a convergence threshold independent of the training horizon and small-momentum divergence on the same fixed problem.

This distinction matters because the failure of a sufficient convergence condition does not imply that the algorithm actually diverges. A smaller momentum coefficient may only lead to a weaker error bound, or it may cause a more fundamental failure of optimization. We therefore ask:

> *Can changing only a fixed momentum coefficient yield vanishing best-iterate stationarity bounds or unbounded divergence on the same smooth finite-sum problem, while the batch size, initialization, and stepsize schedule remain unchanged?*

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.

To our knowledge, we provide the first joint analysis of these two behaviors for Muon within a fixed smooth finite-sum setting. We study the usual exponential moving average of component gradients followed by exact orthogonalization. The objective and its finite-sum decomposition are fixed before choosing β. The momentum coefficient β can be selected independently of the stopping horizon. Our results cover arbitrary epochwise permutations (Algorithm 1) and i.i.d. uniform sampling (Algorithm 2), providing complementary convergence and divergence guarantees within each sampling scheme.

We establish high-momentum stationarity guarantees for a class of smooth finite-sum objectives under an affine growth condition. For each sampling scheme, we give an explicit sufficient threshold below one that depends on the problem class and is independent of the training horizon. Every fixed β above the corresponding threshold yields a best-iterate stationarity bound with an explicit residual and a term vanishing with the horizon under the respective schedules (4) and (6). Under the strong growth condition, the stationarity bounds tend to zero. These guarantees hold for every sequence of epochwise permutations and in expectation under i.i.d. uniform sampling.

For sufficiently small momentum, we construct a smooth, strongly convex quadratic whose components share a minimizer, and prove that Muon diverges to infinity from explicitly specified deterministic initializations. A single fixed quadratic gives divergence for every β below an explicit positive threshold. The result holds on every epochwise permutation path and with arbitrarily high probability under i.i.d. sampling. Its stepsize conditions include the schedules used in our convergence guarantees. Thus, even a strongly convex objective with a common component minimizer can fail under small-momentum Muon.

Crucially, this quadratic also satisfies the strong growth condition, so the high-momentum convergence guarantees apply to the same instance. Within each sampling scheme, the objective, component decomposition, batch size, initialization, and stepsize schedule can remain unchanged across the two regimes. For permutations, the same permutation sequence can also be used; for i.i.d. sampling, the sampling law remains unchanged. Changing only the fixed momentum coefficient therefore separates vanishing best-iterate stationarity bounds from unbounded divergence on a common problem. Experiments on the quadratic, MNIST, and CIFAR-10 demonstrate the sensitivity of optimization behavior to momentum. Due to space constraints, we defer the discussion of related work to Appendix A.

2 Preliminaries

2.1 Mathematical Notation

For $X, Y \in \mathbb{R}^{p\times q}$, write $\langle X, Y\rangle_F := \operatorname{tr}(X^\top Y)$ and $r := \min\{p, q\}$. With singular values $\sigma_1(X) \geq \cdots \geq \sigma_r(X) \geq 0$, including zeros, the nuclear, Frobenius, and operator norms are

$$\|X\|_{\mathrm{nuc}} := \sum_{a=1}^{r} \sigma_a(X), \qquad \|X\|_F := \Big(\sum_{a=1}^{p}\sum_{b=1}^{q} X_{ab}^2\Big)^{1/2}, \qquad \|X\|_{\mathrm{op}} := \sigma_1(X).$$

2.2 Background Settings

Finite-sum objective. We consider the finite-sum optimization problem

$$\min_{X\in\mathbb{R}^{p\times q}} f(X) := \frac{1}{n}\sum_{i=1}^{n} f_i(X), \tag{1}$$

where each $f_i : \mathbb{R}^{p\times q} \to \mathbb{R}$ is continuously differentiable, with gradients taken in the Frobenius inner product. We assume f attains a finite minimum $f^\star := \min_X f(X) > -\infty$.

Mini-batch representation. Many applications of the Muon optimizer, such as LLM pretraining (Liu et al., 2025; Zeng et al., 2025; Team et al., 2025), use stochastic mini-batch updates. One empirical instance of our finite-sum model takes each component f_i to be the loss over a fixed batch. Let $\mathcal{D} = \{z_a\}_{a=1}^N$ be a dataset of indexed samples, partitioned into fixed batches $\mathcal{B}_1, \ldots, \mathcal{B}_n$ of size B, so $N = nB$. Setting $f_i(X) = B^{-1}\sum_{z\in\mathcal{B}_i} \ell(X; z)$ gives the empirical objective $f(X) = N^{-1}\sum_{a=1}^N \ell(X; z_a)$. The two sampling schemes below detail the component-selection rules.

Muon update. We analyze Muon with its orthogonalization step given by the exact polar map (Jordan et al., 2024). Let $M = U\Sigma V^\top$ be its compact SVD over positive singular values. We define $\mathrm{Polar}(M) := UV^\top, \mathrm{Polar}(0) := 0$. The algorithm boxes below give the corresponding explicit procedures, with two different sampling schemes.

Algorithm 1 Muon under Arbitrary Epochwise Permutations

Require: Initial iterate X_0, fixed $\beta \in [0, 1)$, stepsize schedule (η_t)

Initialize $M_{-1} \leftarrow 0$
for $k = 0, 1, 2, \ldots$ **do**
  Choose a permutation π_k of $\{1, \ldots, n\}$
  for $j = 0, 1, \ldots, n-1$ **do**
    $t \leftarrow kn + j$
    $i_t \leftarrow \pi_k(j+1)$
    $M_t \leftarrow \beta M_{t-1} + (1-\beta)\nabla f_{i_t}(X_t)$
    $X_{t+1} \leftarrow X_t - \eta_t\ \mathrm{Polar}(M_t)$
  end for
end for

Remark: Algorithm 1 visits every component once per epoch. The choice of π_k may be deterministic, random, or based on the history before that epoch.

Algorithm 2 Muon under i.i.d. Uniform Sampling

Require: Initial iterate X_0, fixed $\beta \in [0, 1)$, stepsize schedule (η_t)

Initialize $M_{-1} \leftarrow 0$
for $t = 0, 1, 2, \ldots$ **do**
  $i_t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{1, \ldots, n\})$
  $M_t \leftarrow \beta M_{t-1} + (1-\beta)\nabla f_{i_t}(X_t)$
  $X_{t+1} \leftarrow X_t - \eta_t\ \mathrm{Polar}(M_t)$
end for

Remark: In contrast to Algorithm 1, no specific epoch is defined in Algorithm 2, as we draw each i_t independently and uniformly with replacement from $\{1, \ldots, n\}$.

2.3 Assumptions on Objective Function

We impose the following mild assumptions for convergence theorems on our finite-sum problem in (1). Theorem 3.1 takes Assumption 2.1 and Theorem 3.2 requires Assumption 2.2.

Assumption 2.1 (Component-wise smoothness)**.** *There exists a constant $0 < L < \infty$ such that, for every $i \in \{1, \ldots, n\}$ and all $X, Y \in \mathbb{R}^{p\times q}$,*

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_{\mathrm{nuc}} \le L\|X - Y\|_{\mathrm{op}}.$$

Assumption 2.2 (Full-objective smoothness)**.** *There exists a constant $0 < L < \infty$ such that, for all $X, Y \in \mathbb{R}^{p\times q}$,*

$$\|\nabla f(X) - \nabla f(Y)\|_{\mathrm{nuc}} \le L\|X - Y\|_{\mathrm{op}}. \tag{2}$$

This nuclear–operator smoothness condition on the full objective is also used by Shen et al. (2026, Assumption 3.2). Averaging the component gradients shows that Assumption 2.1 implies Assumption 2.2 with the same Lipschitz constant L.

Next we introduce another key assumption, the affine growth condition (Zhang et al., 2022), which is applied in both convergence theorems.

Assumption 2.3 (Affine growth condition (AGC))**.** *There exist constants $D_0, D_1 \ge 0$ such that, for every $X \in \mathbb{R}^{p\times q}$,*

$$\sum_{i=1}^{n}\|\nabla f_i(X)\|_F^2 \le D_1\|\nabla f(X)\|_F^2 + D_0. \tag{3}$$

When $D_0 = 0$, Assumption 2.3 reduces to the standard strong growth condition (SGC) (Schmidt & Le Roux, 2013; Vaswani et al., 2019; Shi et al., 2021; Zhang et al., 2022; 2026); when $D_1 = n$, it reduces to the standard bounded-variance condition $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(X) - \nabla f(X)\|_F^2 \le \frac{D_0}{n}$ (Ghadimi & Lan, 2013). Uniform bounded-variance assumptions are also used in Muon analyses (Li & Hong, 2025; Shen et al., 2026). Thus, our assumption is strictly weaker and covers a broader class of stochastic optimization problems. Further discussion is deferred to Appendix A.

3 Main Results and Discussion

A convergence–divergence phase transition. How does the momentum coefficient β change Muon's behavior on a fixed problem class? Our results give two complementary answers: large momentum yields stationarity bounds throughout the class, while small momentum permits divergence on a single fixed member.

Shared problem class. Write $\mathcal{F}^{n;p,q}_{L,D_0,D_1}$ for the class of finite sums $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ that attain a finite minimum and satisfy Assumptions 2.1 and 2.3. Here n is the number of components (the number of batches when $N = nB$), p, q are the dimensions of the matrix variable X, and $r = \min\{p, q\}$. The constant $L > 0$ is the smoothness constant. $D_0, D_1 \ge 0$ are from the affine growth condition in Assumption 2.3.

We distinguish *problem-class parameters* (n, p, q, L, D_0, D_1) from *runtime parameters* (β, η_0, X_0). With the dataset fixed, batch size B also enters the former through $n = N/B$. We fix the problem class before choosing runtime parameters, modeling the pretraining workflow of fixing data, architecture, and batching before optimizer tuning. These choices are independent of the stopping horizon T.

Convergence. For any $\eta_0 > 0$ and deterministic X_0, use the respective square-root-decay schedules (4) and (6). Theorems 3.1 and 3.2 provide sufficient thresholds $\beta_+^{\mathrm{perm}}(n, r, D_1)$ and $\beta_+^{\mathrm{i.i.d.}}(r, D_1)$, respectively. Every fixed β between the corresponding threshold and 1 yields the stated best-iterate stationarity bounds for every $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$: pathwise under permutations and in expectation under i.i.d. sampling. These bounds vanish when $D_0 = 0$; for $D_0 > 0$, they retain explicit residuals.

Divergence. Under the class conditions of Theorems 3.4 and 3.5, there exists a single quadratic $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$ with the following properties for every stepsize schedule satisfying Assumption 3.3. Under arbitrary epochwise permutations, a suitable diagonal initialization yields triple divergence (8) for all $\beta \in [0, \beta_-(n)]$ on every permutation path (Theorem 3.4). Under i.i.d. sampling, for every $\varepsilon \in (0, 1)$, a suitable deterministic diagonal initialization yields the same divergence for all such β on one event of probability at least $1 - \varepsilon$ (Theorem 3.5). The threshold $\beta_-(n)$ is explicit, strictly positive, and independent of p, q, L, D_0, D_1.

Propositions 3.6 and 3.7 in Section 3.3 establish consistent ordering of these sufficient thresholds. Both square-root schedules satisfy Assumption 3.3, so within each sampling scheme the objective, decomposition, initialization, and stepsize sequence can remain fixed across regimes. Since the common quadratic satisfies the zero-offset condition, changing only β separates vanishing stationarity from triple divergence.

3.1 Large-beta convergence results

We give convergence theorems for both sampling schemes.

Theorem 3.1 (Large-beta convergence result of Algorithm 1)**.** *Let $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$. Use the diminishing stepsize schedule*

$$\eta_t = \frac{\eta_0}{\sqrt{k+1}}, \qquad t = kn + j, \quad 0 \le j < n, \quad \eta_0 > 0. \tag{4}$$

Define $c_{\beta,n} := \frac{n(1-\beta)}{1-\beta^n} - \beta^n$ for $\beta \in [0, 1)$. Let β_+^{perm} be the unique solution in $(0, 1)$ of

$$2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta_+^{\mathrm{perm}},n} = 1.$$

Choose any $\beta \in (\beta_+^{\mathrm{perm}}, 1)$ and run Algorithm 1 on f. With objective f and all parameters fixed, the following bound holds on every sequence of epochwise permutations as $T \to \infty$:

$$\min_{0 \le t < T} \|\nabla f(X_t)\|_{\mathrm{nuc}} \le \frac{2\sqrt{rD_0}\,c_{\beta,n}}{1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}} + O\left(\frac{\log T}{\sqrt{T}}\right). \tag{5}$$

When $D_0 = 0$, the residual vanishes and the bound is $O(\log T/\sqrt{T}) \to 0$.

Section 4 explains the proof idea of Theorem 3.1. Appendix C provides the complete proof.

Remark:

1. The defining equation for β_+^{perm} has a solution, this solution is unique, and $0 < \beta_+^{\mathrm{perm}} < 1$. See Lemma C.10 for proof.
2. $c_{\beta,n}$ bounds the component-gradient discrepancy remaining after frozen epochwise cancellation: it combines the current-epoch prefix contribution with the unequal momentum weights on past complete epochs; see (40).

Theorem 3.2 (Large-beta convergence result of Algorithm 2)**.** *Let $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$. Use the diminishing stepsize schedule*

$$\eta_t = \frac{\eta_0}{\sqrt{t+1}}, \qquad \eta_0 > 0. \tag{6}$$

Let $C_D \geq 1$ be a fixed universal constant in the Hilbert-space Davis inequality (56) (Wollgast & Zorin-Kranich, 2023). Let $\beta_+^{\mathrm{i.i.d.}}$ be the unique solution in $(0,1)$ of

$$2C_D\sqrt{r}\,(\sqrt{D_1}+1)\sqrt{\frac{1-\beta_+^{\mathrm{i.i.d.}}}{1+\beta_+^{\mathrm{i.i.d.}}}} = 1.$$

Choose any $\beta \in (\beta_+^{\mathrm{i.i.d.}}, 1)$ and run Algorithm 2 on f. With objective f and all parameters fixed, the following bound holds in expectation over the i.i.d. samples as $T \to \infty$:

$$\min_{0\leq t<T} \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} \leq \frac{2C_D\sqrt{rD_0}\sqrt{\frac{1-\beta}{1+\beta}}}{1-2C_D\sqrt{r}\,(\sqrt{D_1}+1)\sqrt{\frac{1-\beta}{1+\beta}}} + O\left(\frac{\log T}{\sqrt{T}}\right). \tag{7}$$

When $D_0 = 0$, the residual vanishes and the bound is $O(\log T/\sqrt{T}) \to 0$.

Appendix I explains the proof idea of Theorem 3.2. Appendix D provides the complete proof.

Remark:

1. In fact, Theorem 3.2 also holds for a weaker condition: only Assumption 2.2 is needed, not Assumption 2.1.
2. The constant C_D comes from the Hilbert-space Davis inequality (56),not an algorithmic tuning parameter.

The permutation threshold β_+^{perm} depends on n, r, D_1, while the i.i.d. threshold $\beta_+^{\mathrm{i.i.d.}}$ depends on r, D_1. The $O(\log T/\sqrt{T})$ constants are uniform in T and, for permutations, in the permutation path. For fixed problem parameters with $D_0 > 0$, the red residual terms in (5) and (7) strictly decrease to zero as $\beta \uparrow 1$ within their respective convergence intervals.

3.2 Small-beta divergence results

Before stating the theorems, we specify the stepsize conditions that are shared by both divergence results. Instead of requiring a specific diminishing stepsize schedule, the stepsize requirements for divergence are more general.

Assumption 3.3 (Stepsizes for divergence)**.** *The prescribed deterministic stepsize sequence satisfies*

$$\eta_t > 0, \qquad \eta_{t+1} \leq \eta_t, \qquad \sum_{t=0}^{\infty} \eta_t = \infty.$$

Now we give divergence theorems for both sampling schemes.

Theorem 3.4 (Small-beta divergence result of Algorithm 1)**.** *For any problem class $\mathcal{F}^{n;p,q}_{L,D_0,D_1}$ with $r \geq n \geq 8$ and $D_1 \geq 16n^2 + 2n - 1$, there exists a function $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$ such that, if Algorithm 1 is run with*

(1) any fixed $\beta \in [0, \beta_-]$, where $\beta_- := (70n)^{-1/\lfloor n/8 \rfloor}$;
(2) any stepsize sequence satisfying Assumption 3.3;
(3) a suitable initial iterate $X_0 \in \mathbb{R}^{p\times q}$, which may depend only on η_0,

then it exhibits triple divergence *for every epochwise permutation sequence:*

$$\|X_t\|_F \longrightarrow \infty, \qquad f(X_t) \longrightarrow \infty, \qquad \|\nabla f(X_t)\|_F \longrightarrow \infty \quad (t \to \infty). \tag{8}$$

Section 5 explains the proof idea of Theorem 3.4. Appendix E provides the complete proof.

Theorem 3.5 (Small-beta divergence result of Algorithm 2)**.** *Under the class conditions of Theorem 3.4, the same fixed quadratic witness in* (10) *exhibits the triple divergence in* (8) *under Algorithm 2: for every $\varepsilon \in (0,1)$ and every admissible deterministic stepsize schedule, a suitable deterministic initial iterate and one event of probability at least $1-\varepsilon$ give divergence simultaneously for all $\beta \in [0, \beta_-]$. The full statement and proof are in Appendix F.*

Appendix J explains the proof idea of Theorem 3.5.

We construct a specific function $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$ to prove both divergence theorems.

A quadratic witness. Fix the class parameters in Theorems 3.4 and 3.5, and set

$$c := \frac{L}{(5n-1)^2 + nr}. \tag{9}$$

For $X \in \mathbb{R}^{p\times q}$, define

$$\begin{aligned} f_i(X) &:= \frac{c}{2}\left((4n+1)X_{ii} - \sum_{a=1}^{n} X_{aa}\right)^2 + \frac{cn}{2}\|X\|_F^2, \qquad i = 1,\dots,n, \\ f(X) &:= \frac{1}{n}\sum_{i=1}^{n} f_i(X). \end{aligned} \tag{10}$$

In both sampling schemes, initialize X_0 with its first n diagonal entries equal to s_0 and all other entries zero, and use stepsizes satisfying Assumption 3.3.

Choose s_0 as follows: (1) If we use permutation sampling (Theorem 3.4), set $s_0 \geq 8\lfloor n/8 \rfloor \eta_0$; (2) If we use i.i.d. sampling (Theorem 3.5), then for any $\varepsilon \in (0,1)$, take $s_0 \geq 36\eta_0 \lfloor n/8 \rfloor^2 \log(n/\varepsilon)$. Section 6.1 illustrates the finite-horizon momentum dependence of this same quadratic under random reshuffling and i.i.d. sampling.

The witness (10) belongs to $\mathcal{F}^{n;p,q}_{L,D_0,D_1}$ (Lemma G.3) and is strongly convex, coercive, and of least-squares form (Lemma G.1). Thus its small-momentum divergence occurs despite favorable objective geometry. The class membership and structural properties are proved in Appendix G.

3.3 Discussion of phase transition momentum thresholds

The following propositions establish the strict separation of the sufficient low- and high-momentum thresholds defined above.

Proposition 3.6 (Separation of the permutation sufficient thresholds)**.** *For $n \geq 8$, $r \geq 1$, and $D_1 \geq 0$,*

$$0 < \beta_-(n) < \beta_+^{\mathrm{perm}}(n, r, D_1) < 1.$$

Proposition 3.7 (Separation of the i.i.d. sufficient thresholds)**.** *For $n \geq 8$, $r \geq n$, and $D_1 \geq 0$,*

$$0 < \beta_-(n) < \beta_+^{\mathrm{i.i.d.}}(r, D_1) < 1.$$

Remark 3.8 (Dependence on the number of components and batch size). The sufficient momentum thresholds are independent of the training horizon T. The thresholds $\beta_-(n)$ and $\beta_+^{\mathrm{perm}}(n, r, D_1)$ depend on the number of components n, whereas $\beta_+^{\mathrm{i.i.d.}}(r, D_1)$ has no explicit dependence on n at fixed r, D_1. For a fixed dataset, $n = N/B$; changes in batching can also change D_1.

4 Permutation Convergence: Proof Idea

We only focus on the proof insight for Theorem 3.1 under arbitrary epochwise permutations. The insight for Theorem 3.2 under i.i.d. uniform sampling is similar.

To prove Theorem 3.1, we first bound the weighted average of gradient norms:

$$\min_{0\le t<T}\|\nabla f(X_t)\|_{\mathrm{nuc}} \le \frac{\sum_{t=0}^{T-1}\eta_t\|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1}\eta_t}.$$

For each iteration,

$$\begin{aligned}\eta_t\|\nabla f(X_t)\|_{\mathrm{nuc}} &\overset{(a)}{\le} \eta_t\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F + 2\eta_t\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}}\\ &\overset{(b)}{\le} f(X_t)-f(X_{t+1})+\frac{L}{2}\eta_t^2+2\eta_t\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}}.\end{aligned}$$

Summing and using $f(X_T)\ge f_\star$, with $\Delta := f(X_0)-f_\star$, gives

$$\sum_{t=0}^{T-1}\eta_t\|\nabla f(X_t)\|_{\mathrm{nuc}} \le \Delta+\frac{L}{2}\sum_{t=0}^{T-1}\eta_t^2+2\underbrace{\sum_{t=0}^{T-1}\eta_t\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}}}_{\text{momentum tracking}}.$$

Here (a) uses Lemma 4.1 below, while (b) follows from the descent lemma (Lemma B.5) and $\|\mathrm{Polar}(M_t)\|_{\mathrm{op}}\le 1$. Thus the proof reduces to controlling the momentum-tracking error.

Lemma 4.1. *For every $t\ge 0$,*

$$\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F \ge \|\nabla f(X_t)\|_{\mathrm{nuc}} - 2\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}}. \tag{11}$$

Proof. Split the full gradient into M_t and its difference from M_t. Lemma B.3 identifies the momentum inner product with its nuclear norm; the nuclear–operator Hölder inequality (Lemma B.2) (Recht et al., 2010) bounds the remaining inner product. Using the polar norm bound and then the triangle inequality gives

$$\begin{aligned}\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F &= \left\langle M_t, \mathrm{Polar}(M_t)\right\rangle_F + \left\langle \nabla f(X_t)-M_t, \mathrm{Polar}(M_t)\right\rangle_F\\ &\ge \|M_t\|_{\mathrm{nuc}} - \|\nabla f(X_t)-M_t\|_{\mathrm{nuc}}\|\mathrm{Polar}(M_t)\|_{\mathrm{op}}\\ &\ge \|M_t\|_{\mathrm{nuc}} - \|\nabla f(X_t)-M_t\|_{\mathrm{nuc}}\\ &\ge \big(\|\nabla f(X_t)\|_{\mathrm{nuc}} - \|\nabla f(X_t)-M_t\|_{\mathrm{nuc}}\big) - \|\nabla f(X_t)-M_t\|_{\mathrm{nuc}}\\ &= \|\nabla f(X_t)\|_{\mathrm{nuc}} - 2\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}}.\end{aligned}$$

□

Unrolling the momentum recursion and applying the triangle inequality (Lemma C.1) gives

$$\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}} \le \underbrace{\beta^{t+1}\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}}}_{\text{initialization error}} + \underbrace{\left\|\sum_{s=1}^{t}\beta^{t-s+1}(\nabla f(X_{s-1})-\nabla f(X_s))\right\|_{\mathrm{nuc}}}_{\text{gradient lag}} + \underbrace{(1-\beta)\left\|\sum_{s=0}^{t}\beta^{t-s}(\nabla f_{i_s}(X_s)-\nabla f(X_s))\right\|_{\mathrm{nuc}}}_{\text{component discrepancy}}.$$

The initialization error decays geometrically. Full-gradient smoothness and $\|X_{s+1}-X_s\|_{\mathrm{op}}\le \eta_s$ give an $O(\eta_t)$ gradient lag. To control the component discrepancy, freeze each epoch at its starting iterate. The centered frozen gradients sum to zero over every complete permutation, so only the differences between their momentum weights remain. Affine growth controls these terms; the unfinished epoch is bounded directly, and component smoothness controls the error from moving evaluation points. Combining the three estimates gives (Lemma C.9)

$$\|M_t-\nabla f(X_t)\|_{\mathrm{nuc}} \le \beta^{t+1}\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}} + \sqrt{r}(\sqrt{D_1}+1)c_{\beta,n}\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{rD_0}\,c_{\beta,n} + K\eta_t,$$

where K is independent of t and T. For $\beta>\beta_+^{\mathrm{perm}}$, the gradient coefficient satisfies $2\sqrt{r}(\sqrt{D_1}+1)c_{\beta,n}<1$. Substitution into the accumulated bound therefore allows this gradient-dependent term to be absorbed on the left. Dividing by $\sum_{t<T}\eta_t$ yields Theorem 3.1, since $\sum_{t<T}\eta_t=\Theta(\sqrt{T})$, $\sum_{t<T}\eta_t^2=O(\log T)$, and $\sum_{t<T}\eta_t\beta^{t+1}=O(1)$. The additive tracking term gives the stated residual, which vanishes when $D_0=0$. The complete proof is in Appendix C.

5 Construction of divergence example

The quadratic in (10) exploits a mismatch between gradient magnitude and update frequency. Every component is strongly convex and minimized at zero. Starting with the first n diagonal entries equal to $s_0>0$ and all other entries zero, Muon keeps those other entries zero. We follow $x_{t,a}$, the ath diagonal entry of X_t, where t is the global update index and $a\in\{1,\dots,n\}$. Figure 1 illustrates the three steps of the argument.

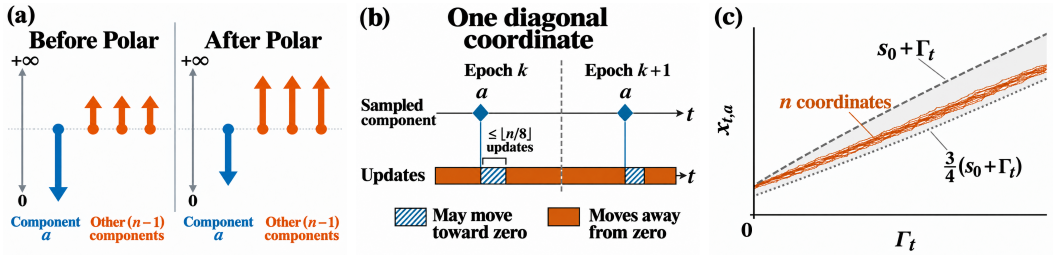

Figure 1: **Mechanism of the divergence construction.** All panels are schematic and not to scale. (a) At an equal-coordinate point with $\beta = 0$, normalization equalizes the magnitudes of the inward and outward coordinate updates. The three orange arrows represent individual directions from the other $n - 1$ components. (b) Each epoch selects component a once. Under the proof's low-momentum and growth-range conditions, each selection permits at most $\lfloor n/8 \rfloor$ potentially inward updates, including the selecting update; outside these intervals, updates move coordinate a outward. (c) The n coordinates remain between the proved bounds $\frac{3}{4}(s_0 + \Gamma_t)$ and $s_0 + \Gamma_t$, where $\Gamma_t := \sum_{s<t} \eta_s$ is the cumulative stepsize.

One large pull, many small pushes. At the initial point, component a gives a large gradient that pulls this coordinate toward zero; each of the other $n - 1$ components gives a smaller gradient that pushes it outward. The large pull outweighs all the small pushes combined, so the full-gradient descent direction points inward. With $\beta = 0$, however, Muon replaces each nonzero diagonal gradient entry by its sign: each of these n coordinate updates has magnitude η_t. Normalization removes the large pull's advantage, leaving one inward direction against $n - 1$ outward directions.

Small momentum cannot sustain the pull. Muon averages gradients before normalization, so a large correcting input can continue to pull the coordinate back for several updates. While the trajectory remains in the range below and $\beta \le \beta_-$, each correcting selection can block outward movement for at most $\lfloor n/8 \rfloor$ updates, including itself. Once component a has been absent for $\lfloor n/8 \rfloor$ consecutive selections, its coordinate moves outward. Each epoch selects a exactly once. Even when selections cluster across epoch boundaries, the nonincreasing stepsizes ensure that the total stepsize weight of potentially inward updates is small; outward movement wins.

The same imbalance persists as the iterate moves. The gradient signs above persist when all n coordinates lie in the range shown below. With the prescribed initialization $s_0 \ge 8\lfloor n/8 \rfloor \eta_0$, counting outward and potentially inward updates forces every coordinate to remain in that range:

$$\frac{3}{4}\left(s_0 + \sum_{s<t} \eta_s\right) \le x_{t,a} \le s_0 + \sum_{s<t} \eta_s, \qquad a = 1, \dots, n.$$

Thus no coordinate can be the first to leave the region supporting the argument. Since the stepsizes have an infinite sum, all these coordinates escape to infinity, although individual updates may move inward. Strong convexity then gives divergence of the objective and full-gradient norm as well. Appendix E proves these claims for every permutation order.

6 Experiments

6.1 Quadratic divergence example

Figure 2 evaluates (10) with $n = r = 8$, five sampling paths, and 10^6 updates. The β scans use schedules (4) and (6), with $\eta_0 = 1$ and $s_0 = 8$ (reshuffling) or $36 \log(80)$ (i.i.d., $\varepsilon = 0.1$). Within each scheme, low momentum produces sustained loss growth, while larger momentum substantially reduces loss. Further results are in Appendix K.

6.2 Momentum sensitivity in neural-network training

MNIST (LeCun et al., 1998). We train a 196–32–32–10 tanh MLP on 2,560 balanced images with Muon using SVD, $B = 32$, and fixed single-class batches in digit order. We use three seeds, 120 epochs, and $\eta_t = 0.03/\sqrt{k+1}$ for $t = kn + j$.

CIFAR-10 (Krizhevsky, 2009). For the upper trajectories, we train ResNet-18 (He et al., 2016) with GroupNorm (Wu & He, 2018) ($B = 8$) and a three-layer CNN of widths 16–32–64 ($B = 128$)

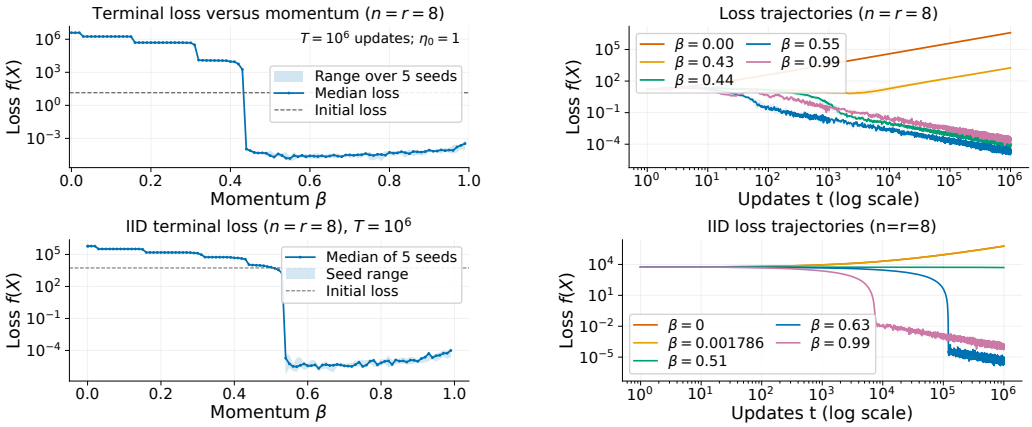

Figure 2: **Momentum-dependent behavior on a fixed quadratic.** Top: random reshuffling; bottom: i.i.d. sampling. Left: terminal loss; right: trajectories. Lines and bands show five-path medians and ranges.

for 50 epochs with the same Muon learning rate schedule. Hidden convolutions use Muon with five Newton–Schulz iterations (ResNet) or SVD (CNN); other parameters use AdamW (Loshchilov & Hutter, 2017). The β sweeps in Figure 3(d–f) show substantially lower training loss at higher momentum.

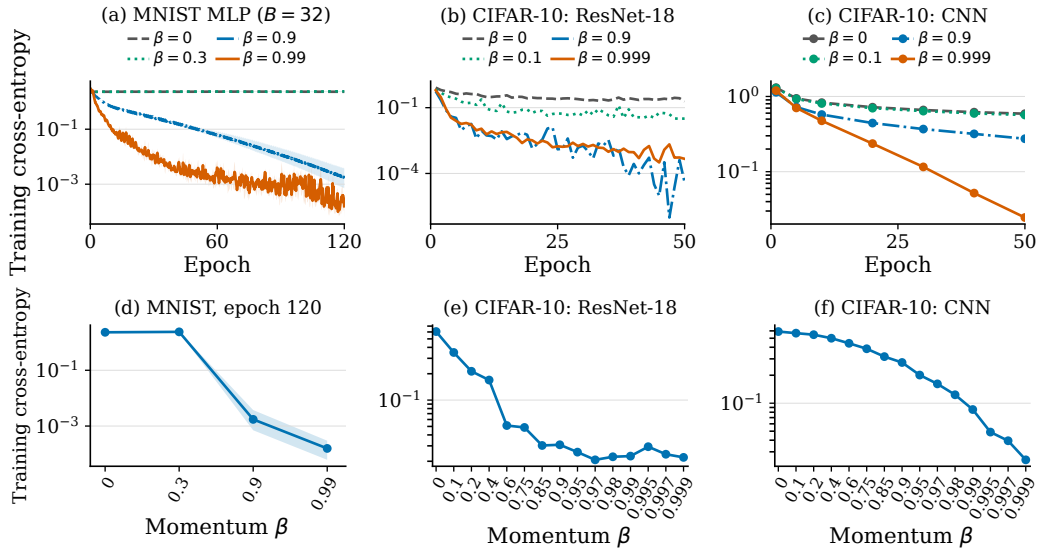

Figure 3: **Momentum-dependent training loss on neural networks.** Top: training loss versus epoch for different fixed β values. Bottom: complete β grids ($4/15/15$ values, at equally spaced categorical positions). We use three seeds. Panel (e) uses $B = 4$, $\eta_0 = 0.003$, and 5 epochs; other settings match the upper panels.

7 Conclusion and Limitations

We establish horizon-independent sufficient momentum regimes for exact-polar Muon: high momentum yields stationarity guarantees, while low momentum causes triple divergence on a smooth strongly convex quadratic. On the same problem, changing only momentum separates these behaviors under both sampling schemes. Our analysis considers a fixed finite-sum decomposition under arbitrary epochwise permutations and i.i.d. uniform sampling; extending the results to non-uniform sampling with replacement remains open. The present study focuses on optimization behavior; its relationship with generalization performance remains to be investigated.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the readability of the manuscript, literature retrieval and discovery of potentially relevant prior work, and feedback during the development and presentation of the theoretical analysis. They were also used to assist with checking mathematical arguments and consistency, refining explanations of proof ideas, and discussing the design and interpretation of supporting experiments. All mathematical claims, proofs, experimental results, and AI-suggested references were independently checked and reviewed by the authors. The authors take full responsibility for the methodology, analysis, experiments, and final content of the paper.

REFERENCES

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.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://arxiv.org/abs/1808.03204.

Florian Hübler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for muon: A theoretical justification of empirical success. *arXiv preprint arXiv:2606.14560*, 2026.

Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex h\"{o} lder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

Won-Jun Jang and Si-Hyeon Lee. Accelerating lmo-based optimization via implicit gradient transport. *arXiv preprint arXiv:2605.05577*, 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*, 6(3):4, 2024.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lexiao Lai, Tianyi Lin, and Jiayu Zhang. Nonsmooth optimization via orthogonalized momentum. *arXiv preprint arXiv:2609.13677*, 2026.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://bottou.org/papers/lecun-98h.

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. *arXiv preprint arXiv:2602.11557*, 2026a.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in muon via spectral filtering. *arXiv preprint arXiv:2606.03899*, 2026b.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Bohao Ma, Junyu Zhang, and Chuan He. Schattor: Schatten-family methods for deep learning optimization. *arXiv preprint arXiv:2606.15702*, 2026.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of muon optimizer for non-convex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M Gower. Muon does not converge on convex lipschitz functions. *arXiv preprint arXiv:2605.08980*, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

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

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835. URL https://doi.org/10.1137/070697835.

Abdurakhmon Sadiev, Artavazd Maranjyan, Ivan Ilin, and Peter Richtárik. Ringmaster lmo: Asynchronous linear minimization oracle momentum method. *arXiv preprint arXiv:2605.18174*, 2026.

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

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

Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2505.23737v3. Version 3, revised July 28, 2026.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. Rmsprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

Egor Shulgin, Mohamed Awad, Peter Richtárik, and Eduard Gorbunov. General analysis of lmo-based optimizers: Beyond bounded variance. In *Forty-third International Conference on Machine Learning*, 2026.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *International Conference on Learning Representations*, volume 2026, pp. 4069–4127, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, volume 2026, pp. 144731–144838, 2026.

Dennis Wollgast and Pavel Zorin-Kranich. Weighted davis inequalities for martingale square functions. *Journal of Theoretical Probability*, 36(3):1520–1533, 2023.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.html.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Yubo Zhang and Junhong Lin. On convergence of muon for nonconvex stochastic optimization under generalized smoothness. *TechRxiv*, 2026. doi: 10.36227/techrxiv.177220100.05629800/v1. URL https://doi.org/10.36227/techrxiv.177220100.05629800/v1.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35: 28386–28399, 2022.

Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, and Ruoyu Sun. Adam converges without any modification on update rules. *arXiv preprint arXiv:2603.02092*, 2026.

A Related Work

The role of momentum and batch size. Momentum and mini-batch size jointly control the averaging of stochastic gradients before orthogonalization. Complementary studies analyze momentum through spectral filtering and develop related methods with variance reduction or implicit gradient transport (Li et al., 2026b; Qian et al., 2025; Jang & Lee, 2026). In Muon and related LMO-based methods, vanishing-error guarantees often use one or both of the following parameter choices:

- **Step-varying or horizon-dependent momentum.** Some guarantees vary the momentum coefficient across iterations (Pethick et al., 2025; Sadiev et al., 2026), while others keep it constant within a run but choose it according to the prescribed training horizon (Shen et al., 2026; Sato et al., 2025; Kim & Oh, 2026; Hübler et al., 2026; Zhang & Lin, 2026; Ma et al., 2026; Jang & Lee, 2026; Tang et al., 2026).
- **Step-varying or horizon-dependent batch size.** Other guarantees increase the batch size during training (Nagashima & Iiduka, 2026; Iiduka, 2026) or choose a constant batch size that grows with the prescribed horizon (Li & Hong, 2025). These choices can control stochastic error even when the momentum coefficient remains fixed.

Among results with both quantities fixed independently of the horizon, stationarity bounds based on a uniform noise level can retain an additive residual (Pethick et al., 2025; Kovalev, 2025; Shen et al., 2026; Sato et al., 2025). This residual term can remain in the bound even with diminishing stepsizes (Nagashima & Iiduka, 2026). Other fixed-parameter results establish approximate max-margin behavior under random reshuffling for separable classification losses (Li et al., 2026a). Our positive guarantees apply to a smooth nonconvex finite-sum class under an affine growth condition, without requiring a particular classification loss. With zero additive offset, sufficiently large fixed β yields vanishing stationarity despite nonidentical component gradients; neither β nor the fixed batch size depends on the horizon. The guarantees hold along every admissible epochwise permutation path, with an additional guarantee in expectation under i.i.d. sampling.

Gradient heterogeneity and noise assumptions. Uniformly bounded stochastic-gradient variance is common in Muon analyses (Shen et al., 2026; Pethick et al., 2025; Kovalev, 2025; Sato et al., 2025; Nagashima & Iiduka, 2026). Broader assumptions for Muon and related methods include moment conditions allowing heavy-tailed noise (Hübler et al., 2026; Iiduka, 2026; Sfyraki & Wang, 2025), affine noise growth under generalized smoothness (Zhang & Lin, 2026), and expected smoothness for LMO-based methods (Shulgin et al., 2026). Our affine growth condition allows component-gradient heterogeneity to scale with the full gradient, covering heterogeneous interpolating quadratic losses whose component variance is globally unbounded. Its zero-offset case still allows stochastic gradients to differ, so the vanishing-stationarity result does not require a noiseless oracle.

Affine growth in finite-sum analyses. Schmidt and Le Roux studied a max-component form of the strong growth condition, while Vaswani, Bach, and Schmidt use its expected-squared-norm form and distinguish it from weaker growth conditions for smooth interpolating finite sums (Schmidt & Le Roux, 2013; Vaswani et al., 2019). The finite-sum analyses of Shi, Li, Hong, and Sun and of Zhang and coauthors use the same affine component-gradient inequality as ours, up to normalization of the objective and constants, to analyze RMSprop and Adam (Shi et al., 2021; Zhang et al., 2022; 2026). Compared with a uniform bounded-variance assumption, finite-sum affine growth allows component variance to scale with the squared full-gradient norm and hence to be globally unbounded; compared with SGC, its positive offset permits nonzero component-gradient variability at a full stationary point.

Divergence of Muon. Nonsmooth counterexamples show that momentum does not in general prevent nonconvergence (Parshakova et al., 2026; Lai et al., 2026). For a generalized Newton–Schulz version of Muon, Do et al. (2026) give batch-dependent error bounds and prove nonconvergence on smooth, strongly convex stochastic quadratics even with diminishing stepsizes. Their counterexamples have different sample minimizers and therefore retain gradient noise at the population minimizer. Their negative result states that the expected truncated distance to the population minimizer does not tend to zero (Do et al., 2026, Theorem 1.3); it does not assert divergence to infinity. In

separable classification, Li et al. (2026a) give a zero-momentum SignSGD counterexample that fails to classify all samples correctly along every reshuffling realization, alongside their positive results with momentum. Our construction establishes unbounded divergence of the parameter norm, objective value, and full-gradient norm on a smooth, strongly convex finite sum whose components share a minimizer. This holds throughout a fixed interval of small momentum coefficients and on every admissible epochwise permutation path. Under i.i.d. sampling, the same construction gives triple divergence on a common event of probability at least $1 - \varepsilon$, simultaneously throughout that interval. The same instance also satisfies our high-momentum convergence guarantee for sufficiently large fixed β, with the objective, decomposition, initialization, stepsize sequence, and permutation path unchanged.

B Auxiliary Results

This appendix collects the matrix inequalities and scalar stepsize estimates used throughout the proofs. For rectangular matrices, $r = \min\{p, q\}$.

Lemma B.1 (Frobenius and nuclear norm comparison)**.** *For every $A \in \mathbb{R}^{p\times q}$,*

$$\|A\|_F \le \|A\|_{\mathrm{nuc}} \le \sqrt{r}\,\|A\|_F. \tag{12}$$

Proof. If $A = 0$, all three quantities vanish. Otherwise let $\rho = \mathrm{rank}(A) > 0$ and let $\sigma_1, \dots, \sigma_\rho$ be its positive singular values. The norm definitions give $\|A\|_F = (\sum_i \sigma_i^2)^{1/2}$ and $\|A\|_{\mathrm{nuc}} = \sum_i \sigma_i$. Expanding the square of the sum and then applying Cauchy–Schwarz yields

$$\begin{aligned}\sum_{i=1}^{\rho} \sigma_i^2 \le \sum_{i=1}^{\rho} \sigma_i^2 + 2\sum_{i<j} \sigma_i\sigma_j &= \left(\sum_{i=1}^{\rho} \sigma_i\right)^2 \\ &\le \left(\sum_{i=1}^{\rho} 1^2\right)\left(\sum_{i=1}^{\rho} \sigma_i^2\right) = \rho\sum_{i=1}^{\rho} \sigma_i^2 \le r\sum_{i=1}^{\rho} \sigma_i^2.\end{aligned}$$

Taking square roots proves both comparisons. □

Lemma B.2 (Nuclear–operator Hölder inequality)**.** *For any $A, D \in \mathbb{R}^{p\times q}$,*

$$|\langle A, D\rangle_F| \le \|A\|_{\mathrm{nuc}}\|D\|_{\mathrm{op}}. \tag{13}$$

Proof. The claim is immediate for $A = 0$. For $A \neq 0$, write its compact singular value decomposition as $A = \sum_{i=1}^{\rho} \sigma_i u_i v_i^\top$, where $\sigma_i > 0$ and $\|u_i\|_2 = \|v_i\|_2 = 1$. Linearity of the trace gives $\langle A, D\rangle_F = \sum_i \sigma_i u_i^\top D v_i$. Hence

$$\begin{aligned}|\langle A, D\rangle_F| &= \left|\sum_{i=1}^{\rho} \sigma_i u_i^\top D v_i\right| \\ &\le \sum_{i=1}^{\rho} \sigma_i |u_i^\top D v_i| \\ &\le \sum_{i=1}^{\rho} \sigma_i \|u_i\|_2 \|D v_i\|_2 \\ &\le \sum_{i=1}^{\rho} \sigma_i \|D\|_{\mathrm{op}} \|v_i\|_2 = \|A\|_{\mathrm{nuc}}\|D\|_{\mathrm{op}}.\end{aligned}$$

The second inequality is the Euclidean Cauchy–Schwarz inequality, and the third is the definition of the operator norm. □

Lemma B.3 (Basic properties of the polar factor)**.** *For every $M \in \mathbb{R}^{p\times q}$ under the convention in Section 2,*

$$\|\mathrm{Polar}(M)\|_{\mathrm{op}} \le 1, \tag{14}$$

and

$$\langle M, \mathrm{Polar}(M)\rangle_F = \|M\|_{\mathrm{nuc}}. \tag{15}$$

Proof. If $M = 0$, then $\mathrm{Polar}(0) = 0$, so its operator norm is $0 \le 1$ and its inner product with M is $0 = \|M\|_{\mathrm{nuc}}$. Suppose $M \neq 0$, let $\rho = \mathrm{rank}(M)$, and write its compact SVD as $M = U\Sigma V^\top$, where $U^\top U = V^\top V = I_\rho$. Set $P = UV^\top$. For every $x \in \mathbb{R}^q$,

$$\|Px\|_2^2 = x^\top VU^\top UV^\top x = x^\top VV^\top x = \|V^\top x\|_2^2.$$

Since the columns of V are orthonormal, $VV^\top$ is an orthogonal projection. More explicitly, $x = VV^\top x + (I - VV^\top)x$ is an orthogonal decomposition, so

$$\|x\|_2^2 = \|VV^\top x\|_2^2 + \|(I - VV^\top)x\|_2^2 \ge \|VV^\top x\|_2^2 = \|V^\top x\|_2^2.$$

Thus $\|Px\|_2 \le \|x\|_2$ for every x, proving $\|P\|_{\mathrm{op}} \le 1$. In fact, taking x to be any column of V gives equality. For the inner product, cyclicity of the trace gives

$$\begin{aligned}\langle M, P\rangle_F &= \mathrm{tr}\big((U\Sigma V^\top)^\top UV^\top\big) \\ &= \mathrm{tr}(V\Sigma U^\top UV^\top) = \mathrm{tr}(V\Sigma V^\top) \\ &= \mathrm{tr}(\Sigma V^\top V) = \mathrm{tr}(\Sigma) = \sum_{i=1}^{\rho} \sigma_i(M) = \|M\|_{\mathrm{nuc}}.\end{aligned}$$

Only the ρ nonzero singular directions were used, so the argument includes rank-deficient matrices. The nonzero singular values of P are all one; consequently $\|P\|_{\mathrm{nuc}} = \rho$, whereas the bound by one concerns its operator norm. □

Lemma B.4 (Polar factors of rectangular diagonal matrices and zero extensions)**.** *Let* $\mathrm{sign}(a) = 1$ *for* $a > 0$, $\mathrm{sign}(a) = -1$ *for* $a < 0$, *and* $\mathrm{sign}(0) = 0$. *For* $y \in \mathbb{R}^r$, *let* $\mathrm{Diag}(y) \in \mathbb{R}^{p\times q}$ *have diagonal* y *and all other entries zero. Then*

$$\mathrm{Polar}(\mathrm{Diag}(y)) = \mathrm{Diag}(\mathrm{sign}(y_1), \dots, \mathrm{sign}(y_r)).$$

Moreover, for $A \in \mathbb{R}^{n\times n}$ *and* $p, q \ge n$, *define its rectangular zero extension by*

$$\mathcal{E}_{p,q}(A) := \begin{pmatrix} A & 0_{n\times(q-n)} \\ 0_{(p-n)\times n} & 0_{(p-n)\times(q-n)} \end{pmatrix}.$$

Then

$$\mathrm{Polar}(\mathcal{E}_{p,q}(A)) = \mathcal{E}_{p,q}(\mathrm{Polar}(A)).$$

Empty padding is allowed when $p = n$ *or* $q = n$; *when* $p = q = n$, *the extension is the identity map.*

Proof. For the diagonal identity, let $J = \{a : y_a \neq 0\}$. If J is empty, both sides are zero by the polar convention. Otherwise, let $e_a^{(p)}$ and $e_a^{(q)}$ be the coordinate unit vectors in the respective spaces. Then

$$\mathrm{Diag}(y) = \sum_{a\in J} |y_a| \left(\mathrm{sign}(y_a) e_a^{(p)}\right)(e_a^{(q)})^\top$$

is a compact singular value decomposition, up to reordering the terms: the left vectors and the right vectors are orthonormal, and each displayed singular value is positive. Replacing those singular values by one gives $\sum_{a\in J} \mathrm{sign}(y_a) e_a^{(p)} (e_a^{(q)})^\top$, which is the claimed rectangular diagonal matrix, with zero at every missing coordinate.

For the extension identity, the case $A = 0$ again follows from $\mathrm{Polar}(0) = 0$. Otherwise write a compact SVD $A = U\Sigma V^\top$ using only its positive singular values. Pad U with $p-n$ zero rows and V with $q-n$ zero rows to obtain $\widetilde{U}$ and $\widetilde{V}$. Their columns remain orthonormal, and $\mathcal{E}_{p,q}(A) = \widetilde{U}\Sigma\widetilde{V}^\top$ is a compact SVD of the extended matrix. Therefore

$$\mathrm{Polar}(\mathcal{E}_{p,q}(A)) = \widetilde{U}\widetilde{V}^\top = \mathcal{E}_{p,q}(UV^\top) = \mathcal{E}_{p,q}(\mathrm{Polar}(A)).$$

The prescribed compact-SVD convention therefore preserves zeros outside the embedded block. □

Lemma B.5 (Objective decrease under an operator-norm step)**.** *Let* $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ *be continuously differentiable and suppose that, for some* $L \ge 0$ *and all* X, Y,

$$\|\nabla f(X) - \nabla f(Y)\|_{\mathrm{nuc}} \le L\|X - Y\|_{\mathrm{op}}. \tag{16}$$

Then for all $X, S \in \mathbb{R}^{p\times q}$,

$$f(X + S) \le f(X) + \langle \nabla f(X), S\rangle_F + \frac{L}{2}\|S\|_{\mathrm{op}}^2. \tag{17}$$

Proof. The function $\tau \mapsto f(X + \tau S)$ is continuously differentiable, with derivative $\langle \nabla f(X + \tau S), S \rangle_F$ by the chain rule. The fundamental theorem of calculus therefore gives

$$f(X + S) - f(X) = \int_0^1 \langle \nabla f(X + \tau S), S \rangle_F \, d\tau.$$

Subtracting $\langle \nabla f(X), S \rangle_F = \int_0^1 \langle \nabla f(X), S \rangle_F \, d\tau$, and using (13) and (16), yields

$$\begin{aligned} f(X + S) - f(X) - \langle \nabla f(X), S \rangle_F &= \int_0^1 \langle \nabla f(X + \tau S) - \nabla f(X), S \rangle_F \, d\tau \\ &\le \int_0^1 | \langle \nabla f(X + \tau S) - \nabla f(X), S \rangle_F | \, d\tau \\ &\le \int_0^1 \|\nabla f(X + \tau S) - \nabla f(X)\|_{\mathrm{nuc}} \|S\|_{\mathrm{op}} \, d\tau \\ &\le \int_0^1 L \|\tau S\|_{\mathrm{op}} \|S\|_{\mathrm{op}} \, d\tau \\ &= L \|S\|_{\mathrm{op}}^2 \int_0^1 \tau \, d\tau = \frac{L}{2} \|S\|_{\mathrm{op}}^2. \end{aligned}$$

Here $\|\tau S\|_{\mathrm{op}} = \tau \|S\|_{\mathrm{op}}$ because $0 \le \tau \le 1$. Rearranging proves the claim. □

Lemma B.6 (Stepsize sums for the square-root epoch schedule)**.** *Let $n \ge 1$ be an integer, $\eta_0 > 0$, and $0 < \beta < 1$, and use the stepsize schedule in (4). For every integer $T \ge 1$,*

$$\sum_{t=0}^{T-1} \eta_t \beta^{t+1} \le \frac{\eta_0 \beta}{1 - \beta}. \tag{18}$$

For horizons consisting of complete epochs, $T \in \{n, 2n, 3n, \dots\}$,

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t &\ge 2n\eta_0 \big(\sqrt{T/n + 1} - 1\big), \\ \sum_{t=0}^{T-1} \eta_t^2 &\le n\eta_0^2 \big(1 + \log(T/n)\big). \end{aligned} \tag{19}$$

For arbitrary integer horizons, with n, η_0, β fixed,

$$\sum_{t=0}^{T-1} \eta_t = \Omega(\sqrt{T}), \quad \sum_{t=0}^{T-1} \eta_t^2 = O(\log T), \quad \sum_{t=0}^{T-1} \eta_t \beta^{t+1} = O(1) \qquad (T \to \infty). \tag{20}$$

These estimates concern the prescribed stepsize sequence alone and do not depend on the iterates or permutation sequence.

Proof. Geometrically weighted sum and complete epochs. Since $\eta_t \le \eta_0$ and all summands are nonnegative,

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \beta^{t+1} &\le \eta_0 \sum_{t=0}^{T-1} \beta^{t+1} = \eta_0 \beta \frac{1 - \beta^T}{1 - \beta} \\ &\le \frac{\eta_0 \beta}{1 - \beta}. \end{aligned}$$

For $T \in n\mathbb{N}$, write $m = T/n \ge 1$, the number of complete epochs. Partitioning the sums into these m epochs gives

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t &= \sum_{u=0}^{m-1} \sum_{j=0}^{n-1} \frac{\eta_0}{\sqrt{u+1}} = n\eta_0 \sum_{v=1}^{m} v^{-1/2}, \\ \sum_{t=0}^{T-1} \eta_t^2 &= \sum_{u=0}^{m-1} \sum_{j=0}^{n-1} \frac{\eta_0^2}{u+1} = n\eta_0^2 \sum_{v=1}^{m} v^{-1}. \end{aligned}$$

For $x \in [v, v+1]$, $x^{-1/2} \le v^{-1/2}$, so summing the corresponding integrals gives

$$\sum_{v=1}^{m} v^{-1/2} \ge \sum_{v=1}^{m} \int_{v}^{v+1} x^{-1/2}\, dx = \int_{1}^{m+1} x^{-1/2}\, dx = 2(\sqrt{m+1} - 1).$$

For $v \ge 2$ and $x \in [v-1, v]$, $x^{-1} \ge v^{-1}$. Thus

$$\sum_{v=1}^{m} v^{-1} = 1 + \sum_{v=2}^{m} v^{-1} \le 1 + \int_{1}^{m} x^{-1}\, dx = 1 + \log m,$$

including $m = 1$, when the latter sum and integral are both zero. Multiplying these estimates by $n\eta_0$ and $n\eta_0^2$, respectively, proves (19).

Arbitrary integer horizons. For $T \ge n$, let $m = \lfloor T/n \rfloor$ and $b = T - mn$, so $0 \le b < n$. The first m epochs are complete and the last b updates have stepsize $\eta_0/\sqrt{m+1}$. Therefore,

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t &= n\eta_0 \sum_{v=1}^{m} v^{-1/2} + \frac{b\eta_0}{\sqrt{m+1}} \ge 2n\eta_0(\sqrt{m+1} - 1), \\
\sum_{t=0}^{T-1} \eta_t^2 &= n\eta_0^2 \sum_{v=1}^{m} v^{-1} + \frac{b\eta_0^2}{m+1} \\
&\le n\eta_0^2 \sum_{v=1}^{m+1} v^{-1} \le n\eta_0^2(1 + \log(m+1)).
\end{aligned}$$

For $m \ge 1$, rationalizing the square root gives

$$\sqrt{m+1} - 1 = \frac{m}{\sqrt{m+1}+1} \ge \frac{\sqrt{m}}{\sqrt{2}+1},$$

because $\sqrt{m+1} \le \sqrt{2m}$ and $1 \le \sqrt{m}$. For $T \ge 2n$, also $m = \lfloor T/n \rfloor \ge T/(2n)$. Consequently the sum $\sum_{t<T} \eta_t$ is bounded below by a positive constant times $\sqrt{T}$, while $\sum_{t<T} \eta_t^2 = O(\log T)$ and the weighted sum in (18) is bounded independently of T. These are precisely the three claims in (20). The estimates above depend only on the prescribed sequence, so they are independent of the iterates and the permutation sequence. □

C Permutation Convergence: Proof

This appendix establishes the explicit finite-horizon estimate in Theorem C.12 and then derives Theorem 3.1. Throughout, assume Assumptions 2.1 and 2.3, and fix an arbitrary permutation sequence and the resulting iterates of Algorithm 1. We use the square-root epoch schedule in (4), $\beta \in (\beta_+^{\mathrm{perm}}, 1)$, $n \ge 1$, and $r = \min\{p, q\}$. For a fixed current iteration, write $t = kn + j$, where $k = \lfloor t/n \rfloor$ and $0 \le j < n$. The index $k' < k$ denotes a past complete epoch. When a formula ranges over both past epochs and the current epoch, a local index h is used, with $s = hn + \ell$ and $0 \le \ell < n$. The horizon T counts updates, and T/n counts complete epochs when $T \in n\mathbb{N}$.

The proof first establishes a complete bound on the momentum-tracking error $\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}}$. It then converts this tracking estimate into a lower bound on the gradient inner product with the Muon direction, and finally into objective decrease and weighted-average stationarity.

Write $\Delta := f(X_0) - f^\star$.

C.1 Decomposing the momentum error

Lemma C.1 (Exact decomposition of the momentum-tracking error)**.** *For every $t \geq 0$, the momentum recursion used in Algorithms 1 and 2 satisfies*

$$\begin{aligned} M_t - \nabla f(X_t) = & \underbrace{\beta^{t+1}\big(M_{-1} - \nabla f(X_0)\big)}_{\textit{initialization error}} \\ & + \underbrace{(1-\beta)\sum_{s=0}^{t} \beta^{t-s}\big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big)}_{\textit{component discrepancy}} \\ & + \underbrace{\sum_{s=1}^{t} \beta^{t-s+1}\big(\nabla f(X_{s-1}) - \nabla f(X_s)\big)}_{\textit{full-gradient lag}}. \end{aligned} \tag{21}$$

This identity uses only the momentum recursion and holds for either sampling scheme and any prescribed stepsize sequence. The final sum is zero when $t = 0$, and the initialization term obeys

$$\big\|\beta^{t+1}\big(M_{-1} - \nabla f(X_0)\big)\big\|_{\mathrm{nuc}} = \beta^{t+1}\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}}. \tag{22}$$

Proof. We first unroll the momentum recursion. For $t \geq 1$, the first two substitutions are

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta)\nabla f_{i_t}(X_t) \\ &= \beta\big[\beta M_{t-2} + (1-\beta)\nabla f_{i_{t-1}}(X_{t-1})\big] + (1-\beta)\nabla f_{i_t}(X_t) \\ &= \beta^2 M_{t-2} + (1-\beta)\big(\beta\nabla f_{i_{t-1}}(X_{t-1}) + \nabla f_{i_t}(X_t)\big). \end{aligned}$$

Continuing one substitution at a time, after $u+1$ substitutions, $0 \leq u \leq t$, the same recursion reads

$$M_t = \beta^{u+1} M_{t-u-1} + (1-\beta)\sum_{v=0}^{u} \beta^v \nabla f_{i_{t-v}}(X_{t-v}).$$

Indeed, if this formula holds for some $u < t$, substituting

$$M_{t-u-1} = \beta M_{t-u-2} + (1-\beta)\nabla f_{i_{t-u-1}}(X_{t-u-1})$$

replaces its first term by

$$\beta^{u+2} M_{t-u-2} + (1-\beta)\beta^{u+1}\nabla f_{i_{t-u-1}}(X_{t-u-1}).$$

Appending the new gradient term to the existing sum gives

$$M_t = \beta^{u+2} M_{t-u-2} + (1-\beta)\sum_{v=0}^{u+1} \beta^v \nabla f_{i_{t-v}}(X_{t-v}),$$

which is exactly the displayed formula with u replaced by $u+1$. Setting $u = t$ and reindexing with $s = t - v$ gives

$$M_t = \beta^{t+1} M_{-1} + (1-\beta)\sum_{s=0}^{t} \beta^{t-s}\nabla f_{i_s}(X_s). \tag{23}$$

For $t = 0$, (23) is exactly the defining recursion.

Subtracting $\nabla f(X_t)$ from (23), adding and subtracting the weighted full-gradient sum, and then inserting $\beta^{t+1}\nabla f(X_0)$ gives, one line at a time,

$$\begin{aligned}
M_t - \nabla f(X_t) &= \beta^{t+1} M_{-1} + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \nabla f_{i_s}(X_s) - \nabla f(X_t) \\
&= \beta^{t+1} M_{-1} + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big) \\
&\quad + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \nabla f(X_s) - \nabla f(X_t) \\
&= \beta^{t+1} \big(M_{-1} - \nabla f(X_0)\big) \\
&\quad + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big) \\
&\quad + \beta^{t+1} \nabla f(X_0) + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \nabla f(X_s) - \nabla f(X_t).
\end{aligned}$$

To identify the last line, use $(1-\beta) = 1-\beta$ and reindex the two geometric sums:

$$\begin{aligned}
&\beta^{t+1} \nabla f(X_0) + (1-\beta)\sum_{s=0}^{t} \beta^{t-s} \nabla f(X_s) - \nabla f(X_t) \\
&\quad = \beta^{t+1} \nabla f(X_0) + \sum_{s=0}^{t} \beta^{t-s} \nabla f(X_s) - \sum_{s=0}^{t} \beta^{t-s+1} \nabla f(X_s) - \nabla f(X_t) \\
&\quad = \sum_{s=0}^{t-1} \beta^{t-s} \nabla f(X_s) - \sum_{s=1}^{t} \beta^{t-s+1} \nabla f(X_s) \\
&\quad = \sum_{s=1}^{t} \beta^{t-s+1} \big(\nabla f(X_{s-1}) - \nabla f(X_s)\big).
\end{aligned}$$

This proves (21); (22) follows by positive homogeneity of the nuclear norm, since $\beta \geq 0$. □

Remark. The decomposition assigns one later task to each term. The initialization error is already geometrically damped and equals $\beta^{t+1}\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}}$ after taking its nuclear norm. Lemma C.2 controls the gradient lag, and lemma C.8 controls the component discrepancy. The permutations may be deterministic or history-dependent; each identity holds on the fixed permutation sequence without probabilistic averaging.

C.2 Bounding gradient lag

Lemma C.2 (Accumulated full-gradient lag)**.** *Define the full-gradient lag term in* (21) *by*

$$\mathcal{L}_t := \sum_{s=1}^{t} \beta^{t-s+1} \big(\nabla f(X_{s-1}) - \nabla f(X_s)\big).$$

Then

$$\|\mathcal{L}_t\|_{\mathrm{nuc}} \leq L \left(\sum_{\ell=1}^{\infty} \beta^{\ell} \sqrt{\ell+1} \right) \eta_t, \tag{24}$$

and the series on the right is finite.

Proof. For any step index $s \geq 0$, the update and (14) give

$$\|X_{s+1} - X_s\|_{\mathrm{op}} = \eta_s \|\mathrm{Polar}(M_s)\|_{\mathrm{op}} \leq \eta_s. \tag{25}$$

For $1 \le s \le t$, full-gradient smoothness and (25) give

$$\|\nabla f(X_{s-1}) - \nabla f(X_s)\|_{\mathrm{nuc}} \le L\|X_{s-1} - X_s\|_{\mathrm{op}} \le L\eta_{s-1}.$$

The remaining task is to compare η_{s-1} with η_t. Put $h = \lfloor (s-1)/n \rfloor$ and $\ell = t - s + 1$. For integer v, the increment $\lfloor v/n \rfloor - \lfloor (v-1)/n \rfloor$ is either zero or one. Consequently,

$$0 \le k - h = \sum_{v=s}^{t} \left(\left\lfloor \frac{v}{n} \right\rfloor - \left\lfloor \frac{v-1}{n} \right\rfloor \right) \le \sum_{v=s}^{t} 1 = t - s + 1 = \ell.$$

The schedule now implies

$$\frac{\eta_{s-1}}{\eta_t} = \sqrt{\frac{k+1}{h+1}} = \sqrt{1 + \frac{k-h}{h+1}} \le \sqrt{1 + k - h} \le \sqrt{\ell + 1},$$

where $h + 1 \ge 1$ was used in the first inequality. Taking norms and substituting these estimates yields

$$\begin{aligned}
\|\mathcal{L}_t\|_{\mathrm{nuc}} &\le \sum_{s=1}^{t} \beta^{t-s+1} \|\nabla f(X_{s-1}) - \nabla f(X_s)\|_{\mathrm{nuc}} \\
&\le L \sum_{s=1}^{t} \beta^{t-s+1} \eta_{s-1} \\
&\le L\eta_t \sum_{s=1}^{t} \beta^{t-s+1} \sqrt{t - s + 2} \\
&= L\eta_t \sum_{\ell=1}^{t} \beta^{\ell} \sqrt{\ell + 1} \\
&\le L\eta_t \sum_{\ell=1}^{\infty} \beta^{\ell} \sqrt{\ell + 1}.
\end{aligned}$$

The change of variables reverses the order $s = 1, \dots, t$ to $\ell = t, \dots, 1$. The consecutive-term ratio in the infinite series is $\beta\sqrt{(\ell+2)/(\ell+1)} \to \beta < 1$, so its tail is bounded by a geometric series. When $t = 0$, $\mathcal{L}_0 = 0$ by the empty-sum convention. □

C.3 Bounding component discrepancy

The next two lemmas serve different purposes. The first controls component discrepancies evaluated at one common point. The second controls distances between iterates and ratios between stepsizes. These facts will be used together after the component discrepancy has been split into frozen terms and a moving-point residual.

Lemma C.3 (Component discrepancies at a common point)**.** *For every $X \in \mathbb{R}^{p \times q}$ and every permutation π of $\{1, \dots, n\}$,*

$$\sum_{j=0}^{n-1} \big(\nabla f_{\pi(j+1)}(X) - \nabla f(X) \big) = 0. \tag{26}$$

Under Assumption 2.3, each $i \in \{1, \dots, n\}$ also satisfies

$$\|\nabla f_i(X) - \nabla f(X)\|_{\mathrm{nuc}} \le \sqrt{r}\,(\sqrt{D_1} + 1)\|\nabla f(X)\|_{\mathrm{nuc}} + \sqrt{rD_0}. \tag{27}$$

Proof. For the first statement, differentiating the finite sum and using the bijectivity of π give

$$\begin{aligned}
\sum_{j=0}^{n-1} \big(\nabla f_{\pi(j+1)}(X) - \nabla f(X) \big) &= \sum_{i=1}^{n} \nabla f_i(X) - n\nabla f(X) \\
&= n \left(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(X) - \nabla f(X) \right) = 0.
\end{aligned}$$

This identity uses only the finite-sum structure and a common evaluation point. For the second statement, all squared norms are nonnegative, so

$$\|\nabla f_i(X)\|_F^2 \le \sum_{i'=1}^{n} \|\nabla f_{i'}(X)\|_F^2 \le D_1 \|\nabla f(X)\|_F^2 + D_0.$$

For $x, y \ge 0$, the identity $(\sqrt{x}+\sqrt{y})^2 = x + y + 2\sqrt{xy} \ge x + y$ implies $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$. Taking square roots therefore yields

$$\|\nabla f_i(X)\|_F \le \sqrt{D_1 \|\nabla f(X)\|_F^2 + D_0} \le \sqrt{D_1}\|\nabla f(X)\|_F + \sqrt{D_0}.$$

Combining this with (12) and the triangle inequality gives every step of the required norm conversion:

$$\begin{aligned}
\|\nabla f_i(X) - \nabla f(X)\|_{\mathrm{nuc}} &\le \sqrt{r}\, \|\nabla f_i(X) - \nabla f(X)\|_F \\
&\le \sqrt{r}\, \big(\|\nabla f_i(X)\|_F + \|\nabla f(X)\|_F\big) \\
&\le \sqrt{r}\, \big((\sqrt{D_1}+1)\|\nabla f(X)\|_F + \sqrt{D_0}\big) \\
&\le \sqrt{r}\, (\sqrt{D_1}+1)\|\nabla f(X)\|_{\mathrm{nuc}} + \sqrt{rD_0}.
\end{aligned}$$

□

Lemma C.4 (Iterate and stepsize comparisons at the current iteration)**.** *Fix $t = kn + j$, where $0 \le j < n$. The current epoch's starting point satisfies*

$$\|X_t - X_{kn}\|_{\mathrm{op}} \le j\eta_t \le (n-1)\eta_t. \tag{28}$$

For every past complete epoch $0 \le k' < k$,

$$\frac{\eta_{k'n}}{\eta_t} = \sqrt{\frac{k+1}{k'+1}} \le \sqrt{k-k'+1}, \tag{29}$$

and

$$\|X_t - X_{k'n}\|_{\mathrm{op}} \le n(k-k'+1)^{3/2}\eta_t. \tag{30}$$

Proof. Using (25), $X_t - X_{kn} = \sum_{s=kn}^{kn+j-1}(X_{s+1} - X_s)$, the triangle inequality, and the constant stepsize within an epoch, we obtain

$$\begin{aligned}
\|X_t - X_{kn}\|_{\mathrm{op}} &\le \sum_{s=kn}^{kn+j-1} \|X_{s+1} - X_s\|_{\mathrm{op}} \\
&\le \sum_{s=kn}^{kn+j-1} \eta_s = j\frac{\eta_0}{\sqrt{k+1}} = j\eta_t \le (n-1)\eta_t.
\end{aligned}$$

When $j = 0$, the sum is empty and both $X_t - X_{kn}$ and $j\eta_t$ are zero. Next, the explicit schedule and $k' + 1 \ge 1$ give

$$\begin{aligned}
\frac{\eta_{k'n}}{\eta_t} &= \frac{\eta_0/\sqrt{k'+1}}{\eta_0/\sqrt{k+1}} = \sqrt{\frac{k+1}{k'+1}} \\
&= \sqrt{1 + \frac{k-k'}{k'+1}} \le \sqrt{1+(k-k')} = \sqrt{k-k'+1}.
\end{aligned}$$

Finally, partition the path from $k'n$ to t into the complete epochs $u = k', \dots, k-1$ and the first j steps of epoch k:

$$\begin{aligned}
\|X_t - X_{k'n}\|_{\mathrm{op}} &\le \sum_{s=k'n}^{t-1} \|X_{s+1} - X_s\|_{\mathrm{op}} \\
&\le \sum_{u=k'}^{k-1} \sum_{\ell=0}^{n-1} \eta_{un+\ell} + \sum_{\ell=0}^{j-1} \eta_{kn+\ell} \\
&= n \sum_{u=k'}^{k-1} \eta_{un} + j\eta_t \\
&\le n(k-k')\eta_{k'n} + j\eta_t \\
&\le \big[n(k-k')\sqrt{k-k'+1} + j\big]\eta_t \\
&\le n\big[(k-k')\sqrt{k-k'+1} + 1\big]\eta_t \\
&\le n(k-k'+1)^{3/2}\eta_t.
\end{aligned}$$

The fourth line uses $\eta_{un} \le \eta_{k'n}$ for $u \ge k'$. The fifth uses (29), and the sixth uses $j \le n$. For the last line, put $d = k - k' \ge 0$ and observe that

$$(d+1)^{3/2} - d\sqrt{d+1} = \sqrt{d+1} \ge 1.$$

Thus the historical-distance estimate follows from the single-step path bound and the stepsize ratio. □

The component-discrepancy contribution in (21) is the matrix

$$\mathcal{C}_t := (1-\beta) \sum_{s=0}^{t} \beta^{t-s} \big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big). \tag{31}$$

Its summands are evaluated at different points. We first replace each evaluation point by the start of its epoch, keeping an explicit residual for the change. The frozen terms can then use the common-point cancellation from lemma C.3.

Lemma C.5 (Exact decomposition by freezing each epoch)**.** *Fix $t = kn + j$. For $h \in \{0, \dots, k\}$ and $0 \le \ell < n$, define*

$$Z_{h,\ell} := \nabla f_{\pi_h(\ell+1)}(X_{hn}) - \nabla f(X_{hn}).$$

For $0 \le s \le t$, write $s = hn + \ell$ and define

$$R_s := \nabla f_{i_s}(X_s) - \nabla f(X_s) - Z_{h,\ell}. \tag{32}$$

Then

$$\begin{aligned}
\mathcal{C}_t = \; & \underbrace{(1-\beta) \sum_{\ell=0}^{j} \beta^{j-\ell} Z_{k,\ell}}_{\textit{current frozen epoch}} \\
& + \underbrace{\sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1} (1-\beta) \sum_{\ell=0}^{n-1} \beta^{n-1-\ell} Z_{k',\ell}}_{\textit{past complete frozen epochs}} \\
& + \underbrace{(1-\beta) \sum_{s=0}^{t} \beta^{t-s} R_s}_{\textit{moving-point residual}}.
\end{aligned} \tag{33}$$

Proof. Every $s \in \{0, \dots, t\}$ belongs to exactly one of the complete epochs $0, \dots, k-1$ or to positions $0, \dots, j$ of epoch k. Substituting $\nabla f_{i_s}(X_s) - \nabla f(X_s) = Z_{h,\ell} + R_s$ into (31) and making

this partition gives

$$\mathcal{C}_t = (1-\beta)\sum_{\ell=0}^{j}\beta^{t-(kn+\ell)}Z_{k,\ell} + (1-\beta)\sum_{k'=0}^{k-1}\sum_{\ell=0}^{n-1}\beta^{t-(k'n+\ell)}Z_{k',\ell} + (1-\beta)\sum_{s=0}^{t}\beta^{t-s}R_s.$$

For the current epoch, $t-(kn+\ell) = j-\ell$. For a past complete epoch,

$$\begin{aligned} t-(k'n+\ell) &= (k-k')n + j - \ell \\ &= (k-k'-1)n + j + 1 + (n-1-\ell). \end{aligned}$$

Factoring the corresponding power of β gives (33). The outer exponent records the age of the last step of that epoch, and the inner exponent records the age of a position relative to that last step. Keeping these factors separate will allow cancellation inside each complete epoch. When $k = 0$, there are no past complete epochs and the middle sum is empty. □

Lemma C.6 (Bounds for the current and past frozen epochs)**.** *With the notation of lemma C.5, the current frozen contribution satisfies*

$$\begin{aligned} \left\|(1-\beta)\sum_{\ell=0}^{j}\beta^{j-\ell}Z_{k,\ell}\right\|_{\mathrm{nuc}} &\le (1-\beta^n)\left[\sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{rD_0}\right] \\ &\quad + \sqrt{r}\,(\sqrt{D_1}+1)L(n-1)(1-\beta^n)\eta_t. \end{aligned} \tag{34}$$

The past complete frozen epochs satisfy

$$\begin{aligned} &\left\|\sum_{k'=0}^{k-1}\beta^{(k-k'-1)n+j+1}(1-\beta)\sum_{\ell=0}^{n-1}\beta^{n-1-\ell}Z_{k',\ell}\right\|_{\mathrm{nuc}} \\ &\le \frac{n(1-\beta)-(1-\beta^n)}{1-\beta^n}\left[\sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{rD_0}\right] \\ &\quad + \sqrt{r}\,(\sqrt{D_1}+1)Ln\big[n(1-\beta)-(1-\beta^n)\big]\left(\sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}\right)\eta_t. \end{aligned} \tag{35}$$

Proof. Current epoch. The common-point estimate (27), evaluated at X_{kn}, gives

$$\|Z_{k,\ell}\|_{\mathrm{nuc}} \le \sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_{kn})\|_{\mathrm{nuc}} + \sqrt{rD_0} \qquad (0 \le \ell \le j).$$

To express this bound at the current iterate, use the triangle inequality, full-gradient smoothness (2), and (28):

$$\begin{aligned} \|\nabla f(X_{kn})\|_{\mathrm{nuc}} &\le \|\nabla f(X_t)\|_{\mathrm{nuc}} + \|\nabla f(X_{kn}) - \nabla f(X_t)\|_{\mathrm{nuc}} \\ &\le \|\nabla f(X_t)\|_{\mathrm{nuc}} + L\|X_{kn} - X_t\|_{\mathrm{op}} \\ &\le \|\nabla f(X_t)\|_{\mathrm{nuc}} + L(n-1)\eta_t. \end{aligned}$$

All EMA weights are nonnegative. Setting $v = j-\ell$ and telescoping the finite geometric sum give

$$(1-\beta)\sum_{\ell=0}^{j}\beta^{j-\ell} = \sum_{v=0}^{j}(\beta^v - \beta^{v+1}) = 1-\beta^{j+1} \le 1-\beta^n,$$

where the last inequality uses $j+1 \le n$ and $0 < \beta < 1$.

Hence

$$\begin{aligned} &\left\|(1-\beta)\sum_{\ell=0}^{j}\beta^{j-\ell}Z_{k,\ell}\right\|_{\mathrm{nuc}} \\ &\le (1-\beta)\sum_{\ell=0}^{j}\beta^{j-\ell}\|Z_{k,\ell}\|_{\mathrm{nuc}} \\ &\le (1-\beta^{j+1})\left[\sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_{kn})\|_{\mathrm{nuc}} + \sqrt{rD_0}\right] \\ &\le (1-\beta^n)\left[\sqrt{r}\,(\sqrt{D_1}+1)\big(\|\nabla f(X_t)\|_{\mathrm{nuc}} + L(n-1)\eta_t\big) + \sqrt{rD_0}\right]. \end{aligned}$$

Distributing $1-\beta^n$ proves (34). No complete-epoch cancellation is needed for this estimate.

Past complete epochs. Fix $k' < k$. Applying (26) at $X_{k'n}$ gives $\sum_{\ell=0}^{n-1} Z_{k',\ell} = 0$. Therefore the weighted sum can be rewritten exactly as

$$\begin{aligned}(1-\beta)\sum_{\ell=0}^{n-1}\beta^{n-1-\ell}Z_{k',\ell} &= (1-\beta)\left[\sum_{\ell=0}^{n-1}\beta^{n-1-\ell}Z_{k',\ell} - \sum_{\ell=0}^{n-1}Z_{k',\ell}\right] \\ &= (1-\beta)\sum_{\ell=0}^{n-1}(\beta^{n-1-\ell}-1)Z_{k',\ell}.\end{aligned}$$

The coefficients $\beta^{n-1-\ell}-1$ are nonpositive. Their absolute values have total mass

$$\begin{aligned}(1-\beta)\sum_{\ell=0}^{n-1}|\beta^{n-1-\ell}-1| &= (1-\beta)\sum_{\ell=0}^{n-1}(1-\beta^{n-1-\ell}) \\ &= n(1-\beta) - (1-\beta)\sum_{v=0}^{n-1}\beta^v \\ &= n(1-\beta) - (1-\beta^n) \geq 0.\end{aligned} \tag{36}$$

The triangle inequality and (27) consequently imply

$$\begin{aligned}&\left\|(1-\beta)\sum_{\ell=0}^{n-1}\beta^{n-1-\ell}Z_{k',\ell}\right\|_{\mathrm{nuc}} \\ &\leq (1-\beta)\sum_{\ell=0}^{n-1}(1-\beta^{n-1-\ell})\|Z_{k',\ell}\|_{\mathrm{nuc}} \\ &\leq \left[n(1-\beta)-(1-\beta^n)\right]\left[\sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_{k'n})\|_{\mathrm{nuc}} + \sqrt{rD_0}\right].\end{aligned}$$

We next compare the historical gradient with the current one. Smoothness and (30) give

$$\begin{aligned}\|\nabla f(X_{k'n})\|_{\mathrm{nuc}} &\leq \|\nabla f(X_t)\|_{\mathrm{nuc}} + \|\nabla f(X_{k'n}) - \nabla f(X_t)\|_{\mathrm{nuc}} \\ &\leq \|\nabla f(X_t)\|_{\mathrm{nuc}} + L\|X_{k'n} - X_t\|_{\mathrm{op}} \\ &\leq \|\nabla f(X_t)\|_{\mathrm{nuc}} + Ln(k-k'+1)^{3/2}\eta_t.\end{aligned}$$

After this substitution, the sum over k' has two scalar factors to estimate. For $k \geq 1$, set $u = k-k'-1$. The indices $k' = 0,\ldots,k-1$ become $u = k-1,\ldots,0$, and $k-k'+1 = u+2$. Reversing the order of a finite sum yields

$$\begin{aligned}\sum_{k'=0}^{k-1}\beta^{(k-k'-1)n+j+1} &= \beta^{j+1}\sum_{u=0}^{k-1}\beta^{un} \\ &= \beta^{j+1}\frac{1-\beta^{kn}}{1-\beta^n} \leq \frac{1}{1-\beta^n},\end{aligned} \tag{37}$$

and

$$\begin{aligned}\sum_{k'=0}^{k-1}\beta^{(k-k'-1)n+j+1}(k-k'+1)^{3/2} &= \beta^{j+1}\sum_{u=0}^{k-1}\beta^{un}(u+2)^{3/2} \\ &\leq \sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}.\end{aligned} \tag{38}$$

Here $\beta^{j+1} \leq 1$, and extending the sum is valid because all terms are nonnegative. The last series is finite: the ratio of consecutive positive summands is

$$\beta^n\left(\frac{u+3}{u+2}\right)^{3/2} \longrightarrow \beta^n < 1.$$

Choose any θ with $\beta^n < \theta < 1$. For all sufficiently large u that ratio is at most θ, so the remaining tail is bounded by a finite multiple of $\sum_{v=0}^{\infty} \theta^v = 1/(1-\theta)$.

To finish, apply the triangle inequality to the outer sum, insert the bound for each complete epoch, and separate the terms involving the current gradient from those involving historical distance:

$$\begin{aligned}
&\left\| \sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1}(1-\beta) \sum_{\ell=0}^{n-1} \beta^{n-1-\ell} Z_{k',\ell} \right\|_{\text{nuc}} \\
&\quad \le \big[n(1-\beta) - (1-\beta^n)\big] \left[\sqrt{r}\,(\sqrt{D_1}+1) \|\nabla f(X_t)\|_{\text{nuc}} + \sqrt{r D_0} \right] \sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1} \\
&\qquad + \sqrt{r}\,(\sqrt{D_1}+1) L n \big[n(1-\beta) - (1-\beta^n)\big] \eta_t \sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1} (k-k'+1)^{3/2}.
\end{aligned}$$

Using (37) and (38) proves (35). If $k = 0$, the contribution is zero and the same inequality holds because its right-hand side is nonnegative. □

Lemma C.7 (Error caused by freezing the evaluation points)**.** *With R_s defined in (32),*

$$\left\| (1-\beta) \sum_{s=0}^{t} \beta^{t-s} R_s \right\|_{\text{nuc}} \le 2L(n-1)(1-\beta^n) \left[1 + \sum_{u=0}^{\infty} \beta^{un} \sqrt{u+2} \right] \eta_t. \tag{39}$$

Proof. Fix $s = hn + \ell \le t$. Since $i_s = \pi_h(\ell+1)$, the definition of R_s gives

$$R_s = \big[\nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{hn})\big] - \big[\nabla f(X_s) - \nabla f(X_{hn})\big].$$

Apply the triangle inequality, component-wise smoothness, full-gradient smoothness, and (28) within epoch h:

$$\begin{aligned}
\|R_s\|_{\text{nuc}} &\le \|\nabla f_{i_s}(X_s) - \nabla f_{i_s}(X_{hn})\|_{\text{nuc}} + \|\nabla f(X_s) - \nabla f(X_{hn})\|_{\text{nuc}} \\
&\le L\|X_s - X_{hn}\|_{\text{op}} + L\|X_s - X_{hn}\|_{\text{op}} \\
&\le 2L\ell\eta_{hn} \le 2L(n-1)\eta_{hn}.
\end{aligned}$$

For the current epoch $h = k$, $\eta_{hn} = \eta_t$. For a past complete epoch $h = k'$, (29) gives $\eta_{k'n} \le \sqrt{k-k'+1}\,\eta_t$. Partitioning the weighted residual sum accordingly gives

$$\begin{aligned}
\left\| (1-\beta) \sum_{s=0}^{t} \beta^{t-s} R_s \right\|_{\text{nuc}} &\le (1-\beta) \sum_{s=0}^{t} \beta^{t-s} \|R_s\|_{\text{nuc}} \\
&\le 2L(n-1) \left[\eta_t (1-\beta) \sum_{\ell=0}^{j} \beta^{j-\ell} + \sum_{k'=0}^{k-1} \eta_{k'n} (1-\beta) \sum_{\ell=0}^{n-1} \beta^{t-(k'n+\ell)} \right].
\end{aligned}$$

The first coefficient is $1 - \beta^{j+1} \le 1 - \beta^n$. For a complete epoch, factoring the exponent as in lemma C.5 gives

$$(1-\beta) \sum_{\ell=0}^{n-1} \beta^{t-(k'n+\ell)} = \beta^{(k-k'-1)n+j+1}(1-\beta) \sum_{\ell=0}^{n-1} \beta^{n-1-\ell} = \beta^{(k-k'-1)n+j+1}(1-\beta^n).$$

Thus the preceding norm is at most

$$2L(n-1)(1-\beta^n)\eta_t \left[1 + \sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1} \sqrt{k-k'+1} \right].$$

For $k \ge 1$, setting $u = k - k' - 1$ gives

$$\begin{aligned}
\sum_{k'=0}^{k-1} \beta^{(k-k'-1)n+j+1} \sqrt{k-k'+1} &= \beta^{j+1} \sum_{u=0}^{k-1} \beta^{un} \sqrt{u+2} \\
&\le \sum_{u=0}^{\infty} \beta^{un} \sqrt{u+2}.
\end{aligned}$$

For $k = 0$, the finite sum is empty and the inequality remains valid. The infinite series converges because its consecutive-term ratio is $\beta^n\sqrt{(u+3)/(u+2)} \to \beta^n < 1$; the same geometric-tail argument used in lemma C.6 applies. This proves (39). □

The two frozen estimates have the same coefficient on the current-gradient envelope after their contributions are added. Define this dimensionless coefficient by

$$c_{\beta,n} := (1-\beta^n) + \frac{n(1-\beta)-(1-\beta^n)}{1-\beta^n} = \frac{n(1-\beta)}{1-\beta^n} - \beta^n. \tag{40}$$

The first form displays the current-epoch contribution and the residual contribution from past complete epochs. To verify the second form, expand the fraction as $n(1-\beta)/(1-\beta^n) - 1$ and cancel $1-1$. Both terms in the first form are nonnegative by (36). This coefficient will determine how much of the gradient term can be absorbed in the descent argument.

Lemma C.8 (Component-discrepancy estimate)**.** *For every $t \geq 0$, the component discrepancy obeys*

$$\begin{aligned}
\|\mathcal{C}_t\|_{\mathrm{nuc}} \leq\ & \underbrace{\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\|\nabla f(X_t)\|_{\mathrm{nuc}}}_{\textit{gradient-dependent component discrepancy}} \\
& + \underbrace{\sqrt{rD_0}\,c_{\beta,n}}_{\textit{additive component discrepancy}} \\
& + \underbrace{\sqrt{r}\,(\sqrt{D_1}+1)L(n-1)(1-\beta^n)\eta_t}_{\textit{current-epoch anchor shift}} \\
& + \underbrace{\sqrt{r}\,(\sqrt{D_1}+1)Ln\big[n(1-\beta)-(1-\beta^n)\big]\left(\sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}\right)\eta_t}_{\textit{past-epoch anchor shifts}} \\
& + \underbrace{2L(n-1)(1-\beta^n)\left[1+\sum_{u=0}^{\infty}\beta^{un}\sqrt{u+2}\right]\eta_t}_{\textit{moving-point residual}}.
\end{aligned} \tag{41}$$

The first four terms are obtained from the frozen-epoch estimates in lemma C.6; the last term is the freezing residual from lemma C.7. The two series are finite, as proved in those lemmas.

Proof. Apply the triangle inequality to the three contributions in (33), then use (34), (35), and (39). This gives

$$\begin{aligned}
\|\mathcal{C}_t\|_{\mathrm{nuc}} \leq\ & \left[(1-\beta^n) + \frac{n(1-\beta)-(1-\beta^n)}{1-\beta^n}\right]\left[\sqrt{r}\,(\sqrt{D_1}+1)\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{rD_0}\right] \\
& + \sqrt{r}\,(\sqrt{D_1}+1)L(n-1)(1-\beta^n)\eta_t \\
& + \sqrt{r}\,(\sqrt{D_1}+1)Ln\big[n(1-\beta)-(1-\beta^n)\big]\left(\sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}\right)\eta_t \\
& + 2L(n-1)(1-\beta^n)\left[1+\sum_{u=0}^{\infty}\beta^{un}\sqrt{u+2}\right]\eta_t.
\end{aligned}$$

The first bracket equals $c_{\beta,n}$ by (40). Distributing it produces the first two terms in (41). The next two terms come from comparing the current and historical frozen anchors with X_t inside lemma C.6. The final term comes from lemma C.7. This proves the five-term bound without introducing an intermediate constant. □

C.4 Combining momentum error estimates

The initialization, full-gradient-lag, and component-discrepancy terms in (21) have now all been bounded. We combine them here. The first three terms in the resulting bound remain explicit

because they have different roles: initialization decays geometrically, the gradient-dependent term must be absorbed in the descent argument, and the additive term contributes the residual in the stationarity bound. Only the four coefficients multiplying the current stepsize are collected into one constant.

Lemma C.9 (Momentum-tracking estimate)**.** *For every $t \geq 0$,*

$$\begin{aligned}\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}} \leq \beta^{t+1}\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} + \sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\|\nabla f(X_t)\|_{\mathrm{nuc}} \\ + \sqrt{rD_0}\,c_{\beta,n} + K\eta_t,\end{aligned} \tag{42}$$

where the single aggregate constant is defined directly by

$$\begin{aligned}K := {} & \underbrace{\sqrt{r}\,(\sqrt{D_1}+1)L(n-1)(1-\beta^n)}_{\textit{current-epoch anchor shift}} \\ & + \underbrace{\sqrt{r}\,(\sqrt{D_1}+1)Ln\big[n(1-\beta)-(1-\beta^n)\big]\sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}}_{\textit{past-epoch anchor shifts}} \\ & + \underbrace{2L(n-1)(1-\beta^n)\left[1+\sum_{u=0}^{\infty}\beta^{un}\sqrt{u+2}\right]}_{\textit{moving-point residual}} \\ & + \underbrace{L\sum_{\ell=1}^{\infty}\beta^{\ell}\sqrt{\ell+1}}_{\textit{full-gradient lag}}.\end{aligned} \tag{43}$$

This finite nonnegative constant depends only on β, n, r, D_1, L and is independent of D_0, η_0, t, T and the permutation sequence.

Proof. The exact decomposition (21) reads

$$M_t - \nabla f(X_t) = \beta^{t+1}\big(M_{-1} - \nabla f(X_0)\big) + \mathcal{C}_t + \mathcal{L}_t.$$

Apply the triangle inequality and pull the nonnegative scalar β^{t+1} outside the first norm:

$$\begin{aligned}\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}} &\leq \big\|\beta^{t+1}\big(M_{-1} - \nabla f(X_0)\big)\big\|_{\mathrm{nuc}} \\ &\quad + \|\mathcal{C}_t\|_{\mathrm{nuc}} + \|\mathcal{L}_t\|_{\mathrm{nuc}} \\ &= \beta^{t+1}\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \\ &\quad + \|\mathcal{C}_t\|_{\mathrm{nuc}} + \|\mathcal{L}_t\|_{\mathrm{nuc}}.\end{aligned}$$

The equality uses the absolute homogeneity of a norm, $\|\alpha A\| = |\alpha|\|A\|$, with $\alpha = \beta^{t+1} \geq 0$. Inserting the five-term bound (41) and the lag estimate (24) gives

$$\begin{aligned}\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}} \leq {} & \beta^{t+1}\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \\ & + \sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{rD_0}\,c_{\beta,n} \\ & + \sqrt{r}\,(\sqrt{D_1}+1)L(n-1)(1-\beta^n)\eta_t \\ & + \sqrt{r}\,(\sqrt{D_1}+1)Ln\big[n(1-\beta)-(1-\beta^n)\big]\left(\sum_{u=0}^{\infty}\beta^{un}(u+2)^{3/2}\right)\eta_t \\ & + 2L(n-1)(1-\beta^n)\left[1+\sum_{u=0}^{\infty}\beta^{un}\sqrt{u+2}\right]\eta_t \\ & + L\left(\sum_{\ell=1}^{\infty}\beta^{\ell}\sqrt{\ell+1}\right)\eta_t.\end{aligned}$$

The last four lines sum to $K\eta_t$ by (43), proving (42). Convergence of all three series was established in the corresponding error estimates. Nonnegativity follows from (36) and the nonnegative factors in each summand. The displayed definition also shows the claimed parameter dependence. Every estimate holds on each fixed permutation sequence. □

C.5 BOUNDING ONE-STEP VALUE DECREASE

The momentum-tracking estimate now provides the error term needed to control the actual Muon direction.

For the update $X_{t+1} = X_t - \eta_t \operatorname{Polar}(M_t)$, the linear contribution to the objective change is $-\eta_t \langle \nabla f(X_t), \operatorname{Polar}(M_t) \rangle_F$. Lemma 4.1 in Section 4 lower-bounds the gradient inner product itself and therefore upper-bounds its negative multiple in the objective-change estimate. Thus no estimate of the distance between the full gradient and the Muon update direction is required.

Lemma C.10 (A momentum threshold ensuring a positive descent coefficient)**.** *The equation defining β_+^{perm} in Theorem 3.1 has a unique solution in $(0,1)$. For fixed $r \geq 1$ and $D_1 \geq 0$, this solution $\beta_+^{\mathrm{perm}}(n, r, D_1)$ is strictly increasing in the positive integer n. Consequently, every $\beta \in (\beta_+^{\mathrm{perm}}, 1)$ satisfies*

$$1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n} > 0. \tag{44}$$

Proof. The finite geometric identity $(1-\beta)\sum_{j=0}^{n-1} \beta^j = 1 - \beta^n$ gives

$$c_{\beta,n} = \frac{n}{\sum_{j=0}^{n-1} \beta^j} - \beta^n.$$

The denominator is positive on $[0,1]$, so this formula extends $c_{\beta,n}$ continuously to that closed interval. Its derivative on $(0,1)$ is

$$\frac{d}{d\beta} c_{\beta,n} = -\frac{n \sum_{j=1}^{n-1} j\beta^{j-1}}{(\sum_{j=0}^{n-1} \beta^j)^2} - n\beta^{n-1} < 0.$$

The first term is nonpositive, and the second is strictly negative; when $n = 1$, the first sum is empty and the derivative is -1. At the endpoints,

$$c_{0,n} = n, \qquad c_{1,n} = \frac{n}{n} - 1 = 0.$$

Since $r \geq 1$ and $D_1 \geq 0$, it follows that

$$2\sqrt{r}\,(\sqrt{D_1} + 1)c_{0,n} = 2n\sqrt{r}\,(\sqrt{D_1} + 1) \geq 2 > 1, \qquad 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{1,n} = 0 < 1.$$

Continuity gives a solution to $2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n} = 1$ in $(0,1)$, and strict decrease gives uniqueness.

To compare the roots across epoch lengths, fix $0 < \beta < 1$ and write $b_n := n^{-1} \sum_{j=0}^{n-1} \beta^j$. Every term in this average exceeds β^n, and hence

$$b_{n+1} = \frac{nb_n + \beta^n}{n+1} < b_n, \qquad c_{\beta,n+1} = \frac{1}{b_{n+1}} - \beta^{n+1} > \frac{1}{b_n} - \beta^n = c_{\beta,n}.$$

At $\beta = \beta_+^{\mathrm{perm}}(n, r, D_1)$, the value of $c_{\beta,n+1}$ therefore exceeds the fixed target $1/[2\sqrt{r}(\sqrt{D_1} + 1)]$. Since $c_{\beta,n+1}$ strictly decreases in β, its unique root satisfies $\beta_+^{\mathrm{perm}}(n+1, r, D_1) > \beta_+^{\mathrm{perm}}(n, r, D_1)$.

To obtain the positive coefficient used in the convergence proof, take any $\beta > \beta_+^{\mathrm{perm}}$. Strict decrease of $c_{\beta,n}$ in β and the defining equation imply

$$2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n} < 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta_+^{\mathrm{perm}},n} = 1.$$

Subtracting from one proves (44). This sign conclusion uses strict monotonicity together with the defining equation, rather than uniqueness alone. □

Lemma C.11 (One-step objective decrease)**.** *For every iteration $t \geq 0$, Algorithm 1 satisfies*

$$\begin{aligned} f(X_{t+1}) - f(X_t) \leq & -\left[1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}\right]\eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}} \\ & + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \eta_t \beta^{t+1} + 2\sqrt{rD_0}\, c_{\beta,n}\eta_t + \left(2K + \frac{L}{2}\right)\eta_t^2. \end{aligned} \tag{45}$$

Proof. From momentum tracking to a gradient inner-product bound. Combining (11) and (42) gives

$$\langle \nabla f(X_t), \mathrm{Polar}(M_t)\rangle_F \geq \|\nabla f(X_t)\|_{\mathrm{nuc}} - 2\begin{bmatrix}\beta^{t+1}\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}} \\ +\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\|\nabla f(X_t)\|_{\mathrm{nuc}} \\ +\sqrt{rD_0}\,c_{\beta,n}+K\eta_t\end{bmatrix}.$$

Distribute -2 and collect the two occurrences of $\|\nabla f(X_t)\|_{\mathrm{nuc}}$:

$$\begin{aligned}\langle \nabla f(X_t), \mathrm{Polar}(M_t)\rangle_F \geq \big[1-2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\big]\|\nabla f(X_t)\|_{\mathrm{nuc}} \\ -2\beta^{t+1}\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}}-2\sqrt{rD_0}\,c_{\beta,n}-2K\eta_t.\end{aligned} \tag{46}$$

From the inner-product bound to objective decrease. By (2), f satisfies the smoothness condition of lemma B.5. Apply (17) with $X=X_t$ and $S=-\eta_t\,\mathrm{Polar}(M_t)$. The update gives $X+S=X_{t+1}$, while (14) gives $\|S\|_{\mathrm{op}}^2\leq\eta_t^2$. Therefore,

$$\begin{aligned}f(X_{t+1})-f(X_t) &\leq -\eta_t\,\langle\nabla f(X_t),\mathrm{Polar}(M_t)\rangle_F+\frac{L}{2}\eta_t^2\|\mathrm{Polar}(M_t)\|_{\mathrm{op}}^2 \\ &\leq -\eta_t\,\langle\nabla f(X_t),\mathrm{Polar}(M_t)\rangle_F+\frac{L}{2}\eta_t^2.\end{aligned}$$

Multiplying (46) by $-\eta_t<0$ reverses its inequality. Substituting the result in the preceding upper bound gives

$$\begin{aligned}f(X_{t+1})-f(X_t)\leq\ &-\big[1-2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}\big]\eta_t\|\nabla f(X_t)\|_{\mathrm{nuc}} \\ &+2\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}}\eta_t\beta^{t+1}+2\sqrt{rD_0}\,c_{\beta,n}\eta_t+\left(2K+\frac{L}{2}\right)\eta_t^2.\end{aligned}$$

This proves (45). The derivation combines lemmas B.3, B.5, C.9 and 4.1; positivity of the gradient coefficient is used later when summing and dividing in the proof of Theorem C.12. □

C.6 Completing the proof

Theorem C.12 (Explicit finite-horizon stationarity bound)**.** *Suppose Assumptions 2.1 and 2.3 hold. Let $r=\min\{p,q\}$ and let β_+^{perm} be the unique solution in $(0,1)$ of*

$$2\sqrt{r}\,(\sqrt{D_1}+1)\left(\frac{n(1-\beta_+^{\mathrm{perm}})}{1-(\beta_+^{\mathrm{perm}})^n}-(\beta_+^{\mathrm{perm}})^n\right)=1.$$

Fix $\beta\in(\beta_+^{\mathrm{perm}},1)$ and $\eta_0>0$, and run Algorithm 1 with (4). Write $c_{\beta,n}=n(1-\beta)/(1-\beta^n)-\beta^n$. Then $1-2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}>0$. For every $T\in n\mathbb{N}$, on every permutation path,

$$\begin{aligned}&\frac{\sum_{t=0}^{T-1}\eta_t\|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1}\eta_t} \\ &\quad\leq\frac{2\sqrt{rD_0}\,c_{\beta,n}}{1-2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}} \\ &\qquad+\frac{1}{1-2\sqrt{r}\,(\sqrt{D_1}+1)c_{\beta,n}}\left[\frac{\Delta+2\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}}\eta_0\beta/(1-\beta)}{2n\eta_0\big(\sqrt{T/n+1}-1\big)}\right. \\ &\qquad\qquad\left.+\frac{\eta_0(2K+L/2)\big(1+\log(T/n)\big)}{2\big(\sqrt{T/n+1}-1\big)}\right].\end{aligned} \tag{47}$$

Here $\Delta=f(X_0)-f^\star$, and the finite constant K in (43) depends only on β,n,r,D_1,L. In particular, K is independent of D_0,η_0,T and of the permutation path. For every integer horizon $T\geq 1$, including incomplete epochs, the bound (49) below holds.

Proof of Theorem C.12. By lemma C.10, the choice $\beta > \beta_+^{\mathrm{perm}}$ ensures

$$1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n} > 0.$$

We combine the one-step estimate in lemma C.11 with the stepsize sums in lemma B.6.

Step 1: telescope the one-step bound. Fix any integer $T \geq 1$ and apply lemma C.11. Summing the left-hand side of (45) gives

$$\sum_{t=0}^{T-1} \big[f(X_{t+1}) - f(X_t)\big] = (f(X_1) - f(X_0)) + \cdots + (f(X_T) - f(X_{T-1})) = f(X_T) - f(X_0).$$

Summing (45) and moving the negative gradient term to the left thus yields

$$\begin{aligned}
&\big[1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}\big] \sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}} \\
&\quad \leq f(X_0) - f(X_T) \\
&\qquad + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \sum_{t=0}^{T-1} \eta_t \beta^{t+1} \\
&\qquad + 2\sqrt{rD_0}\, c_{\beta,n} \sum_{t=0}^{T-1} \eta_t + \left(2K + \frac{L}{2}\right) \sum_{t=0}^{T-1} \eta_t^2.
\end{aligned}$$

Since $f(X_T) \geq f^\star$ and $\Delta = f(X_0) - f^\star$, we have $f(X_0) - f(X_T) \leq \Delta$. Hence

$$\begin{aligned}
&\big[1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}\big] \sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}} \\
&\quad \leq \Delta \\
&\qquad + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \sum_{t=0}^{T-1} \eta_t \beta^{t+1} \\
&\qquad + 2\sqrt{rD_0}\, c_{\beta,n} \sum_{t=0}^{T-1} \eta_t + \left(2K + \frac{L}{2}\right) \sum_{t=0}^{T-1} \eta_t^2.
\end{aligned} \tag{48}$$

Both the square-bracket coefficient and $\sum_{t=0}^{T-1} \eta_t$ are positive. Dividing by them gives the bound valid for every integer horizon:

$$\begin{aligned}
\frac{\sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1} \eta_t} &\leq \frac{2\sqrt{rD_0}\, c_{\beta,n}}{1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}} \\
&\quad + \frac{\Delta + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \sum_{t=0}^{T-1} \eta_t \beta^{t+1}}{\big[1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}\big] \sum_{t=0}^{T-1} \eta_t} \\
&\quad + \frac{\left(2K + \frac{L}{2}\right) \sum_{t=0}^{T-1} \eta_t^2}{\big[1 - 2\sqrt{r}\,(\sqrt{D_1} + 1)c_{\beta,n}\big] \sum_{t=0}^{T-1} \eta_t}.
\end{aligned} \tag{49}$$

Step 2: obtain the explicit complete-epoch bound. For $T \in n\mathbb{N}$, lemma B.6 provides (18) and (19). In (49), use the lower bound for each denominator and the upper bounds for the nonnegative numerators. The initialization and squared-stepsize terms become, respectively,

$$\frac{\Delta + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \sum_{t=0}^{T-1} \eta_t \beta^{t+1}}{\sum_{t=0}^{T-1} \eta_t} \leq \frac{\Delta + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \eta_0 \beta/(1-\beta)}{2n\eta_0(\sqrt{T/n + 1} - 1)}$$

and

$$\begin{aligned}
\left(2K + \frac{L}{2}\right) \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} &\leq \left(2K + \frac{L}{2}\right) \frac{n\eta_0^2(1 + \log(T/n))}{2n\eta_0(\sqrt{T/n + 1} - 1)} \\
&= \frac{\eta_0(2K + L/2)(1 + \log(T/n))}{2(\sqrt{T/n + 1} - 1)}.
\end{aligned}$$

Substituting these two estimates and $c_{\beta,n} = n(1-\beta)/(1-\beta^n) - \beta^n$ proves the explicit complete-epoch bound in the theorem.

The bound (49) was obtained for every integer $T \geq 1$ in Step 1, so it also covers incomplete epochs. □

Proof of Theorem 3.1. Apply Theorem C.12. For arbitrary integer horizons, Lemma B.6 gives $\sum_{t<T} \eta_t \gtrsim \sqrt{T}$, $\sum_{t<T} \eta_t^2 = O(\log T)$, and $\sum_{t<T} \eta_t \beta^{t+1} = O(1)$. Inserting these estimates into (49) yields

$$\frac{\sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1} \eta_t} \leq \frac{2\sqrt{r D_0}\, c_{\beta,n}}{1 - 2\sqrt{r}\,(\sqrt{D_1}+1) c_{\beta,n}} + O\left(\frac{\log T}{\sqrt{T}}\right).$$

The remainder constant depends only on $\beta, n, r, D_1, L, \eta_0, \Delta, \|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}}$ and is independent of T and the permutation path. Since all stepsizes are positive,

$$\min_{0 \leq t < T} \|\nabla f(X_t)\|_{\mathrm{nuc}} \leq \frac{\sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1} \eta_t}.$$

Combining the two inequalities proves (5). For fixed n, r, D_1, the identity

$$c_{\beta,n} = \frac{n}{\sum_{j=0}^{n-1} \beta^j} - \beta^n \longrightarrow 0 \qquad (\beta \uparrow 1)$$

shows that the numerator of the residual term in (5) tends to zero and its denominator tends to one. For fixed $D_0 > 0$, Lemma C.10 shows that the positive numerator strictly decreases and the positive denominator strictly increases on $(\beta_+^{\mathrm{perm}}, 1)$, proving the strict decrease of this asymptotic residual bound to zero.

Setting $D_0 = 0$ removes the constant term from the last bound, giving

$$0 \leq \frac{\sum_{t=0}^{T-1} \eta_t \|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1} \eta_t} = O\left(\frac{\log T}{\sqrt{T}}\right) \longrightarrow 0 \qquad (T \to \infty,\ T \in \mathbb{N}).$$

The minimum-to-average inequality above then gives $\min_{0 \leq t < T} \|\nabla f(X_t)\|_{\mathrm{nuc}} = O(\log T/\sqrt{T}) \to 0$. All conclusions hold for every permutation sequence with the same fixed $\beta \in (\beta_+^{\mathrm{perm}}, 1)$. □

D I.I.D. CONVERGENCE: PROOF

We use the i.i.d. sampling scheme in Algorithm 2, deterministic initial matrices, and the assumptions of Theorem 3.2. Throughout this appendix, $\beta \in (0,1)$ and the stepsizes are the positive deterministic sequence specified in (6). Write $\Delta := f(X_0) - f^\star$. The proof follows the momentum-tracking framework of Appendix C. Section D.1 reuses the decomposition and gradient-lag argument of Sections C.1 and C.2, with the stepsize comparison adapted to Algorithm 2. Section D.2 replaces the permutation-specific argument of Section C.3 by a martingale estimate. Section D.3 then follows Sections C.4–C.6, retaining the expectation and stepsize changes needed for the i.i.d. conclusion. Throughout, we use full-objective smoothness and the affine growth condition assumed in Theorem 3.2.

D.1 SHARED ESTIMATES AND I.I.D. PRELIMINARIES

The identity in Lemma C.1 depends only on the momentum recursion and therefore applies unchanged. With $\mathcal{L}_t$ as defined in Lemma C.2, denote the middle term of (21) by

$$\mathcal{C}_t := (1-\beta) \sum_{s=0}^{t} \beta^{t-s} \big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big).$$

Thus the common decomposition reads

$$M_t - \nabla f(X_t) = \beta^{t+1}(M_{-1} - \nabla f(X_0)) + \mathcal{C}_t + \mathcal{L}_t. \tag{50}$$

The initialization term is exactly (22).

Let $\mathscr{F}_t := \sigma(i_0, \dots, i_t)$ for $t \geq 0$, with $\mathscr{F}_{-1}$ trivial. The canonical polar map is Borel measurable: it is the pointwise limit, as $\varepsilon \downarrow 0$, of the continuous maps $M \mapsto M(M^\top M + \varepsilon I_q)^{-1/2}$. This limit agrees with the compact-SVD convention, including at zero. The recursion therefore makes X_t and M_{t-1} $\mathscr{F}_{t-1}$-measurable. Independence and uniformity give

$$\mathbb{E}[\nabla f_{i_t}(X_t) \mid \mathscr{F}_{t-1}] = \nabla f(X_t), \qquad \mathbb{E}[\nabla f_{i_t}(X_t) - \nabla f(X_t) \mid \mathscr{F}_{t-1}] = 0. \tag{51}$$

Indeed, conditioning on $\mathscr{F}_{t-1}$ leaves X_t fixed and averages $\nabla f_{i_t}(X_t)$ over all n index choices.

The single-step bound (25) uses only the update and (14), so it also holds under i.i.d. sampling. Summing it and applying full-gradient smoothness gives, for $0 \leq s \leq t$,

$$\|X_t - X_s\|_{\mathrm{op}} \leq \sum_{u=s}^{t-1} \eta_u, \qquad \|\nabla f(X_s)\|_{\mathrm{nuc}} \leq \|\nabla f(X_t)\|_{\mathrm{nuc}} + L \sum_{u=s}^{t-1} \eta_u. \tag{52}$$

The second inequality uses the triangle inequality and (2). The samplewise affine growth argument in the proof of Lemma C.3, before its conversion to the nuclear norm, also applies unchanged and gives

$$\|\nabla f_{i_s}(X_s) - \nabla f(X_s)\|_F \leq (\sqrt{D_1} + 1)\|\nabla f(X_s)\|_{\mathrm{nuc}} + \sqrt{D_0}. \tag{53}$$

These estimates require neither a complete permutation nor component-wise smoothness; the zero-sum identity (26) is specific to complete permutations and is replaced here by (51).

Every random quantity used below is integrable on finite horizons. In fact, (52) bounds the full gradient by the deterministic number $\|\nabla f(X_0)\|_{\mathrm{nuc}} + L \sum_{u<t} \eta_u$. The affine growth condition then bounds every realized component gradient. The smooth objective is also bounded on the finite-radius set containing the iterates. Thus expectations, finite sums, and the martingale inequality below are well defined.

For the full-gradient lag, the proof of Lemma C.2 needs only the single-step bound, full-gradient smoothness, and an earlier-to-current stepsize comparison. Under (6), for $0 \leq s \leq t$,

$$\frac{\eta_s}{\eta_t} = \sqrt{\frac{t+1}{s+1}} \leq \sqrt{t-s+1}, \qquad \sum_{u=s}^{t-1} \eta_u \leq (t-s)\sqrt{t-s+1}\eta_t. \tag{54}$$

The first inequality follows from $t + 1 \leq (s + 1)(t - s + 1)$, and the second uses the nonincreasing stepsizes. In particular, for $1 \leq s \leq t$, $\eta_{s-1} \leq \sqrt{t-s+2}\eta_t$. Substituting this comparison into the same norm estimate and reindexing as in Lemma C.2 gives

$$\|\mathcal{L}_t\|_{\mathrm{nuc}} \leq L \sum_{s=1}^{t} \beta^{t-s+1}\eta_{s-1} \leq L\eta_t \sum_{\ell=1}^{\infty} \beta^\ell \sqrt{\ell+1}. \tag{55}$$

The series is the same convergent series as in that lemma. Initialization and lag are now controlled; the remaining sampling-dependent task is to bound $\mathbb{E}\|\mathcal{C}_t\|_{\mathrm{nuc}}$.

D.2 The component discrepancy under i.i.d. sampling

We use the Hilbert-space L^1 Davis inequality: there is a universal $C_D \geq 1$ such that every finite martingale-difference sequence (d_s) with integrable square function satisfies

$$\mathbb{E}\left\|\sum_s d_s\right\|_F \leq C_D\, \mathbb{E}\left(\sum_s \|d_s\|_F^2\right)^{1/2}. \tag{56}$$

Here matrices form a real Hilbert space under the Frobenius inner product. This terminal-value inequality follows from the maximal version for a martingale starting at zero, since the terminal norm is at most the running maximum; see the discussion following equation (1.3) in Wollgast & Zorin-Kranich (2023).

Lemma D.1 (Expected component-discrepancy bound)**.** *Define*

$$\nu_\beta := C_D\sqrt{r}\sqrt{\frac{1-\beta}{1+\beta}}, \qquad \rho_\beta := (\sqrt{D_1} + 1)\nu_\beta.$$

For every $t \geq 0$,

$$\begin{aligned}\mathbb{E}\left\|(1-\beta)\sum_{s=0}^{t}\beta^{t-s}\big(\nabla f_{i_s}(X_s)-\nabla f(X_s)\big)\right\|_{\mathrm{nuc}} \\ \leq \rho_\beta \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} + \nu_\beta\sqrt{D_0} \\ + C_D\sqrt{r}(\sqrt{D_1}+1)L(1-\beta)\left[\sum_{s=0}^{t}\beta^{2(t-s)}\left(\sum_{u=s}^{t-1}\eta_u\right)^2\right]^{1/2}.\end{aligned} \tag{57}$$

Proof. For a fixed terminal time t, the sequence $d_s = (1-\beta)\beta^{t-s}\big(\nabla f_{i_s}(X_s)-\nabla f(X_s)\big), 0 \leq s \leq t$, consists of martingale differences by (51). Its partial sums, starting from zero before $s=0$, form a Frobenius-space martingale. Applying (56) and $\|A\|_{\mathrm{nuc}} \leq \sqrt{r}\|A\|_F$ gives

$$\begin{aligned}\mathbb{E}\left\|(1-\beta)\sum_{s=0}^{t}\beta^{t-s}\big(\nabla f_{i_s}(X_s)-\nabla f(X_s)\big)\right\|_{\mathrm{nuc}} \\ \leq C_D\sqrt{r}(1-\beta)\,\mathbb{E}\left[\sum_{s=0}^{t}\beta^{2(t-s)}\|\nabla f_{i_s}(X_s)-\nabla f(X_s)\|_F^2\right]^{1/2}.\end{aligned}$$

After applying Davis, insert the shared samplewise envelope (53) and then (52) to obtain the pathwise bound

$$\begin{aligned}\|\nabla f_{i_s}(X_s)-\nabla f(X_s)\|_F &\leq (\sqrt{D_1}+1)\|\nabla f(X_s)\|_{\mathrm{nuc}} + \sqrt{D_0} \\ &\leq (\sqrt{D_1}+1)\|\nabla f(X_t)\|_{\mathrm{nuc}} + \sqrt{D_0} + (\sqrt{D_1}+1)L\sum_{u=s}^{t-1}\eta_u.\end{aligned}$$

Apply the Euclidean triangle inequality to the weighted sequence with weights β^{t-s}, separately to these three terms. Since

$$(1-\beta)\left(\sum_{s=0}^{t}\beta^{2(t-s)}\right)^{1/2} \leq \sqrt{\frac{1-\beta}{1+\beta}},$$

the first two contributions have coefficients ρ_β and ν_β. The third contribution is the deterministic square-function term in (57). Taking expectations proves the claim. The terminal gradient was inserted only after applying the martingale inequality, through a pathwise envelope; it was not treated as predictable at earlier times. □

D.3 From component discrepancy to the i.i.d. convergence theorem

We now follow the sequence of Sections C.4–C.6: combine the three errors, use the gradient inner-product bound, and sum objective decrease. The permutation-specific coefficients in Lemma C.9 are replaced by those of Lemma D.1, and the resulting bounds hold in expectation. We retain the stepsize weights explicitly in the finite-horizon estimate.

Proposition D.2 (Expected finite-horizon bound)**.** *Put $a_\beta := 1 - 2\rho_\beta$, and suppose $a_\beta > 0$. For $t \geq 0$ define the deterministic quantity*

$$\begin{aligned}\mathcal{E}_t := L\sum_{s=1}^{t}\beta^{t-s+1}\eta_{s-1} \\ + C_D\sqrt{r}(\sqrt{D_1}+1)L(1-\beta)\left[\sum_{s=0}^{t}\beta^{2(t-s)}\left(\sum_{u=s}^{t-1}\eta_u\right)^2\right]^{1/2}.\end{aligned} \tag{58}$$

Then

$$\begin{aligned}\mathbb{E}\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}} \leq \beta^{t+1}\|M_{-1}-\nabla f(X_0)\|_{\mathrm{nuc}} + \rho_\beta\mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} \\ + \nu_\beta\sqrt{D_0} + \mathcal{E}_t,\end{aligned} \tag{59}$$

and, for every $T \geq 1$,

$$\frac{\sum_{t<T} \eta_t \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t<T} \eta_t} \leq \frac{2\nu_\beta}{a_\beta}\sqrt{D_0} + \frac{\Delta + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \sum_{t<T} \eta_t \beta^{t+1}}{a_\beta \sum_{t<T} \eta_t} + \frac{2\sum_{t<T} \eta_t \mathcal{E}_t + (L/2)\sum_{t<T} \eta_t^2}{a_\beta \sum_{t<T} \eta_t}. \tag{60}$$

Proof. As in the proof of Lemma C.9, apply the triangle inequality to (50). Taking expectations and using (22), (55), and Lemma D.1 gives (59). The proof of Lemma C.11 then applies with this expected tracking bound: its two ingredients, Lemma I.1 and (17), hold on each realized trajectory under full-gradient smoothness. Taking expectations and collecting terms yields

$$a_\beta \eta_t \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} \leq \mathbb{E}f(X_t) - \mathbb{E}f(X_{t+1}) + 2\|M_{-1} - \nabla f(X_0)\|_{\mathrm{nuc}} \eta_t \beta^{t+1} + 2\nu_\beta \sqrt{D_0}\eta_t + 2\eta_t \mathcal{E}_t + \frac{L}{2}\eta_t^2.$$

The telescoping and division in the proof of Theorem C.12 now apply to the expected objective values: their sum is $f(X_0) - \mathbb{E}f(X_T) \leq \Delta$. Division by the positive number $a_\beta \sum_{t<T} \eta_t$ proves (60). The stepsizes are deterministic, so they remain outside every expectation. □

The remaining changes to the completion argument in Section C.6 are the i.i.d. threshold, its drift remainder, and the per-iteration stepsize sums.

Proof of Theorem 3.2. First, the threshold in the theorem is exactly the solution of

$$a_\beta > 0 \quad \Longleftrightarrow \quad 4C_D^2 r(\sqrt{D_1} + 1)^2(1 - \beta) < 1 + \beta \quad \Longleftrightarrow \quad \beta > \beta_+^{\mathrm{i.i.d.}}.$$

Since $C_D \geq 1$, $r \geq 1$, and $D_1 \geq 0$, this threshold lies in $(0, 1)$. Also $\nu_\beta \to 0$ and $a_\beta \to 1$ as $\beta \uparrow 1$. On $(\beta_+^{\mathrm{i.i.d.}}, 1)$, the positive quantity ν_β strictly decreases and a_β strictly increases, so for fixed $D_0 > 0$ the asymptotic residual bound $2\nu_\beta\sqrt{D_0}/a_\beta$ strictly decreases to zero.

For the lag term in (58), use the stepsize comparison in (55). For its square-function term, use (54) and set $\ell = t - s$. These are the i.i.d. counterparts of collecting the stepsize coefficients in Lemma C.9; they give $\mathcal{E}_t \leq K_{\mathrm{i.i.d.}} \eta_t$, where

$$K_{\mathrm{i.i.d.}} := L\sum_{\ell=1}^{\infty} \beta^\ell \sqrt{\ell + 1} + C_D\sqrt{r}(\sqrt{D_1} + 1)L(1 - \beta)\left[\sum_{\ell=0}^{\infty} \ell^2(\ell + 1)\beta^{2\ell}\right]^{1/2}. \tag{61}$$

Both series converge: consecutive-term ratios tend to β and β^2, respectively. The constant depends only on β, r, D_1, L. The scalar integral comparisons and geometric-sum argument in the proof of Lemma B.6 apply to $\eta_t = \eta_0/\sqrt{t + 1}$ directly, without grouping updates into epochs, and give

$$\sum_{t<T} \eta_t \geq 2\eta_0(\sqrt{T + 1} - 1), \qquad \sum_{t<T} \eta_t^2 \leq \eta_0^2(1 + \log T), \qquad \sum_{t<T} \eta_t \beta^{t+1} \leq \frac{\eta_0 \beta}{1 - \beta}.$$

Substitution into (60) proves a weighted expected-gradient bound with residual $2\nu_\beta\sqrt{D_0}/a_\beta$ (the explicit term in (7)) and remainder $O(\log T/\sqrt{T})$, for the infinite-horizon run in Algorithm 2.

Positivity of the deterministic stepsizes implies

$$\min_{0 \leq t < T} \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} \leq \frac{\sum_{t<T} \eta_t \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t<T} \eta_t}.$$

This proves the stated bounds. Setting $D_0 = 0$ removes the residual. Finally, the pathwise minimum is bounded by each individual gradient norm, so taking expectations and then minimizing gives

$$\mathbb{E} \min_{0 \leq t < T} \|\nabla f(X_t)\|_{\mathrm{nuc}} \leq \min_{0 \leq t < T} \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}}.$$

□

E PERMUTATION DIVERGENCE: PROOF

This appendix proves the permutation divergence claim for the shared quadratic witness in (10) through the pathwise escape mechanism and then deduces Theorem 3.4 using function-class membership. The ordering and parameter dependence of the phase thresholds are established in Section 3.3. The global regularity, conditioning, and function-class membership of the quadratic family are proved separately in Appendix G. All matrices and gradients in the escape argument are generated by Algorithm 1; no prescribed surrogate trajectory is substituted for the actual iterates.

For the divergence proofs, write $\operatorname{diag}(X) = (X_{11}, \dots, X_{rr})^\top$ for $X \in \mathbb{R}^{p\times q}$, and let $\operatorname{Diag}(v) \in \mathbb{R}^{p\times q}$ place the entries of $v \in \mathbb{R}^r$ on the diagonal and zero elsewhere. Let P_n have ones in its first n diagonal entries and zeros elsewhere. With e_i the standard basis of $\mathbb{R}^r$, put $u_i = (4n+1)e_i - \sum_{a=1}^n e_a$ for $1 \le i \le n$.

E.1 FIXED CONSTRUCTION, INDICES, AND THE UNIFORM MEMORY WINDOW

Fix $r \ge n \ge 8$ and an arbitrary $c > 0$, and let f_i, f be the quadratic functions in (10). The dynamical estimates below hold for this fixed c; the theorem-level specialization uses (9). Let k denote an epoch, $j \in \{0, \dots, n-1\}$ a position within that epoch, $i \in \{1, \dots, n\}$ a finite-sum index, and $a \in \{1, \dots, n\}$ an active diagonal coordinate. Inner-iteration indices are s, t. As in Section 2,

$$t = kn + j, \qquad k = \lfloor t/n \rfloor, \qquad i_t = \pi_k(j+1).$$

Throughout, the stepsizes satisfy Assumption 3.3; in particular, $\eta_t \le \eta_0$. The permutations are arbitrary. Set

$$\beta_- = (70n)^{-1/\lfloor n/8 \rfloor}, \qquad s_0 \ge 8\lfloor n/8 \rfloor \eta_0, \tag{62}$$

and fix any $\beta \in [0, \beta_-]$. The initial iterate is $X_0 = s_0 P_n$. The proof uses a window of $\lfloor n/8 \rfloor$ updates, fixed by n.

Lemma E.1 (A uniform forgetting window)**.** *The parameters in (62) satisfy*

$$\lfloor n/8 \rfloor \ge 1, \qquad n \ge 8\lfloor n/8 \rfloor, \qquad \beta^{\lfloor n/8 \rfloor} \le \frac{1}{70n}. \tag{63}$$

For $0 < \beta \le \beta_-$, the shortest integer window obtained from the geometric-tail condition obeys

$$\left\lceil \frac{\log(70n)}{-\log \beta} \right\rceil \le \lfloor n/8 \rfloor.$$

Proof. Since $n \ge 8$, its quotient by eight has floor at least one, and $\lfloor n/8 \rfloor \le n/8$. Raising $0 \le \beta \le \beta_-$ to the positive integer power $\lfloor n/8 \rfloor$ gives $\beta^{\lfloor n/8 \rfloor} \le \beta_-^{\lfloor n/8 \rfloor} = (70n)^{-1}$. If $\beta > 0$, this inequality implies $\lfloor n/8 \rfloor(-\log \beta) \ge \log(70n)$. Dividing by the positive number $-\log \beta$ and taking the ceiling gives the last assertion because $\lfloor n/8 \rfloor$ is an integer. If $\beta = 0$, the power inequality holds directly and no logarithm is used. □

E.2 GRADIENT FORMULA AND INVARIANT ACTIVE COORDINATES

Lemma E.2 (Exact reduction to the active diagonal dynamics)**.** *For every $X \in \mathbb{R}^{p\times q}$ and every finite-sum index i,*

$$\nabla f_i(X) = c\left[\left(u_i^\top \operatorname{diag}(X)\right) \operatorname{Diag}(u_i) + nX\right]. \tag{64}$$

Along the initialized Muon trajectory there are vectors $x_t, m_t, \zeta_t \in \mathbb{R}^n$ such that

$$X_t = \operatorname{Diag}(x_{t,1}, \dots, x_{t,n}, \underbrace{0, \dots, 0}_{r-n}),$$

where the trailing zero block is absent when $r = n$. The momentum matrix M_t and selected component gradient $\nabla f_{i_t}(X_t)$ have the same form, with active entries m_t and ζ_t, respectively. Define

$$q_i(x) := 4nx_i - \sum_{a \ne i} x_a, \qquad x \in \mathbb{R}^n. \tag{65}$$

Then

$$\zeta_{t,a} = \begin{cases} c\big(4nq_{i_t}(x_t) + nx_{t,i_t}\big), & a = i_t, \\ c\big(-q_{i_t}(x_t) + nx_{t,a}\big), & a \neq i_t, \end{cases} \tag{66}$$

and the actual coordinate recursion is

$$m_{t,a} = \beta m_{t-1,a} + (1-\beta)\zeta_{t,a}, \qquad x_{t+1,a} = x_{t,a} - \eta_t \operatorname{sign}(m_{t,a}), \tag{67}$$

with $x_{0,a} = s_0$ and $m_{-1,a} = 0$. Here $\zeta_{t,a} := [\nabla f_{i_t}(X_t)]_{aa}$ is the ath diagonal entry of the selected component gradient, while the full gradient is written $\nabla f(X_t)$.

Proof. Let $b_i(X) := u_i^\top \operatorname{diag}(X)$. This is a linear functional. For any perturbation Z,

$$\frac{d}{d\varepsilon} f_i(X + \varepsilon Z)\Big|_{\varepsilon=0} = c\, b_i(X) b_i(Z) + cn \langle X, Z \rangle_F .$$

The identity $b_i(Z) = \langle \operatorname{Diag}(u_i), Z \rangle_F$ converts the right-hand side to $\langle c[b_i(X)\operatorname{Diag}(u_i) + nX], Z \rangle_F$. Since this holds for every Z, it proves (64).

The initial iterate and buffer are supported only on the first n diagonal positions of the rectangular matrix. Suppose X_t and M_{t-1} have this property. Each u_i is supported on the first n coordinates, so (64) shows that $\nabla f_{i_t}(X_t)$ also has the property. The linear momentum update preserves it. By Lemma B.4, the rectangular polar factor is supported on the same diagonal positions, with entries $\operatorname{sign}(m_{t,a})$ in the active coordinates and zero elsewhere. Thus the iterate update preserves the property as well. Induction proves it at every time.

For an active vector x, the coordinates of u_i are $4n$ in position i, -1 in the other active positions, and zero in the inactive positions. Hence $u_i^\top \operatorname{diag}(X) = 4nx_i - \sum_{a \neq i} x_a = q_i(x)$. Substituting these entries into (64) proves (66). Applying the diagonal polar formula in the update gives (67). □

Consequently, the n active coordinates obey the same recursion as the $n \times n$ member of this quadratic family at the same fixed scale c. Lemma B.4 identifies the rectangular iterates as its zero extension, with every other entry remaining zero. The regularizer $cn\|X\|_F^2/2$ acts on the entire matrix space, but its gradient vanishes on the inactive coordinates of this particular trajectory.

E.3 An expanding cone and finite-memory forgetting

Define the cumulative stepsize and expanding radius by

$$\Gamma_t := \sum_{s=0}^{t-1} \eta_s, \qquad \rho_t := s_0 + \Gamma_t, \qquad \Gamma_0 = 0. \tag{68}$$

The radius is positive and nondecreasing. Let

$$\mathcal{Q}_t := \left\{ \operatorname{Diag}(x_1, \dots, x_n, 0, \dots, 0) : \frac{3}{4}\rho_t \le x_a \le \rho_t \text{ for every } a \le n \right\}. \tag{69}$$

We first derive estimates conditional on $X_s \in \mathcal{Q}_s$ at the relevant times. Proposition E.6 will establish this condition for the actual trajectory by a first-exit argument.

Lemma E.3 (Signs and sizes of component gradients in the cone)**.** *If $X_t \in \mathcal{Q}_t$, then for every finite-sum index i,*

$$(2n+1)\rho_t \le q_i(x_t) \le 4n\rho_t. \tag{70}$$

The ith active entry of $\nabla f_i(X_t)$ is positive, whereas each entry indexed by $a \neq i$ is at most $-c(n+1)\rho_t$. Moreover,

$$\big|[\nabla f_i(X_t)]_{aa}\big| \le 17cn^2\rho_t \quad \textit{for every } a \le n. \tag{71}$$

Proof. The cone inequalities give

$$q_i(x_t) = 4nx_{t,i} - \sum_{a \neq i} x_{t,a} \ge 4n\frac{3}{4}\rho_t - (n-1)\rho_t = (2n+1)\rho_t.$$

All active coordinates are positive. Dropping the subtracted sum therefore gives $q_i(x_t) \le 4nx_{t,i} \le 4n\rho_t$. This proves (70). The distinguished gradient entry $c(4nq_i(x_t) + nx_{t,i})$ is positive because $c > 0$ and both summands are positive. For $a \neq i$,

$$c(-q_i(x_t) + nx_{t,a}) \le -c(2n+1)\rho_t + cn\rho_t = -c(n+1)\rho_t < 0.$$

For the distinguished entry, positivity and the upper bounds give

$$|c(4nq_i(x_t) + nx_{t,i})| \le c(16n^2 + n)\rho_t \le 17cn^2\rho_t.$$

For every other active entry, its negativity gives

$$|c(-q_i(x_t) + nx_{t,a})| = c(q_i(x_t) - nx_{t,a}) \le cq_i(x_t) \le 4cn\rho_t \le 17cn^2\rho_t.$$

These inequalities prove the envelope for all active entries. □

Lemma E.4 (Forgetting a coordinate's last correcting input)**.** *Suppose $X_s \in \mathcal{Q}_s$ for every $0 \le s \le t$. Fix an active coordinate a. If*

$$i_s \neq a \quad \textit{for every } s \in \{\max(0, t - \lfloor n/8 \rfloor + 1), \dots, t\}, \tag{72}$$

then $m_{t,a} < 0$, and hence $x_{t+1,a} = x_{t,a} + \eta_t$.

Proof. First suppose $t \ge \lfloor n/8 \rfloor - 1$. For each $s \in \{t - \lfloor n/8 \rfloor + 1, \dots, t\}$, the assumption that index a is not selected in this interval and Lemma E.3 imply $\zeta_{s,a} \le -c(n+1)\rho_s$. The past radius is comparable to the current one. Indeed,

$$0 \le \rho_t - \rho_s = \sum_{q=s}^{t-1} \eta_q \le \lfloor n/8 \rfloor \eta_0, \qquad \rho_t \ge s_0 \ge 8\lfloor n/8 \rfloor \eta_0.$$

It follows that $\rho_s \ge \rho_t - \lfloor n/8 \rfloor \eta_0 \ge (7/8)\rho_t \ge \rho_t/2$, so

$$\zeta_{s,a} \le -\frac{cn}{2}\rho_t \quad (t - \lfloor n/8 \rfloor + 1 \le s \le t). \tag{73}$$

Write $A_n := 17n^2$. Before these $\lfloor n/8 \rfloor$ inputs, the older momentum has magnitude at most $cA_n\rho_t$. To see this, if $t - \lfloor n/8 \rfloor = -1$ the buffer is zero. Otherwise unrolling the recursion from $m_{-1,a} = 0$ gives

$$m_{t-\lfloor n/8 \rfloor,a} = (1-\beta) \sum_{q=0}^{t-\lfloor n/8 \rfloor} \beta^{t-\lfloor n/8 \rfloor - q} \zeta_{q,a}.$$

Every coefficient is nonnegative, and their sum is $1 - \beta^{t-\lfloor n/8 \rfloor + 1} \le 1$. By the cone history and (71), $|\zeta_{q,a}| \le cA_n\rho_q \le cA_n\rho_t$. The triangle inequality therefore gives $|m_{t-\lfloor n/8 \rfloor,a}| \le cA_n\rho_t$ in both cases. Integer powers with exponent zero in these finite sums are interpreted as one, so the same recursion identities hold at $\beta = 0$.

Unrolling just the most recent $\lfloor n/8 \rfloor$ steps, and using (73), now yields

$$\begin{aligned} m_{t,a} &= \beta^{\lfloor n/8 \rfloor} m_{t-\lfloor n/8 \rfloor,a} + (1-\beta) \sum_{s=t-\lfloor n/8 \rfloor+1}^{t} \beta^{t-s} \zeta_{s,a} \\ &\le c \left[\beta^{\lfloor n/8 \rfloor} A_n - (1 - \beta^{\lfloor n/8 \rfloor}) \frac{n}{2} \right] \rho_t \\ &\le c \left[\frac{17n}{70} - \frac{n}{2} + \frac{1}{140} \right] \rho_t = -\frac{c(36n-1)}{140} \rho_t < 0. \end{aligned} \tag{74}$$

The last inequality uses (63); the geometric sum of the recent coefficients equals $1 - \beta^{\lfloor n/8 \rfloor}$ also when $\beta = 0$.

It remains to consider $t < \lfloor n/8 \rfloor - 1$. In that case (72) says index a has never been selected in updates $0, \dots, t$. Every input $\zeta_{s,a}$ in that range is strictly negative by Lemma E.3. Thus

$$m_{t,a} = (1-\beta) \sum_{s=0}^{t} \beta^{t-s} \zeta_{s,a} < 0 :$$

all summands are nonpositive and the term with $s = t$ is strictly negative since $1 - \beta > 0$. Finally, in either time range, (67) gives $x_{t+1,a} = x_{t,a} - \eta_t(-1) = x_{t,a} + \eta_t$. □

E.4 WEIGHTED COUNTING AND CLOSURE OF THE CONE

For a fixed active coordinate a, call update s *bad* if index a is selected at least once in $\{\max(0, s - \lfloor n/8 \rfloor + 1), \dots, s\}$, and *good* otherwise. These terms concern the update of that coordinate. One selection of index a can make its own update and the next $\lfloor n/8 \rfloor - 1$ updates bad. Good updates move the coordinate outward by exactly η_s whenever the required cone history holds. For $t \geq 0$, define the occurrence weight and bad-update weight by

$$v_{t,a} := \sum_{\substack{0 \leq s < t \\ i_s = a}} \eta_s, \qquad w_{t,a} := \sum_{\substack{0 \leq s < t \\ s \text{ bad for } a}} \eta_s. \tag{75}$$

Both are zero at $t = 0$.

Lemma E.5 (Weighted contamination bound)**.** *For every $t \geq 0$ and active coordinate a,*

$$w_{t,a} \leq \lfloor n/8 \rfloor v_{t,a} \leq \lfloor n/8 \rfloor \left(\frac{\Gamma_t}{n} + \eta_0 \right). \tag{76}$$

This estimate uses only the permutation order and stepsize schedule, without a cone assumption.

Proof. Write $t = kn + j$ with $0 \leq j < n$. Each completed epoch selects index a exactly once, and the first j updates of the current epoch select it at most once. The occurrence in epoch zero, if it precedes t, has weight at most η_0. For every later epoch $h \geq 1$ with an occurrence $q_h < t$, nonincreasing stepsizes give

$$\eta_{q_h} \leq \eta_{hn} \leq \frac{1}{n} \sum_{s=(h-1)n}^{hn-1} \eta_s.$$

The preceding epoch is complete before t. Distinct occurrence epochs map to distinct preceding epochs, so summing these bounds gives

$$v_{t,a} \leq \eta_0 + \frac{1}{n} \sum_{s=0}^{kn-1} \eta_s \leq \eta_0 + \frac{\Gamma_t}{n}.$$

For $k = 0$, the sum is empty and the bound follows from the single possible occurrence in the first partial epoch. Thus no constancy of the stepsizes within an epoch is needed.

An occurrence at time $q < t$ can contaminate only times $q, \dots, q + \lfloor n/8 \rfloor - 1$. Since the stepsizes are nonincreasing, the weight of those times that precede t is at most $\lfloor n/8 \rfloor \eta_q$. Every bad update before t belongs to at least one such window generated by an occurrence before t. Summing the weights of all windows can count an update more than once, so it is an upper bound on their union weight. Hence

$$w_{t,a} \leq \sum_{\substack{0 \leq q < t \\ i_q = a}} \sum_{s=q}^{\min(t-1, q+\lfloor n/8 \rfloor - 1)} \eta_s \leq \sum_{\substack{0 \leq q < t \\ i_q = a}} \lfloor n/8 \rfloor \eta_q = \lfloor n/8 \rfloor v_{t,a}.$$

Combining the two bounds proves (76), including windows crossing epoch boundaries. □

Proposition E.6 (Invariance of the expanding cone)**.** *For every permutation sequence and every $t \geq 0$, $X_t \in \mathcal{Q}_t$. In particular, for each active coordinate a,*

$$\frac{3}{4}(s_0 + \Gamma_t) \leq x_{t,a} \leq s_0 + \Gamma_t. \tag{77}$$

Proof. At time zero every active coordinate equals $s_0 = \rho_0$, and all inactive entries are zero, so $X_0 \in \mathcal{Q}_0$. If there were a failure, the set of failing integer times would have a least element $t_* \geq 1$. The coordinate-support property would still hold by Lemma E.2, and $X_s \in \mathcal{Q}_s$ for every $s < t_*$ by minimality. Thus Lemma E.4 applies to every good update $s < t_*$.

Fix a. Its increment is exactly $+\eta_s$ at every good update. At every bad update, (67) and $|\operatorname{sign}(m_{s,a})| \leq 1$ give an increment at least $-\eta_s$. Relative to a sequence of only good increments,

replacing one update by a bad one therefore costs at most $2\eta_s$. Summing these actual increment bounds gives

$$\begin{aligned} x_{t_*,a} &\geq s_0 + \sum_{\substack{s<t_* \\ s \text{ good for } a}} \eta_s - \sum_{\substack{s<t_* \\ s \text{ bad for } a}} \eta_s \\ &= s_0 + \Gamma_{t_*} - 2w_{t_*,a} \\ &\geq s_0 - 2\lfloor n/8 \rfloor \eta_0 + \left(1 - \frac{2\lfloor n/8 \rfloor}{n}\right) \Gamma_{t_*}. \end{aligned}$$

By $s_0 \geq 8\lfloor n/8 \rfloor \eta_0$ and $n \geq 8\lfloor n/8 \rfloor$,

$$s_0 - 2\lfloor n/8 \rfloor \eta_0 \geq \frac{3}{4} s_0, \qquad 1 - \frac{2\lfloor n/8 \rfloor}{n} \geq \frac{3}{4}.$$

Since $\Gamma_{t_*} \geq 0$, these inequalities imply $x_{t_*,a} \geq (3/4)(s_0 + \Gamma_{t_*})$. Conversely, every coordinate increment is at most η_s, so summing the upper bounds gives $x_{t_*,a} \leq s_0 + \Gamma_{t_*}$. Both inequalities hold for every active coordinate, proving $X_{t_*} \in \mathcal{Q}_{t_*}$, a contradiction. Thus there is no first exit and (77) holds for every time. □

Corollary E.7 (Escape of every active coordinate)**.** *For every active coordinate a, $x_{t,a} \to +\infty$, and consequently $\|X_t\|_F \to \infty$.*

Proof. By Assumption 3.3,

$$\Gamma_t = \sum_{s=0}^{t-1} \eta_s \longrightarrow \infty.$$

Proposition E.6 now implies $x_{t,a} \geq (3/4)(s_0 + \Gamma_t) \to \infty$. Diagonal support also gives

$$\|X_t\|_F^2 = \sum_{a=1}^{n} x_{t,a}^2 \geq \frac{9n}{16}(s_0 + \Gamma_t)^2 \longrightarrow \infty.$$

Taking square roots proves the iterate-norm conclusion. □

E.5 Permutation divergence of the shared quadratic witness and completion of the theorem

Proof of the permutation divergence claim for the shared quadratic witness. Take the functions f_i, f in (10), with c chosen by (9). Assumption 2.1 specifies $L > 0$, so $c > 0$. This function and its scale are fixed by L, n, p, q before any algorithmic choices.

For any stepsize sequence satisfying Assumption 3.3, choose $s_0 \geq 8\lfloor n/8 \rfloor \eta_0$ and set $X_0 = s_0 P_n$. This initial iterate is independent of β and the permutation sequence. Now fix any $\beta \in [0, \beta_-]$ and any permutation sequence. Lemma E.1 gives all uniform window conditions needed in the escape proof. The preceding lemmas apply to the chosen fixed $c > 0$, so Corollary E.7 directly yields $\|X_t\|_F \to \infty$ for these functions.

To obtain objective divergence, average the component lower bounds from Lemma G.1:

$$f(X_t) = \frac{1}{n} \sum_{i=1}^{n} f_i(X_t) \geq \frac{cn}{2} \|X_t\|_F^2 \longrightarrow \infty.$$

For gradient divergence, the same lemma gives $H := \nabla^2 f \succeq cnI$. Since $\nabla f(X_t) = H[X_t]$, Cauchy–Schwarz gives

$$cn\|X_t\|_F^2 \leq \langle X_t, H[X_t] \rangle_F = \langle X_t, \nabla f(X_t) \rangle_F \leq \|X_t\|_F \|\nabla f(X_t)\|_F.$$

The cone lower bound ensures $X_t \neq 0$ at every time. Dividing by $\|X_t\|_F$ gives $\|\nabla f(X_t)\|_F \geq cn\|X_t\|_F \to \infty$, because $c > 0$ is fixed. This proves all three limits in (8) for the same function and trajectory.

The function and its scale c are independent of the entire stepsize sequence, β, the stopping horizon, and the permutation path. The initialization may depend on the first stepsize η_0 but is uniform over

all sequences satisfying Assumption 3.3 with that first stepsize, and is fixed before β and the path are chosen. The positive factor c does not affect the momentum signs, and the only restriction on β used is $\beta^{\lfloor n/8 \rfloor} \le (70n)^{-1}$. Thus the same function and initial state work throughout the stated low-momentum interval. Every argument holds on each permutation path, so the result also holds almost surely under random reshuffling. □

Proof of Theorem 3.4. Fix any class $\mathcal{F}^{n;p,q}_{L,D_0,D_1}$ in the stated regime, and choose f_i, f and c by (10) and (9). Lemma G.3 in Appendix G.3 verifies $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$. For every stepsize sequence satisfying Assumption 3.3, the preceding proof supplies an initial iterate $X_0 = s_0 P_n$, with $s_0 \ge 8\lfloor n/8 \rfloor \eta_0$, for which Algorithm 1 exhibits triple divergence for every fixed $\beta \in [0, \beta_-]$ and every permutation path. The function and its decomposition are independent of the entire stepsize sequence, β, and the path. For a fixed first stepsize η_0, the same initial iterate works for every sequence satisfying Assumption 3.3 with that first stepsize, independently of β and the path. This establishes the claimed existence statement. □

F I.I.D. DIVERGENCE: FULL STATEMENT AND PROOF

This appendix proves the i.i.d. divergence claim for the shared quadratic witness in (10) through the pathwise escape mechanism and then deduces Theorem 3.5 using function-class membership. The ordering and parameter dependence of the phase thresholds are established in Section 3.3. The global regularity, conditioning, and function-class membership of the quadratic family are proved separately in Appendix G. All matrices and gradients in the escape argument are generated by Algorithm 2; no prescribed surrogate trajectory is substituted for the actual iterates.

Theorem F.1 (Full i.i.d. low-momentum divergence statement)**.** *For any problem class $\mathcal{F}^{n;p,q}_{L,D_0,D_1}$ with $r \ge n \ge 8$ and $D_1 \ge 16n^2 + 2n - 1$, there is a fixed quadratic function $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$ with a fixed finite-sum decomposition, independent of ε, the stepsize sequence, β, and the samples, with the following property. For every $\varepsilon \in (0,1)$ and every deterministic stepsize sequence satisfying Assumption 3.3, there exists a deterministic initial iterate $X_0 \in \mathbb{R}^{p\times q}$ depending only on η_0, ε, and the fixed class parameters. Couple the runs of Algorithm 2 for all $\beta \in [0, \beta_-]$, where $\beta_- := (70n)^{-1/\lfloor n/8 \rfloor}$, through one i.i.d. index sequence. There is an event $\mathcal{E}_\varepsilon$, determined only by that sequence and the prescribed stepsizes, with $\mathbb{P}(\mathcal{E}_\varepsilon) \ge 1 - \varepsilon$, on which all these runs satisfy*

$$\|X_t\|_F \longrightarrow \infty, \qquad f(X_t) \longrightarrow \infty, \qquad \|\nabla f(X_t)\|_F \longrightarrow \infty \quad (t \to \infty) \tag{78}$$

simultaneously for every $\beta \in [0, \beta_-]$. The initial iterate and event are independent of β.

F.1 FIXED CONSTRUCTION, INDICES, AND THE UNIFORM MEMORY WINDOW

Fix $r \ge n \ge 8$, an arbitrary $c > 0$, and $\varepsilon \in (0,1)$, and let f_i, f be the quadratic functions in (10). The dynamical estimates below hold for this fixed c; the theorem-level specialization uses (9). Let i denote a finite-sum index, $a \in \{1, \dots, n\}$ an active diagonal coordinate, and s, t iteration indices. At each update, i_t is sampled independently and uniformly from $\{1, \dots, n\}$. The stepsizes are deterministic and satisfy Assumption 3.3; in particular, $\eta_t \le \eta_0$. Set

$$\begin{aligned} &K := \lfloor n/8 \rfloor, \qquad \beta_- := (70n)^{-1/K}, \\ &u_\varepsilon := 6\eta_0 K^2 \log(n/\varepsilon), \qquad s_0 \ge \max\{8K\eta_0, 6u_\varepsilon\}. \end{aligned} \tag{79}$$

The initial iterate is $X_0 = s_0 P_n$ and $M_{-1} = 0$. The window length K is fixed by n. All low-momentum trajectories are coupled to the same i.i.d. index sequence; their common event will be defined using only that sequence and the prescribed stepsizes.

Lemma F.2 (A uniform forgetting window)**.** *For every $\beta \in [0, \beta_-]$, the parameters in (79) satisfy*

$$\lfloor n/8 \rfloor \ge 1, \qquad n \ge 8\lfloor n/8 \rfloor, \qquad \beta^{\lfloor n/8 \rfloor} \le \frac{1}{70n}. \tag{80}$$

For $0 < \beta \le \beta_-$, the shortest integer window obtained from the geometric-tail condition obeys

$$\left\lceil \frac{\log(70n)}{-\log \beta} \right\rceil \le \lfloor n/8 \rfloor.$$

Proof. Since $n \geq 8$, its quotient by eight has floor at least one, and $\lfloor n/8 \rfloor \leq n/8$. Raising $0 \leq \beta \leq \beta_-$ to the positive integer power $\lfloor n/8 \rfloor$ gives $\beta^{\lfloor n/8 \rfloor} \leq \beta_-^{\lfloor n/8 \rfloor} = (70n)^{-1}$. If $\beta > 0$, this inequality implies $\lfloor n/8 \rfloor(-\log \beta) \geq \log(70n)$. Dividing by the positive number $-\log \beta$ and taking the ceiling gives the last assertion because $\lfloor n/8 \rfloor$ is an integer. If $\beta = 0$, the power inequality holds directly and no logarithm is used. □

F.2 Gradient formula and invariant active coordinates

Lemma F.3 (Exact reduction to the active diagonal dynamics)**.** *For every $X \in \mathbb{R}^{p \times q}$ and every finite-sum index i,*

$$\nabla f_i(X) = c\left[\left(u_i^\top \operatorname{diag}(X)\right) \operatorname{Diag}(u_i) + nX\right]. \tag{81}$$

Along the initialized Muon trajectory there are vectors $x_t, m_t, \zeta_t \in \mathbb{R}^n$ such that

$$X_t = \operatorname{Diag}(x_{t,1}, \ldots, x_{t,n}, \underbrace{0, \ldots, 0}_{r-n}),$$

where the trailing zero block is absent when $r = n$. The momentum matrix M_t and selected component gradient $\nabla f_{i_t}(X_t)$ have the same form, with active entries m_t and ζ_t, respectively. Define

$$q_i(x) := 4nx_i - \sum_{a \neq i} x_a, \qquad x \in \mathbb{R}^n. \tag{82}$$

Then

$$\zeta_{t,a} = \begin{cases} c\big(4nq_{i_t}(x_t) + nx_{t,i_t}\big), & a = i_t, \\ c\big(-q_{i_t}(x_t) + nx_{t,a}\big), & a \neq i_t, \end{cases} \tag{83}$$

and the actual coordinate recursion is

$$m_{t,a} = \beta m_{t-1,a} + (1-\beta)\zeta_{t,a}, \qquad x_{t+1,a} = x_{t,a} - \eta_t \operatorname{sign}(m_{t,a}), \tag{84}$$

with $x_{0,a} = s_0$ and $m_{-1,a} = 0$. Here $\zeta_{t,a} := [\nabla f_{i_t}(X_t)]_{aa}$ is the ath diagonal entry of the selected component gradient, while the full gradient is written $\nabla f(X_t)$.

Proof. Let $b_i(X) := u_i^\top \operatorname{diag}(X)$. This is a linear functional. For any perturbation Z,

$$\frac{d}{dh} f_i(X + hZ)\Big|_{h=0} = c\, b_i(X) b_i(Z) + cn \langle X, Z \rangle_F .$$

The identity $b_i(Z) = \langle \operatorname{Diag}(u_i), Z \rangle_F$ converts the right-hand side to $\langle c[b_i(X) \operatorname{Diag}(u_i) + nX], Z \rangle_F$. Since this holds for every Z, it proves (81).

The initial iterate and buffer are supported only on the first n diagonal positions of the rectangular matrix. Suppose X_t and M_{t-1} have this property. Each u_i is supported on the first n coordinates, so (81) shows that $\nabla f_{i_t}(X_t)$ also has the property. The linear momentum update preserves it. By Lemma B.4, the rectangular polar factor is supported on the same diagonal positions, with entries $\operatorname{sign}(m_{t,a})$ in the active coordinates and zero elsewhere. Thus the iterate update preserves the property as well. Induction proves it at every time.

For an active vector x, the coordinates of u_i are $4n$ in position i, -1 in the other active positions, and zero in the inactive positions. Hence $u_i^\top \operatorname{diag}(X) = 4nx_i - \sum_{a \neq i} x_a = q_i(x)$. Substituting these entries into (81) proves (83). Applying the diagonal polar formula in the update gives (84). □

Consequently, the n active coordinates obey the same recursion as the $n \times n$ member of this quadratic family at the same fixed scale c. Lemma B.4 identifies the rectangular iterates as its zero extension, with every other entry remaining zero. The regularizer $cn\|X\|_F^2/2$ acts on the entire matrix space, but its gradient vanishes on the inactive coordinates of this particular trajectory.

F.3 AN EXPANDING CONE AND FINITE-MEMORY FORGETTING

Define the cumulative stepsize and expanding radius by

$$\Gamma_t := \sum_{s=0}^{t-1} \eta_s, \qquad \rho_t := s_0 + \Gamma_t, \qquad \Gamma_0 = 0. \tag{85}$$

The radius is positive and nondecreasing. Let

$$\mathcal{Q}_t := \left\{ \mathrm{Diag}(x_1, \dots, x_n, 0, \dots, 0) : \frac{2}{3}\rho_t \le x_a \le \rho_t \text{ for every } a \le n \right\}. \tag{86}$$

We first derive estimates conditional on $X_s \in \mathcal{Q}_s$ at the relevant times. Proposition F.8 will establish this condition for the actual trajectory by a first-exit argument.

Lemma F.4 (Signs and sizes of component gradients in the cone)**.** *If $X_t \in \mathcal{Q}_t$, then for every finite-sum index i,*

$$(5n/3 + 1)\rho_t \le q_i(x_t) \le 4n\rho_t. \tag{87}$$

The ith active entry of $\nabla f_i(X_t)$ is positive, whereas each entry indexed by $a \neq i$ is at most $-c(2n/3+1)\rho_t$. Moreover,

$$\left|[\nabla f_i(X_t)]_{aa}\right| \le 17cn^2\rho_t \quad \textit{for every } a \le n. \tag{88}$$

Proof. The cone inequalities give

$$q_i(x_t) = 4nx_{t,i} - \sum_{a \neq i} x_{t,a} \ge 4n\frac{2}{3}\rho_t - (n-1)\rho_t = (5n/3+1)\rho_t.$$

All active coordinates are positive. Dropping the subtracted sum therefore gives $q_i(x_t) \le 4nx_{t,i} \le 4n\rho_t$. This proves (87). The distinguished gradient entry $c(4nq_i(x_t) + nx_{t,i})$ is positive because $c > 0$ and both summands are positive. For $a \neq i$,

$$c(-q_i(x_t) + nx_{t,a}) \le -c(5n/3+1)\rho_t + cn\rho_t = -c(2n/3+1)\rho_t < 0.$$

For the distinguished entry, positivity and the upper bounds give

$$|c(4nq_i(x_t) + nx_{t,i})| \le c(16n^2 + n)\rho_t \le 17cn^2\rho_t.$$

For every other active entry, its negativity gives

$$|c(-q_i(x_t) + nx_{t,a})| = c(q_i(x_t) - nx_{t,a}) \le cq_i(x_t) \le 4cn\rho_t \le 17cn^2\rho_t.$$

These inequalities prove the envelope for all active entries. □

Lemma F.5 (Forgetting a coordinate's last correcting input)**.** *Suppose $X_s \in \mathcal{Q}_s$ for every $0 \le s \le t$. Fix an active coordinate a. If*

$$i_s \neq a \quad \textit{for every } s \in \{\max(0, t - \lfloor n/8 \rfloor + 1), \dots, t\}, \tag{89}$$

then $m_{t,a} < 0$, and hence $x_{t+1,a} = x_{t,a} + \eta_t$.

Proof. First suppose $t \ge \lfloor n/8 \rfloor - 1$. For each $s \in \{t - \lfloor n/8 \rfloor + 1, \dots, t\}$, the assumption that index a is not selected in this interval and Lemma F.4 imply $\zeta_{s,a} \le -c(2n/3+1)\rho_s$. The past radius is comparable to the current one. Indeed,

$$0 \le \rho_t - \rho_s = \sum_{q=s}^{t-1} \eta_q \le \lfloor n/8 \rfloor \eta_0, \qquad \rho_t \ge s_0 \ge 8\lfloor n/8 \rfloor \eta_0.$$

It follows that $\rho_s \ge \rho_t - \lfloor n/8 \rfloor \eta_0 \ge (7/8)\rho_t$. Since $(2n/3+1)(7/8) \ge n/2$,

$$\zeta_{s,a} \le -\frac{cn}{2}\rho_t \quad (t - \lfloor n/8 \rfloor + 1 \le s \le t). \tag{90}$$

Write $A_n := 17n^2$. Before these $\lfloor n/8 \rfloor$ inputs, the older momentum has magnitude at most $cA_n\rho_t$. To see this, if $t - \lfloor n/8 \rfloor = -1$ the buffer is zero. Otherwise unrolling the recursion from $m_{-1,a} = 0$ gives

$$m_{t-\lfloor n/8 \rfloor,a} = (1-\beta) \sum_{q=0}^{t-\lfloor n/8 \rfloor} \beta^{t-\lfloor n/8 \rfloor - q} \zeta_{q,a}.$$

Every coefficient is nonnegative, and their sum is $1 - \beta^{t-\lfloor n/8 \rfloor + 1} \le 1$. By the cone history and (88), $|\zeta_{q,a}| \le cA_n\rho_q \le cA_n\rho_t$. The triangle inequality therefore gives $|m_{t-\lfloor n/8 \rfloor,a}| \le cA_n\rho_t$ in both cases. Integer powers with exponent zero in these finite sums are interpreted as one, so the same recursion identities hold at $\beta = 0$.

Unrolling just the most recent $\lfloor n/8 \rfloor$ steps, and using (90), now yields

$$\begin{aligned} m_{t,a} &= \beta^{\lfloor n/8 \rfloor} m_{t-\lfloor n/8 \rfloor,a} + (1-\beta) \sum_{s=t-\lfloor n/8 \rfloor+1}^{t} \beta^{t-s} \zeta_{s,a} \\ &\le c \left[\beta^{\lfloor n/8 \rfloor} A_n - (1-\beta^{\lfloor n/8 \rfloor}) \frac{n}{2} \right] \rho_t \\ &\le c \left[\frac{17n}{70} - \frac{n}{2} + \frac{1}{140} \right] \rho_t = -\frac{c(36n-1)}{140} \rho_t < 0. \end{aligned} \tag{91}$$

The last inequality uses (80); the geometric sum of the recent coefficients equals $1 - \beta^{\lfloor n/8 \rfloor}$ also when $\beta = 0$.

It remains to consider $t < \lfloor n/8 \rfloor - 1$. In that case (89) says index a has never been selected in updates $0, \dots, t$. Every input $\zeta_{s,a}$ in that range is strictly negative by Lemma F.4. Thus

$$m_{t,a} = (1-\beta) \sum_{s=0}^{t} \beta^{t-s} \zeta_{s,a} < 0 :$$

all summands are nonpositive and the term with $s = t$ is strictly negative since $1 - \beta > 0$. Finally, in either time range, (84) gives $x_{t+1,a} = x_{t,a} - \eta_t(-1) = x_{t,a} + \eta_t$. □

F.4 Weighted counting and closure of the cone

For a fixed active coordinate a, call update s *bad* if index a is selected at least once in $\{\max(0, s - \lfloor n/8 \rfloor + 1), \dots, s\}$, and *good* otherwise. These terms concern the update of that coordinate. One selection of index a can make its own update and the next $\lfloor n/8 \rfloor - 1$ updates bad. Good updates move the coordinate outward by exactly η_s whenever the required cone history holds. For $t \ge 0$, define the occurrence weight and bad-update weight by

$$v_{t,a} := \sum_{\substack{0 \le s < t \\ i_s = a}} \eta_s, \qquad w_{t,a} := \sum_{\substack{0 \le s < t \\ s \text{ bad for } a}} \eta_s. \tag{92}$$

Both are zero at $t = 0$.

Lemma F.6 (Weighted contamination bound)**.** *For every index sequence, every $t \ge 0$, and every active coordinate a,*

$$w_{t,a} \le K v_{t,a}. \tag{93}$$

This estimate uses only the nonincreasing stepsizes, without a cone assumption.

Proof. An occurrence at time $q < t$ can contaminate only times $q, \dots, q+K-1$. Since the stepsizes are nonincreasing, the weight of those times that precede t is at most $K\eta_q$. Every bad update before t belongs to at least one such window generated by an occurrence before t. Summing the weights of all windows can count an update more than once, so it is an upper bound on their union weight. Hence

$$w_{t,a} \le \sum_{\substack{0 \le q < t \\ i_q = a}} \sum_{s=q}^{\min(t-1, q+K-1)} \eta_s \le \sum_{\substack{0 \le q < t \\ i_q = a}} K\eta_q = K v_{t,a}.$$

□

Lemma F.7 (A time-uniform i.i.d. occupation event)**.** *There is an event $\mathcal{E}_\varepsilon$, determined only by the i.i.d. index sequence and the prescribed stepsizes, such that $\mathbb{P}(\mathcal{E}_\varepsilon) \geq 1-\varepsilon$ and*

$$w_{t,a} \leq \frac{1}{6}\Gamma_t + u_\varepsilon \quad \textit{for every } t \geq 0,\ a \leq n \quad \textit{on } \mathcal{E}_\varepsilon. \tag{94}$$

The event is independent of the momentum parameter β.

Proof. Put $\mathscr{F}_t := \sigma(i_0, \ldots, i_{t-1})$, with $\mathscr{F}_0$ trivial, and

$$Y_{t,a} := K\mathbf{1}\{i_t = a\} - \frac{1}{6}, \qquad \lambda := \frac{1}{6\eta_0 K^2}.$$

Given $\mathscr{F}_t$, the variable $Y_{t,a}$ has range length K and mean $K/n - 1/6 \leq -1/24$, because $K/n \leq 1/8$. The bounded-variable exponential-moment inequality (Hoeffding, 1963) $\mathbb{E}e^{h(Z-\mathbb{E}Z)} \leq e^{h^2(b-a)^2/8}$ for $Z \in [a,b]$ gives

$$\mathbb{E}\big[e^{\lambda\eta_t Y_{t,a}} \mid \mathscr{F}_t\big] \leq \exp\left(-\frac{\lambda\eta_t}{24} + \frac{\lambda^2\eta_t^2 K^2}{8}\right) \leq \exp(-\lambda\eta_t/48) \leq 1,$$

where $\lambda\eta_t K^2/8 \leq 1/48$ uses $\eta_t \leq \eta_0$. For completeness, the exponential-moment inequality follows by writing $\psi(h) = \log \mathbb{E}e^{hZ}$: its second derivative is the variance under the exponentially tilted distribution, at most $(b-a)^2/4$, and integrating twice from 0 gives the stated bound.

Thus

$$L_{t,a} := \exp\left(\lambda \sum_{s<t} \eta_s Y_{s,a}\right) = \exp(\lambda[K v_{t,a} - \Gamma_t/6])$$

is a nonnegative $\mathscr{F}_t$-supermartingale with $L_{0,a} = 1$. Stopping it at its first crossing of $e^{\lambda u_\varepsilon}$, truncated at a deterministic time T, gives a crossing probability at most $e^{-\lambda u_\varepsilon}$; letting $T \to \infty$ yields (Howard et al., 2020)

$$\mathbb{P}\left(\sup_{t\geq 0}\{K v_{t,a} - \Gamma_t/6\} > u_\varepsilon\right) \leq e^{-\lambda u_\varepsilon} = \frac{\varepsilon}{n}.$$

Define

$$\mathcal{E}_\varepsilon := \{\forall a \leq n,\ \forall t \geq 0 : K v_{t,a} - \Gamma_t/6 \leq u_\varepsilon\}. \tag{95}$$

This is a measurable event with countably many constraints. A union bound over a gives probability at least $1-\varepsilon$. Lemma F.6 gives (94). Neither the event nor its proof uses β or the iterates. □

Proposition F.8 (Invariance of the expanding cone)**.** *On every index sequence in $\mathcal{E}_\varepsilon$, simultaneously for every $\beta \in [0, \beta_-]$ and every $t \geq 0$, $X_t \in \mathcal{Q}_t$. In particular, for each active coordinate a,*

$$\frac{2}{3}(s_0 + \Gamma_t) \leq x_{t,a} \leq s_0 + \Gamma_t. \tag{96}$$

Proof. Fix one index sequence in $\mathcal{E}_\varepsilon$ and any $\beta \in [0, \beta_-]$. At time zero every active coordinate equals $s_0 = \rho_0$, and all inactive entries are zero, so $X_0 \in \mathcal{Q}_0$. If there were a failure, the set of failing integer times would have a least element $t_* \geq 1$. The coordinate-support property would still hold by Lemma F.3, and $X_s \in \mathcal{Q}_s$ for every $s < t_*$ by minimality. Thus Lemma F.5 applies to every good update $s < t_*$.

Fix a. Its increment is exactly $+\eta_s$ at every good update. At every bad update, (84) and $|\operatorname{sign}(m_{s,a})| \leq 1$ give an increment at least $-\eta_s$. Relative to a sequence of only good increments, replacing one update by a bad one therefore costs at most $2\eta_s$. Summing these actual increment bounds gives

$$\begin{aligned} x_{t_*,a} &\geq s_0 + \sum_{\substack{s<t_* \\ s \text{ good for } a}} \eta_s - \sum_{\substack{s<t_* \\ s \text{ bad for } a}} \eta_s \\ &= s_0 + \Gamma_{t_*} - 2w_{t_*,a} \\ &\geq s_0 - 2u_\varepsilon + \frac{2}{3}\Gamma_{t_*}. \end{aligned}$$

Here the last inequality uses Lemma F.7. By $s_0 \geq 6u_\varepsilon$,

$$s_0 - 2u_\varepsilon \geq \frac{2}{3}s_0.$$

Since $\Gamma_{t_*} \geq 0$, these inequalities imply $x_{t_*,a} \geq (2/3)(s_0 + \Gamma_{t_*})$. Conversely, every coordinate increment is at most η_s, so summing the upper bounds gives $x_{t_*,a} \leq s_0 + \Gamma_{t_*}$. Both inequalities hold for every active coordinate, proving $X_{t_*} \in \mathcal{Q}_{t_*}$, a contradiction. Thus there is no first exit and (96) holds for every time. The event does not mention β, and the argument uses only the common bound $\beta^K \leq (70n)^{-1}$. It therefore holds for all low-momentum parameters on this one event, without a union bound over β. □

Corollary F.9 (Escape of every active coordinate)**.** *On $\mathcal{E}_\varepsilon$, for every $\beta \in [0, \beta_-]$ and every active coordinate a, $x_{t,a} \to +\infty$, and consequently $\|X_t\|_F \to \infty$.*

Proof. By Assumption 3.3,

$$\Gamma_t = \sum_{s=0}^{t-1} \eta_s \longrightarrow \infty.$$

Proposition F.8 now implies $x_{t,a} \geq (2/3)(s_0 + \Gamma_t) \to \infty$. Diagonal support also gives

$$\|X_t\|_F^2 = \sum_{a=1}^{n} x_{t,a}^2 \geq \frac{4n}{9}(s_0 + \Gamma_t)^2 \longrightarrow \infty.$$

Taking square roots proves the iterate-norm conclusion. □

F.5 I.I.D. DIVERGENCE OF THE SHARED QUADRATIC WITNESS AND COMPLETION OF THE THEOREM

Proof of the i.i.d. divergence claim for the shared quadratic witness. Take the functions f_i, f in (10), with c chosen by (9). Assumption 2.1 specifies $L > 0$, so $c > 0$. This function and its scale are fixed by L, n, p, q before any algorithmic choices.

For any $\varepsilon \in (0, 1)$ and any deterministic stepsize sequence satisfying Assumption 3.3, choose s_0 as in (79) and set $X_0 = s_0 P_n$. This initial iterate is independent of β and the index sequence. Lemma F.7 supplies a sample-only event of probability at least $1 - \varepsilon$. On this event, Corollary F.9 gives $\|X_t\|_F \to \infty$ simultaneously for all $\beta \in [0, \beta_-]$.

To obtain objective divergence, average the component lower bounds from Lemma G.1:

$$f(X_t) = \frac{1}{n}\sum_{i=1}^{n} f_i(X_t) \geq \frac{cn}{2}\|X_t\|_F^2 \longrightarrow \infty.$$

For gradient divergence, the same lemma gives $H := \nabla^2 f \succeq cnI$. Since $\nabla f(X_t) = H[X_t]$, Cauchy–Schwarz gives

$$cn\|X_t\|_F^2 \leq \langle X_t, H[X_t]\rangle_F = \langle X_t, \nabla f(X_t)\rangle_F \leq \|X_t\|_F \|\nabla f(X_t)\|_F.$$

The cone lower bound ensures $X_t \neq 0$ at every time on the event. Dividing by $\|X_t\|_F$ gives $\|\nabla f(X_t)\|_F \geq cn\|X_t\|_F \to \infty$, because $c > 0$ is fixed. This proves all three limits in (78) for the same function and trajectory.

The function and its scale are independent of ε, the entire stepsize sequence, β, the stopping horizon, and the samples. For fixed class parameters, the initialization depends only on η_0, ε and works for every admissible deterministic schedule with that first stepsize. The same event works simultaneously for all $\beta \in [0, \beta_-]$. This proves the stated simultaneous probability guarantee. □

Proof of Theorems 3.5 and F.1. Fix any class $\mathcal{F}_{L,D_0,D_1}^{n;p,q}$ in the stated regime, and choose f_i, f and c by (10) and (9). Lemma G.3 verifies $f \in \mathcal{F}_{L,D_0,D_1}^{n;p,q}$. For each prescribed deterministic stepsize sequence satisfying Assumption 3.3 and each $\varepsilon \in (0, 1)$, the preceding proof supplies one deterministic initial iterate and one event with probability at least $1 - \varepsilon$ on which Algorithm 2 exhibits triple divergence simultaneously for every $\beta \in [0, \beta_-]$. The construction and initialization have exactly the dependencies stated in the theorem. □

F.6 EXPECTED-GRADIENT CONSEQUENCES

Proposition F.10 (Expected-gradient lower bounds). *Under the i.i.d. initialization in (79), put* $A_\varepsilon := \frac{2}{3}(1-\varepsilon)cn^{3/2} > 0$. *For every* $\beta \in [0, \beta_-]$,

$$\mathbb{E}_\beta \|\nabla f(X_t)\|_{\mathrm{nuc}} \geq A_\varepsilon(s_0 + \Gamma_t), \qquad t \geq 0. \tag{97}$$

Consequently, for $T \geq 1$,

$$\begin{aligned} \min_{0\leq t<T} \mathbb{E}_\beta \|\nabla f(X_t)\|_{\mathrm{nuc}} &\geq A_\varepsilon s_0 > 0, \\ \frac{\sum_{t<T} \eta_t \mathbb{E}_\beta \|\nabla f(X_t)\|_{\mathrm{nuc}}}{\Gamma_T} &\geq A_\varepsilon \left(s_0 + \frac{\Gamma_T}{2} - \frac{\sum_{t<T} \eta_t^2}{2\Gamma_T} \right). \end{aligned} \tag{98}$$

For $\eta_t = \eta_0/\sqrt{t+1}$, *the weighted average is* $\Omega(\sqrt{T})$, *uniformly over the low-momentum interval.*

Proof. On $\mathcal{E}_\varepsilon$, Proposition F.8 gives $\|X_t\|_F \geq (2/3)\sqrt{n}(s_0 + \Gamma_t)$. The quadratic Hessian bound gives $\|\nabla f(X_t)\|_{\mathrm{nuc}} \geq \|\nabla f(X_t)\|_F \geq cn\|X_t\|_F$. The norm is nonnegative off the event, so multiplying this lower bound by its probability proves (97). The expectations are finite at every fixed time: the update length gives $\|X_t\|_F \leq \sqrt{n}(s_0 + \Gamma_t)$ on every index sequence, and the full gradient is a fixed linear map.

The minimum bound follows from $\Gamma_t \geq 0$. Expanding the square of $\Gamma_T = \sum_{t<T} \eta_t$ gives

$$\sum_{t<T} \eta_t \Gamma_t = \frac{1}{2} \left(\Gamma_T^2 - \sum_{t<T} \eta_t^2 \right).$$

Multiply (97) by η_t, sum, and divide by $\Gamma_T > 0$ to obtain (98). For the square-root schedule, integral comparison gives $\Gamma_T = \Theta(\sqrt{T})$ and $\sum_{t<T} \eta_t^2 = O(\log T)$, proving the stated rate. All lower-bound constants are independent of β in the specified interval. □

G Quadratic Divergence Witness: Properties

Fix $r \geq n \geq 8$ and $c > 0$, and let f_i, f be the quadratic functions in (10). We verify the advertised regularity, common minimizer, least-squares structure, full-Hessian condition number, and Assumptions 2.1 and 2.3. These are global properties of the functions and do not require a particular initialization or an invariant trajectory. Hessian conditioning is measured in Frobenius geometry.

G.1 Strong convexity, common minimizer, and least-squares structure

All regularity statements below hold on the entire space $\mathbb{R}^{p\times q}$, not just on the invariant active block. This is needed to place the constructed family in the problem class used by the positive theorem.

Lemma G.1 (Strong convexity, interpolation, and least-squares structure)**.** *Let $H_i := \nabla^2 f_i$ denote the constant Hessian as a linear operator on the Frobenius inner-product space. For every $Z \in \mathbb{R}^{p\times q}$,*

$$H_i[Z] = c\left[\left(u_i^\top \operatorname{diag}(Z)\right)\operatorname{Diag}(u_i) + nZ\right], \tag{99}$$

and, with $a_n := 16n^2 + 2n - 1$,

$$cn\|Z\|_F^2 \leq \langle Z, H_i[Z]\rangle_F \leq ca_n\|Z\|_F^2. \tag{100}$$

Each f_i is globally Frobenius-smooth with constant ca_n and cn-strongly convex, is coercive, and has the unique minimizer 0. Moreover, $f_i(X) = \frac{1}{2}\|\mathcal{A}_i(X)\|_2^2$ for a linear map $\mathcal{A}_i : \mathbb{R}^{p\times q} \to \mathbb{R}^{1+pq}$, and all functions f_i share this minimizer. The full objective has the same strong-convexity, coercivity, minimizer, and least-squares properties; in particular,

$$H := \nabla^2 f = \frac{1}{n}\sum_{i=1}^n H_i \succeq cnI.$$

Proof. Differentiating the quadratic formula in (10) gives $\nabla f_i(X) = c[(u_i^\top \operatorname{diag}(X))\operatorname{Diag}(u_i) + nX]$. Differentiating once more in the direction Z gives (99). Taking the Frobenius inner product with Z gives

$$\langle Z, H_i[Z]\rangle_F = c\left(u_i^\top \operatorname{diag}(Z)\right)^2 + cn\|Z\|_F^2.$$

Since $c > 0$, the square term is nonnegative, which proves the lower bound. By Cauchy–Schwarz and the fact that the diagonal entries form a subset of all matrix entries,

$$|u_i^\top \operatorname{diag}(Z)| \leq \|u_i\|_2\|\operatorname{diag}(Z)\|_2 \leq \|u_i\|_2\|Z\|_F.$$

There is one coordinate of size $4n$ and $n-1$ coordinates of size one in u_i, so $\|u_i\|_2^2 = 16n^2 + n - 1$. Substitution proves the upper bound with $a_n = n + \|u_i\|_2^2$. The Hessian is self-adjoint; its eigenvalues therefore lie in $[cn, ca_n]$. This gives cn-strong convexity and, since $\nabla f_i(X) - \nabla f_i(Y) = H_i[X - Y]$, a global Frobenius gradient Lipschitz constant ca_n.

The formula for the function itself gives $f_i(X) \geq (cn/2)\|X\|_F^2 \geq 0$, with equality only at $X = 0$. Thus 0 is the unique minimizer, and $f_i(X) \to \infty$ as $\|X\|_F \to \infty$, establishing coercivity. Finally define

$$\mathcal{A}_i(X) := \begin{pmatrix} \sqrt{c}\, u_i^\top \operatorname{diag}(X) \\ \sqrt{cn}\, \operatorname{vec}(X)\end{pmatrix},$$

where $\operatorname{vec}(X)$ stacks the entries of X into a vector. This map is linear, and

$$\frac{1}{2}\|\mathcal{A}_i(X)\|_2^2 = \frac{c}{2}\left(u_i^\top \operatorname{diag}(X)\right)^2 + \frac{cn}{2}\|\operatorname{vec}(X)\|_2^2 = f_i(X).$$

The equality $\|\operatorname{vec}(X)\|_2 = \|X\|_F$ proves the claim. Since each f_i has the unique minimizer 0, their minimizer sets intersect in $\{0\}$, giving interpolation. Averaging the Hessian bounds gives $H \succeq cnI$, so f is cn-strongly convex. Averaging the function lower bounds gives $f(X) \geq (cn/2)\|X\|_F^2$, and the stacked linear map

$$\mathcal{A}(X) := n^{-1/2}\begin{pmatrix}\mathcal{A}_1(X)\\ \vdots \\ \mathcal{A}_n(X)\end{pmatrix}$$

satisfies $f(X) = \frac{1}{2}\|\mathcal{A}(X)\|_2^2$. Thus f is also a homogeneous least-squares objective, is coercive, and has the unique minimizer 0. □

G.2 Full Hessian spectrum and condition number

Lemma G.2 (Conditioning of the full Hessian)**.** *The Hessian H of the full objective satisfies $\kappa_F(H) < 19$.*

Proof. Decompose the matrix space into the orthogonal subspaces

$$\begin{aligned}\mathcal{S}_0 &:= \{Z : Z_{aa} = 0 \text{ for } 1 \le a \le n\},\\ \mathcal{S}_1 &:= \operatorname{span}\{P_n\},\\ \mathcal{S}_\perp &:= \{\operatorname{Diag}(z_1, \dots, z_n, 0, \dots, 0) : \textstyle\sum_{a=1}^n z_a = 0\}.\end{aligned}$$

Here each diagonal vector has $r - n$ trailing zeros, with the trailing block absent when $r = n$. Every $p \times q$ matrix is the sum of its active diagonal part and a matrix in $\mathcal{S}_0$. Its active diagonal vector further decomposes into a multiple of $\mathbf{1} \in \mathbb{R}^n$ and a vector orthogonal to $\mathbf{1}$. These observations prove that the three subspaces span the matrix space and are mutually orthogonal, with dimensions $pq - n$, 1, and $n - 1$, respectively.

If $Z \in \mathcal{S}_0$, then $u_i^\top \operatorname{diag}(Z) = 0$ for every i, so (99) gives $H[Z] = cnZ$. On the active diagonal subspace, collect the first n coordinates of $u_i^\top$ as the rows of

$$U_n = (4n + 1)I_n - \mathbf{1}\mathbf{1}^\top.$$

Writing $Z = \operatorname{Diag}(z_1, \dots, z_n, 0, \dots, 0)$, averaging (99) shows that the active diagonal vector of $H[Z]$ equals $c(U_n^\top U_n/n + nI_n)z$. Since

$$U_n\mathbf{1} = (3n + 1)\mathbf{1}, \qquad U_n z = (4n + 1)z \quad \text{if } \mathbf{1}^\top z = 0,$$

and U_n is symmetric, $U_n^\top U_n$ has eigenvalues $(3n+1)^2$ and $(4n+1)^2$ on those two subspaces. Thus H acts by $c[n + (3n + 1)^2/n]$ on $\mathcal{S}_1$ and by $c[n + (4n + 1)^2/n]$ on $\mathcal{S}_\perp$. All three subspaces are nonzero because $r \ge n \ge 8$, so the smallest eigenvalue is cn and the largest is $c[n + (4n + 1)^2/n]$. Their ratio is

$$\kappa_F(H) = \frac{c[n + (4n + 1)^2/n]}{cn} = 17 + \frac{8}{n} + \frac{1}{n^2} \le 17 + 1 + \frac{1}{64} < 19.$$

□

G.3 Assumptions and function-class membership

Lemma G.3 (Global component-wise smoothness and finite-sum growth)**.** *For every $i \in \{1, \dots, n\}$ and all $X, Y \in \mathbb{R}^{p\times q}$,*

$$\begin{aligned}\|\nabla f_i(X) - \nabla f_i(Y)\|_{\mathrm{nuc}} &\le c\big[(5n - 1)^2 + nr\big]\|X - Y\|_{\mathrm{op}},\\ \sum_{i=1}^n \|\nabla f_i(X)\|_F^2 &\le (16n^2 + 2n - 1)\|\nabla f(X)\|_F^2.\end{aligned} \tag{101}$$

Consequently, if the prescribed class parameters satisfy $c[(5n - 1)^2 + nr] \le L$, $D_1 \ge 16n^2 + 2n - 1$, and $D_0 \ge 0$, then $f \in \mathcal{F}^{n;p,q}_{L,D_0,D_1}$. In particular, choosing c by (9) satisfies the smoothness requirement for any prescribed $L > 0$.

Proof. Let $Z = X - Y$. For $1 \le a \le r$, let $e_a^{(p)}$ and $e_a^{(q)}$ be the coordinate unit vectors in the respective spaces. Each diagonal entry obeys $|Z_{aa}| = |(e_a^{(p)})^\top Z e_a^{(q)}| \le \|Z\|_{\mathrm{op}}$. Consequently,

$$|u_i^\top \operatorname{diag}(Z)| \le \|u_i\|_1\|Z\|_{\mathrm{op}}, \qquad \|\operatorname{Diag}(u_i)\|_{\mathrm{nuc}} = \|u_i\|_1 = 5n - 1.$$

The r singular values of Z are each at most $\|Z\|_{\mathrm{op}}$, so $\|Z\|_{\mathrm{nuc}} \le r\|Z\|_{\mathrm{op}}$. Using $\nabla f_i(X) - \nabla f_i(Y) = H_i[Z]$, (99), and the triangle inequality gives

$$\begin{aligned}\|\nabla f_i(X) - \nabla f_i(Y)\|_{\mathrm{nuc}} &\le c|u_i^\top \operatorname{diag}(Z)|\|\operatorname{Diag}(u_i)\|_{\mathrm{nuc}} + cn\|Z\|_{\mathrm{nuc}}\\ &\le c\big((5n - 1)^2 + nr\big)\|X - Y\|_{\mathrm{op}}.\end{aligned}$$

This is the required component-wise operator-to-nuclear bound on the whole space.

For the growth bound, write $a_n = 16n^2 + 2n - 1$ as in Lemma G.1. We use inequalities for self-adjoint operators on the Frobenius space: $A \preceq B$ means $\langle Z, (B - A)[Z] \rangle_F \geq 0$ for every Z. Each eigenvalue λ of H_i lies in $[cn, ca_n]$, so $\lambda^2 \leq ca_n\lambda$. Expanding an arbitrary X in an orthonormal eigenbasis of H_i therefore gives

$$\left\langle X, H_i^2[X] \right\rangle_F \leq ca_n \left\langle X, H_i[X] \right\rangle_F .$$

Likewise Lemma G.1 gives that every eigenvalue of H is at least cn, so expansion in an orthonormal eigenbasis of H gives

$$\left\langle X, H^2[X] \right\rangle_F \geq cn \left\langle X, H[X] \right\rangle_F .$$

No common eigenbasis for the different H_i is required for these separate inequalities. Since the functions are homogeneous quadratics, $\nabla f_i(X) = H_i[X]$ and $\nabla f(X) = H[X]$. Their Hessians are self-adjoint, and $\sum_i H_i = nH$. Thus

$$\begin{aligned} \sum_{i=1}^{n} \|\nabla f_i(X)\|_F^2 &= \sum_{i=1}^{n} \left\langle X, H_i^2[X] \right\rangle_F \\ &\leq ca_n \sum_{i=1}^{n} \left\langle X, H_i[X] \right\rangle_F = ca_n n \left\langle X, H[X] \right\rangle_F \\ &\leq a_n \left\langle X, H^2[X] \right\rangle_F = a_n \|\nabla f(X)\|_F^2. \end{aligned} \tag{102}$$

The factor c cancels between the component-Hessian upper bound and the full-Hessian lower bound, giving the unnormalized affine growth bound with coefficient a_n and zero offset. Comparing these bounds with the prescribed L, D_0, D_1 proves Assumptions 2.1 and 2.3. Attainment of a finite minimum and continuous differentiability follow from Lemma G.1, completing membership in the stated class. □

H Sufficient Momentum Thresholds: Separation Proofs

Proof of Proposition 3.6. Lemma C.10 proves existence and uniqueness of $\beta_+^{\mathrm{perm}} \in (0,1)$ and strict decrease of $c_{\beta,n}$ in β. The definition gives $0 < \beta_- < 1$, and the geometric identity yields

$$c_{\beta,n} = \frac{n}{\sum_{j=0}^{n-1} \beta^j} - \beta^n \geq 1 - \beta^n \qquad (0 < \beta < 1).$$

Since $n/\lfloor n/8 \rfloor \geq 8$, we have $\beta_-^n \leq (70n)^{-8} < 1/2$. Consequently,

$$c_{\beta_-,n} \geq 1 - (70n)^{-8} > \frac{1}{2} \geq \frac{1}{2\sqrt{r}(\sqrt{D_1}+1)} = c_{\beta_+^{\mathrm{perm}},n}.$$

Strict decrease in β implies $\beta_- < \beta_+^{\mathrm{perm}}$. □

Proof of Proposition 3.7. Put $K = \lfloor n/8 \rfloor$ and $A = 4C_D^2 r(\sqrt{D_1}+1)^2 \geq 4n$. Then $\beta_+^{\mathrm{i.i.d.}} = 1 - 2/(A+1) \in (0,1)$. Bernoulli's inequality gives

$$(\beta_+^{\mathrm{i.i.d.}})^K \geq 1 - \frac{2K}{A+1} > \frac{3}{4} > \frac{1}{70n} = \beta_-^K.$$

Here $2K/(A+1) \leq (n/4)/(4n+1) < 1/4$. Taking positive Kth roots proves the strict ordering. □

H.1 Dependence of the sufficient thresholds on components and batch size

The low-momentum threshold β_-. As the number of components increases,

$$\beta_-(n) \longrightarrow 1, \qquad 1 - \beta_-(n) \sim \frac{8\log(70n)}{n} \quad \text{as } n \to \infty. \tag{103}$$

Thus the certified divergence interval can reach arbitrarily close to 1 across admissible problems with growing n. The floor in β_- prevents monotonicity at consecutive integers. This limit is realized by the witness family with $r \geq n$ and $D_1 \geq 16n^2 + 2n - 1$.

The i.i.d. high-momentum threshold $\beta_+^{\mathrm{i.i.d.}}$. Solving its defining equation gives

$$\beta_+^{\mathrm{i.i.d.}}(r, D_1) = 1 - \frac{2}{4C_D^2 r(\sqrt{D_1}+1)^2 + 1}.$$

Hence, at fixed r, D_1, this sufficient threshold is independent of n.

The permutation high-momentum threshold β_+^{perm}. At fixed r, D_1, this threshold is strictly increasing in n by Lemma C.10.

Relation to batch size. For a fixed dataset size N and equal batches of size B, $n = N/B$. Holding r, D_1 fixed, larger batches lower the permutation high-momentum threshold and leave the i.i.d. threshold unchanged. Rebatching may change D_1 along with the finite-sum decomposition.

I I.I.D. CONVERGENCE: PROOF IDEA

Under the setup of Theorem 3.2, we first bound the weighted average of expected gradient norms. Since the stepsizes are positive and deterministic,

$$\min_{0\le t<T} \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}} \le \frac{\sum_{t=0}^{T-1} \eta_t \mathbb{E}\|\nabla f(X_t)\|_{\mathrm{nuc}}}{\sum_{t=0}^{T-1} \eta_t}.$$

It therefore suffices to control the expected gradient norm at each step in a form that can be summed with these weights; Section D.3 carries out this reduction.

We seek a bound that controls each weighted expected gradient norm by the expected one-step objective decrease plus controlled error terms. Summing makes the expected objective differences telescope, and lower boundedness of the objective bounds their total. Proposition D.2 establishes the required one-step inequality and sums and normalizes it.

To obtain such a one-step bound, use the polar-factor property in Lemma B.3,

$$\|\mathrm{Polar}(M_t)\|_{\mathrm{op}} \le 1.$$

The descent lemma (Lemma B.5) applies to the Muon update on every realized trajectory under full-objective smoothness. Taking expectations gives

$$\begin{aligned}\mathbb{E}f(X_{t+1}) - \mathbb{E}f(X_t) &\le -\eta_t \mathbb{E}\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F + \frac{L}{2}\eta_t^2 \mathbb{E}\|\mathrm{Polar}(M_t)\|_{\mathrm{op}}^2 \\ &\le -\eta_t \mathbb{E}\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F + \frac{L}{2}\eta_t^2.\end{aligned}$$

Thus the expected polar alignment must exceed a positive multiple of the expected gradient norm, up to controlled errors. Section D.3 makes this threshold precise.

The following inequality holds on each realized trajectory and turns the alignment requirement into a bound on the distance between the momentum and the full gradient.

Lemma I.1. *For every $t \ge 0$,*

$$\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F \ge \|\nabla f(X_t)\|_{\mathrm{nuc}} - 2\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}}. \tag{104}$$

Proof. Split the full gradient into M_t and its difference from M_t. Lemma B.3 identifies the momentum inner product with its nuclear norm; the nuclear–operator Hölder inequality (Lemma B.2) bounds the remaining inner product. Using the polar norm bound and then the triangle inequality gives

$$\begin{aligned}\left\langle \nabla f(X_t), \mathrm{Polar}(M_t)\right\rangle_F &= \left\langle M_t, \mathrm{Polar}(M_t)\right\rangle_F + \left\langle \nabla f(X_t) - M_t, \mathrm{Polar}(M_t)\right\rangle_F \\ &\ge \|M_t\|_{\mathrm{nuc}} - \|\nabla f(X_t) - M_t\|_{\mathrm{nuc}}\|\mathrm{Polar}(M_t)\|_{\mathrm{op}} \\ &\ge \|M_t\|_{\mathrm{nuc}} - \|\nabla f(X_t) - M_t\|_{\mathrm{nuc}} \\ &\ge \big(\|\nabla f(X_t)\|_{\mathrm{nuc}} - \|\nabla f(X_t) - M_t\|_{\mathrm{nuc}}\big) - \|\nabla f(X_t) - M_t\|_{\mathrm{nuc}} \\ &= \|\nabla f(X_t)\|_{\mathrm{nuc}} - 2\|M_t - \nabla f(X_t)\|_{\mathrm{nuc}}.\end{aligned}$$

□

Taking expectations in this inequality reduces the task to an expected tracking bound in which the coefficient of the expected full-gradient norm is strictly smaller than one half, with controlled remaining terms. Proposition D.2 and the high-momentum condition in Theorem 3.2 supply this bound.

Unrolling the momentum recursion separates the tracking error into three terms on every realized trajectory (50):

$$M_t - \nabla f(X_t) = \underbrace{\beta^{t+1}(M_{-1} - \nabla f(X_0))}_{\text{initialization error}} + \underbrace{\sum_{s=1}^{t} \beta^{t-s+1}\big(\nabla f(X_{s-1}) - \nabla f(X_s)\big)}_{\mathcal{L}_t\text{: gradient lag}} + \underbrace{(1-\beta)\sum_{s=0}^{t} \beta^{t-s}\big(\nabla f_{i_s}(X_s) - \nabla f(X_s)\big)}_{\mathcal{C}_t\text{: component discrepancy}}.$$

The triangle inequality followed by expectation reduces the task to bounding the expected sizes of the three terms. The initialization error decays geometrically, and full-gradient smoothness together with the bounded update length gives an $O(\eta_t)$ gradient lag (Section D.1, (55)). The remaining task is to bound the expected size of the component discrepancy $\mathcal{C}_t$.

Given all past samples, averaging the next sampled gradient over its possible choices gives the full gradient. Each new sampling error therefore has mean zero given the past, although the size of their accumulated sum need not vanish. The proof controls its expected size by adding the squared errors with squared momentum weights, taking a square root, and then an expectation. For large momentum, these weights are spread over many updates, reducing the resulting error coefficient. The growth assumption bounds the individual errors, and smoothness controls the change from earlier full gradients to the current one.

Section D.2 gives the detailed component-discrepancy bound in Lemma D.1. Section D.3 combines the three estimates and turns the resulting expected tracking bound into Theorem 3.2.

J I.I.D. DIVERGENCE: PROOF IDEA

Fix $\varepsilon \in (0,1)$. Throughout, use deterministic stepsizes satisfying Assumption 3.3. We prove the existence claim in Theorem 3.5 through the shared quadratic witness in (10), with the fixed scale $c > 0$ chosen by (9). Lemma G.3 verifies its membership in the prescribed class. The common positive factor c multiplies the gradient and momentum magnitude estimates and does not affect the sign conditions driving escape. The mechanism is that polar normalization removes the magnitude advantage of a rare correcting gradient, while a short momentum memory allows that correction to be forgotten. The proof must show that this effect persists along the actual moving trajectory on one i.i.d. sampling event of probability at least $1-\varepsilon$, simultaneously for every $\beta \in [0, \beta_-]$. Its structure is

$$\begin{aligned} \text{invariant active coordinates} &\implies \text{forgetting after a short window} \\ &\implies \text{few bad updates} \implies \text{escape.} \end{aligned}$$

Appendix F proves the i.i.d. dynamical claims for this witness and deduces Theorem 3.5, and Appendix G verifies Assumptions 2.1 and 2.3 and the other properties of the quadratic example.

J.1 WHY INDIVIDUAL POLAR DIRECTIONS CAN POINT OUTWARD

Consider the functions f_i, f in (10) at $X = sP_n$ with $s > 0$. Since $u_i^\top \operatorname{diag}(sP_n) = (3n+1)s$, the ith active entry of $\nabla f_i(sP_n)$ is positive, whereas every other active entry equals $-c(2n+1)s$. The inactive entries are zero. Let $E_{ii} := e_i^{(p)}(e_i^{(q)})^\top \in \mathbb{R}^{p\times q}$, where the two vectors are the coordinate unit vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Since $c > 0$, the rectangular diagonal polar identity in Lemma B.4 gives

$$\operatorname{Polar}(\nabla f_i(sP_n)) = 2E_{ii} - P_n.$$

The full gradient and the average of the polar directions of ∇f_i consequently satisfy

$$\begin{aligned} \nabla f(sP_n) &= c\left(n + \frac{(3n+1)^2}{n}\right) sP_n, \\ \operatorname{Polar}(\nabla f(sP_n)) &= P_n, \\ \frac{1}{n}\sum_{i=1}^{n} \operatorname{Polar}(\nabla f_i(sP_n)) &= \left(\frac{2}{n} - 1\right) P_n. \end{aligned}$$

The coefficient in the last line is negative for $n > 2$. Thus subtracting the average of these polar directions points outward, even though subtracting the full-gradient polar direction points toward the common minimizer. This static calculation motivates the example. The proof of divergence must additionally control momentum and show that moving away from the exact ray does not destroy the sign pattern.

J.2 CONTROLLING THE ACTUAL MOMENTUM ON AN EXPANDING CONE

Starting from $X_0 = s_0 P_n$, the iterates, component gradients, and momentum matrices are supported only on the first n diagonal positions in $\mathbb{R}^{p\times q}$. The gradient formula preserves this support, the momentum recursion preserves it, and Lemma B.4 proves that the rectangular polar factor preserves it. Thus the other entries remain zero under the ordinary Muon update. Write $x_{t,a}$ and $m_{t,a}$ for an active iterate and momentum coordinate. Then

$$x_{t+1,a} = x_{t,a} - \eta_t \operatorname{sign}(m_{t,a}), \qquad \operatorname{sign}(0) = 0.$$

Every coordinate moves by at most η_t. Define

$$\Gamma_t := \sum_{s<t} \eta_s, \qquad \rho_t := s_0 + \Gamma_t.$$

The proof seeks to keep every active coordinate in $[2\rho_t/3, \rho_t]$. Throughout this cone, $\nabla f_a(X_t)$ has a positive entry in coordinate a, while each $\nabla f_i(X_t)$ with $i \neq a$ has a negative entry in that coordinate, of magnitude at least $c(2n/3+1)\rho_t$. All component-gradient entries have magnitude at most $17cn^2\rho_t$.

If index a has not been selected in the last $\lfloor n/8 \rfloor$ updates, those updates supply only negative inputs to its momentum. The initialization condition $s_0 \geq 8\lfloor n/8 \rfloor \eta_0$ ensures that the radii at these recent times are at least $7/8$ of the current radius. Consequently each recent negative input is at most $-c(2n/3+1)(7/8)\rho_t \leq -cn\rho_t/2$. The contribution of the older momentum is multiplied by $\beta^{\lfloor n/8 \rfloor}$. Unrolling the recursion therefore yields the bound

$$m_{t,a} \leq c\left[17n^2\beta^{\lfloor n/8 \rfloor} - \frac{n}{2}(1-\beta^{\lfloor n/8 \rfloor})\right]\rho_t.$$

For every $\beta \leq \beta_-$ we have $\beta^{\lfloor n/8 \rfloor} \leq (70n)^{-1}$, making the bracket strictly negative. Since $c > 0$, the actual momentum is negative as well. Such an update moves coordinate a outward by exactly η_t. At early times with fewer than $\lfloor n/8 \rfloor$ previous updates, the same conclusion follows directly from the zero momentum buffer $M_{-1} = 0$ if index a has not yet been selected. This is the finite-memory forgetting argument in Lemma F.5.

J.3 Counting the remaining updates and closing the argument

Call an update bad for coordinate a if index a is selected in its most recent $K := \lfloor n/8 \rfloor$ positions; all other updates are good. A single selection of index a contaminates at most K updates, including itself. Nonincreasing stepsizes bound the weight of its contamination window by K times the stepsize at that occurrence. If $v_{t,a} := \sum_{s<t} \eta_s \mathbf{1}\{i_s = a\}$ and $w_{t,a}$ is the total stepsize weight of bad updates before time t, Lemma F.6 gives $w_{t,a} \leq K v_{t,a}$.

Independent uniform sampling gives $\mathbb{E}[K\mathbf{1}\{i_t = a\} - 1/6] = K/n - 1/6 \leq -1/24$. The bounded-variable exponential-moment estimate (Hoeffding, 1963) makes

$$\exp\left(\frac{Kv_{t,a} - \Gamma_t/6}{6\eta_0 K^2}\right)$$

a nonnegative supermartingale. Its maximal inequality (Howard et al., 2020), followed by a union bound over the n coordinates, supplies an event $\mathcal{E}_\varepsilon$ of probability at least $1-\varepsilon$ on which

$$w_{t,a} \leq \frac{1}{6}\Gamma_t + u_\varepsilon \quad \text{for every } t, a, \qquad u_\varepsilon := 6\eta_0 K^2 \log(n/\varepsilon).$$

Lemma F.7 proves this time-uniform statement. The event uses only the sample indices and the prescribed stepsizes; it does not depend on β. Choose $s_0 \geq \max\{8K\eta_0, 6u_\varepsilon\}$ and fix any sample path in this event.

Under the cone history, every good update adds η_s to the coordinate, and a bad update can subtract at most η_s. Relative to an all-good trajectory, each bad update therefore loses at most $2\eta_s$. Summing yields

$$\begin{aligned} x_{t,a} &\geq s_0 + \Gamma_t - 2w_{t,a} \\ &\geq s_0 - 2u_\varepsilon + \frac{2}{3}\Gamma_t \geq \frac{2}{3}(s_0 + \Gamma_t). \end{aligned}$$

The last step uses $s_0 \geq 6u_\varepsilon$. The upper bound $x_{t,a} \leq s_0 + \Gamma_t$ follows from the size of each coordinate step. To avoid assuming what is to be proved, apply these inequalities at a hypothetical first exit from the cone: every earlier update satisfies the required cone history, and the inequalities show that the alleged exiting iterate is still inside. Thus no first exit exists. Since the event is independent of β and the forgetting bound is uniform over $[0, \beta_-]$, the same event works simultaneously for every low-momentum parameter.

Assumption 3.3 gives $\Gamma_t \to \infty$, so every active coordinate and $\|X_t\|_F$ diverge. The global bounds

$$f(X) \geq \frac{cn}{2}\|X\|_F^2, \qquad \|\nabla f(X)\|_F \geq cn\|X\|_F$$

from Lemma G.1 give objective and gradient divergence because $c > 0$ is fixed. The global regularity and growth calculations in Appendix G place this same objective in the class covered by Theorem 3.2.

J.4 WHY THE THRESHOLD INVOLVES THE NUMBER OF COMPONENTS

Under the stated initialization condition, the window conditions $\beta^{\lfloor n/8 \rfloor} \leq (70n)^{-1}$ and $\lfloor n/8 \rfloor / n \leq 1/8$ control forgetting and the mean frequency of correcting inputs, respectively. Each index is sampled with probability $1/n$ at every update. A window of $\lfloor n/8 \rfloor$ updates gives the sufficient endpoint β_-. The positive factor c multiplies both terms of the forgetting estimate, so it does not enter this sufficient threshold. These conditions involve the number of components n and no stopping horizon. The embedding requires $p, q \geq n$, equivalently $r \geq n$; the additional rows and columns leave the active-coordinate dynamics unchanged. Thus one fixed problem and stepsize schedule exhibit divergence throughout the stated interval on a common event of probability at least $1 - \varepsilon$. Proposition 3.7 proves that this interval lies strictly below the high-momentum endpoint for i.i.d. sampling.

K ADDITIONAL EXPERIMENTS

This appendix collects the additional experiments on the quadratic construction. The figures retain their original settings, recorded points, and sampling-path summaries. Random reshuffling permutes fixed components within each epoch; i.i.d. sampling draws a component independently at each update. Unless a panel states otherwise, the respective schedules are $\eta_t = 1/\sqrt{\lfloor t/n \rfloor + 1}$ and $\eta_t = 1/\sqrt{t+1}$. Where shown, bands denote the range over sampling paths.

K.1 FIXED OBJECTIVES, BATCH SIZE, AND MOMENTUM

The first group fixes the elementary losses before grouping them into batches. Thus changing B retains the full objective. Figure 8 includes momentum values in both certified regimes; the remaining sweeps examine dependence on momentum and batch size.

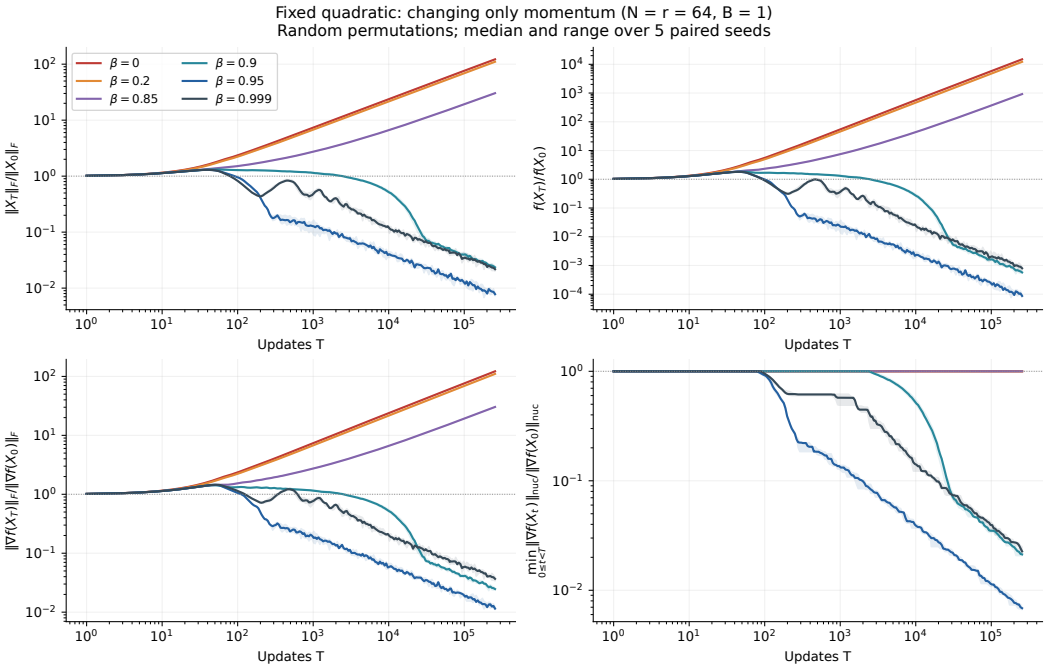

Figure 4: Parameter, objective, and gradient trajectories on the fixed quadratic with $N = r = 64$ and $B = 1$.

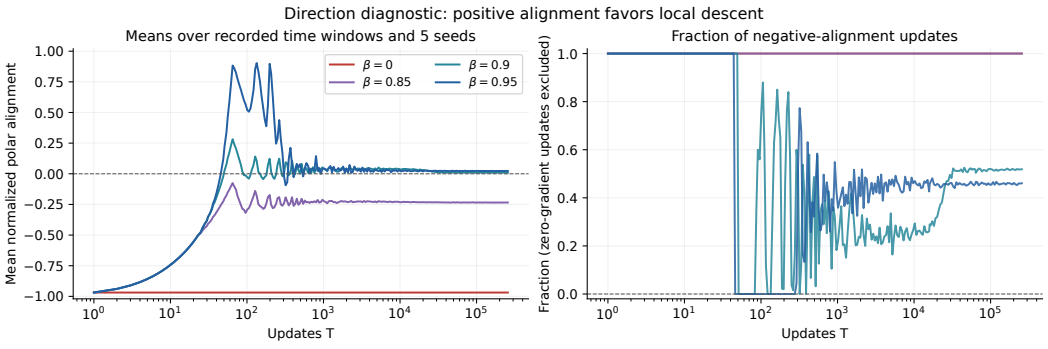

Figure 5: Alignment between the full gradient and the Muon direction on the same fixed quadratic.

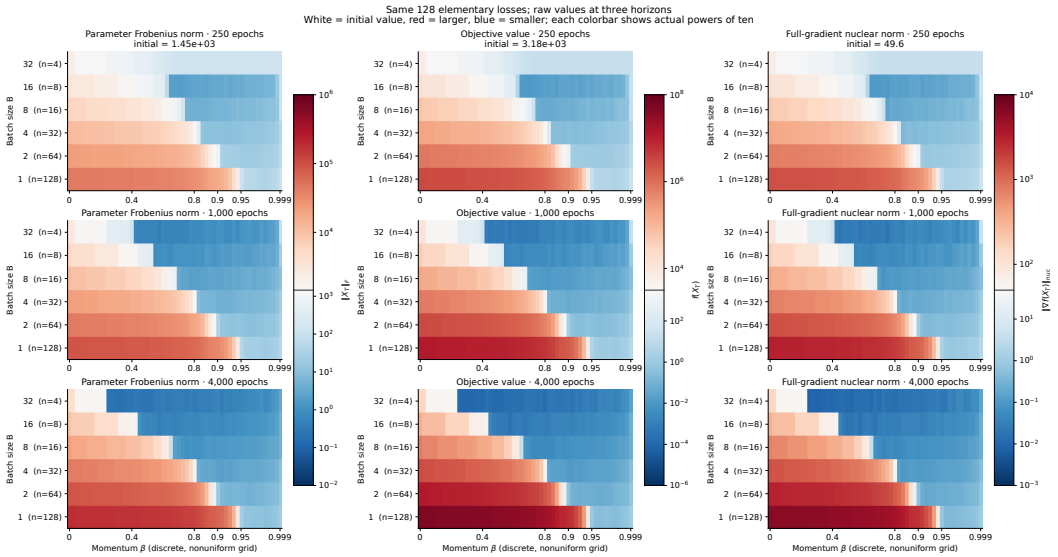

Figure 6: Momentum and batch-size sweeps at several training horizons, holding the elementary losses and full objective fixed.

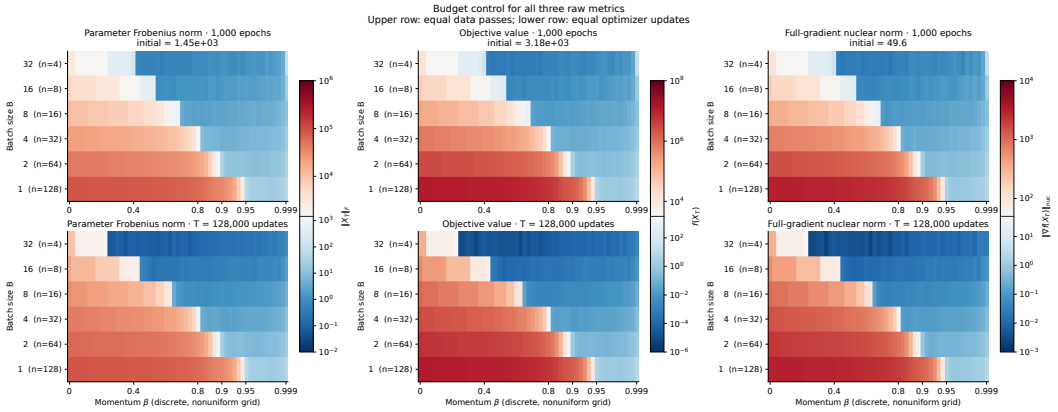

Figure 7: Comparisons at a common epoch budget and at a common update budget.

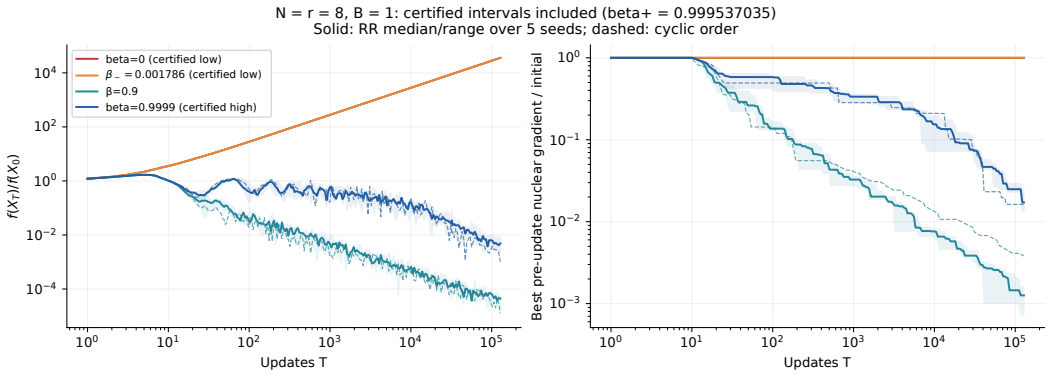

Figure 8: A small permutation instance including momentum values in the sufficient low- and high-momentum ranges.

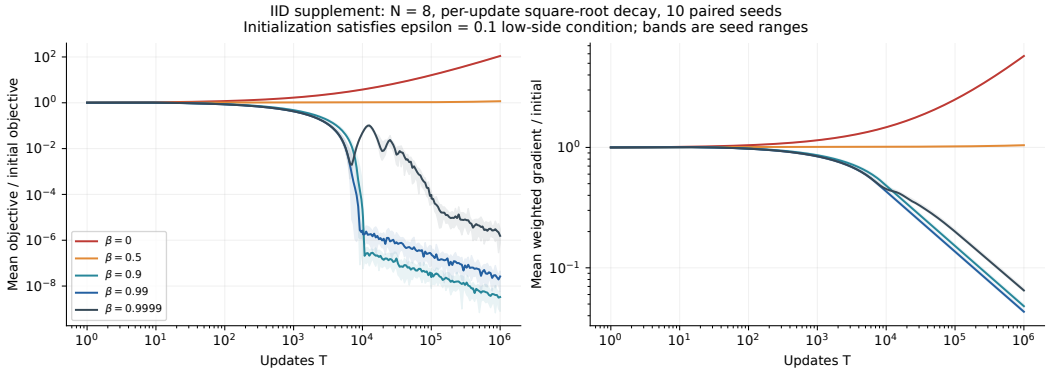

Figure 9: An i.i.d. example with the prescribed initialization and updatewise square-root stepsizes.

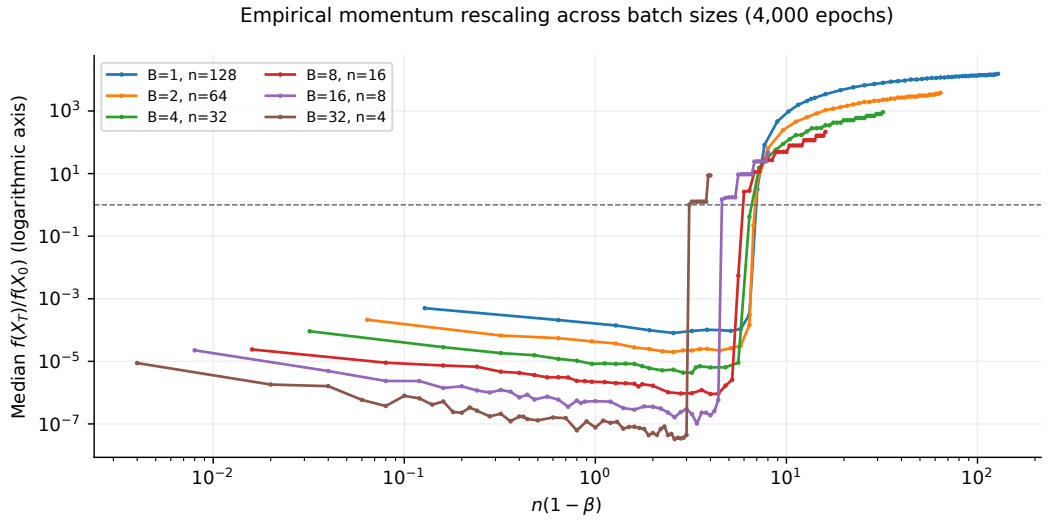

Figure 10: The batch-size sweep after $4{,}000$ epochs, displayed against the rescaled momentum coordinate $n(1-\beta)$.

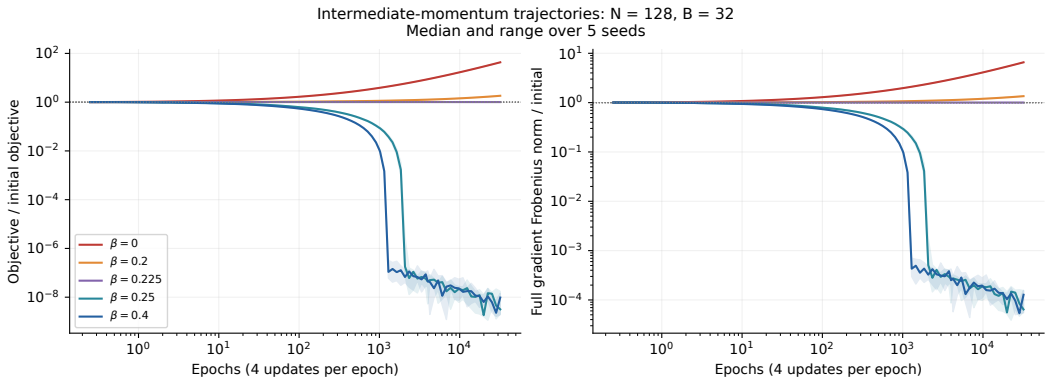

Figure 11: A representative intermediate-momentum trajectory and neighboring momentum values.

K.2 Larger quadratic instances

The following scans use $n = r = 64$. Terminal-loss scans use 10^6 updates. The i.i.d. trajectory panel extends the indicated representative configurations to 10^8 updates, retaining their common shorter prefixes.

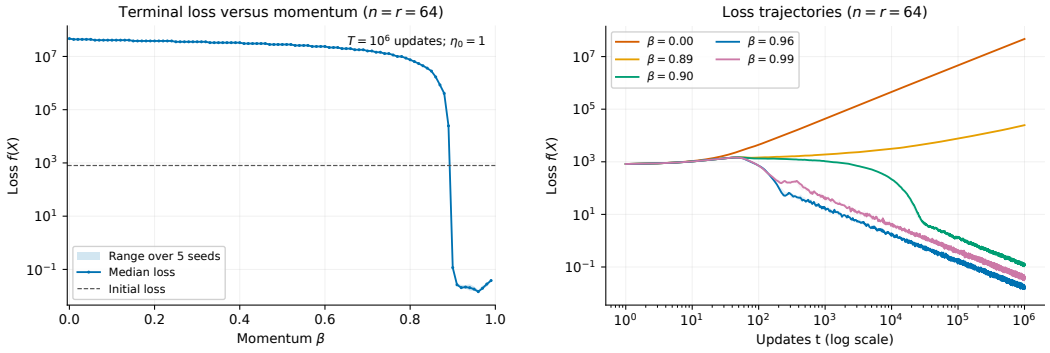

Figure 12: Terminal losses and representative trajectories under random reshuffling, $n = r = 64$.

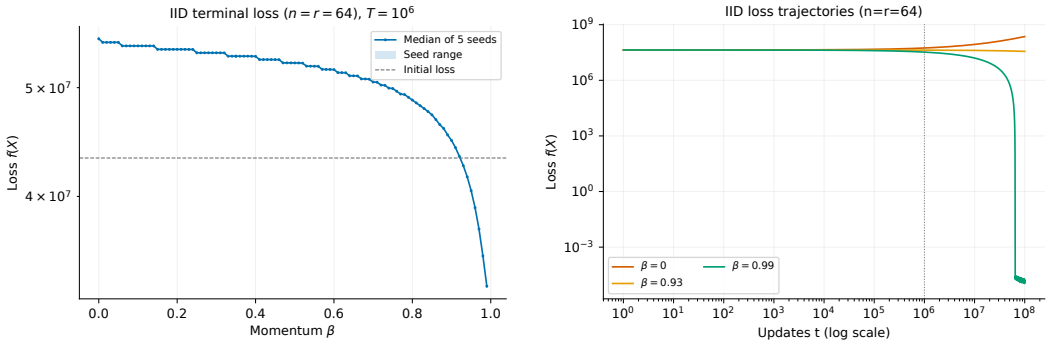

Figure 13: Terminal losses and representative trajectories under i.i.d. sampling, $n = r = 64$.

K.3 Initialization and matrix dimension

The initialization sweep varies the initial diagonal radius relative to the prescribed permutation value. The dimension sweep uses zero padding for $r > n$ and a coordinate-sharing extension for $r < n$. In this extension, set $m = \min\{n, r\}$, replace the active coordinate i in (10) by $q(i) = 1 + ((i - 1) \bmod m)$, sum over the first m diagonal coordinates, and use $c = ((4n + m - 1)^2 + nr)^{-1}$.

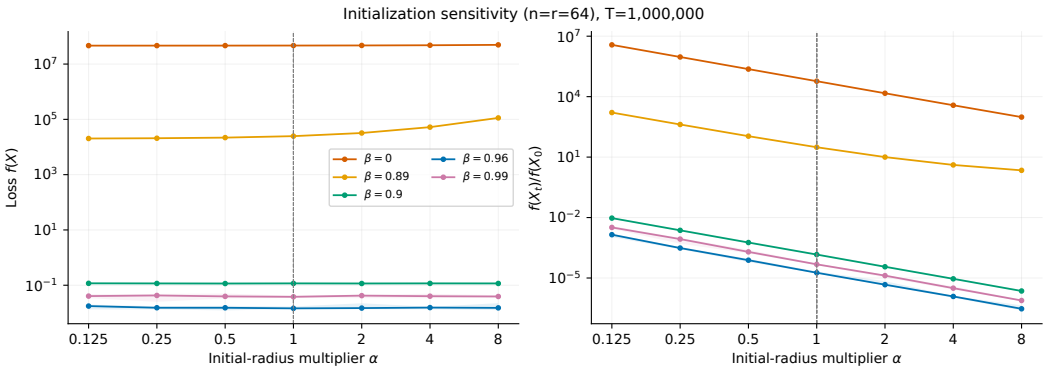

Figure 14: Terminal loss versus initial radius for $n = r = 64$.

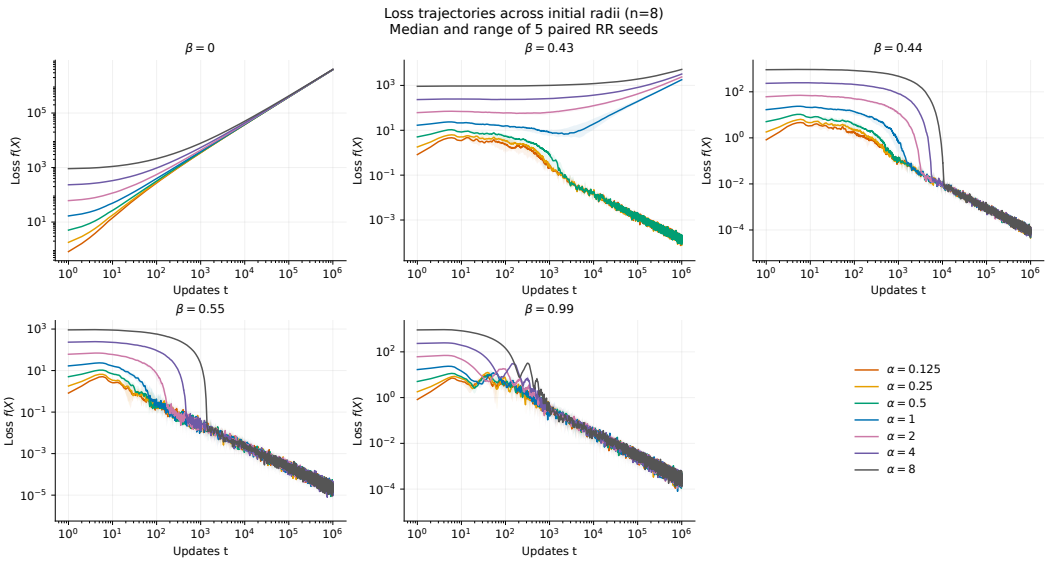

Figure 15: Trajectories at different initial radii for $n = r = 8$.

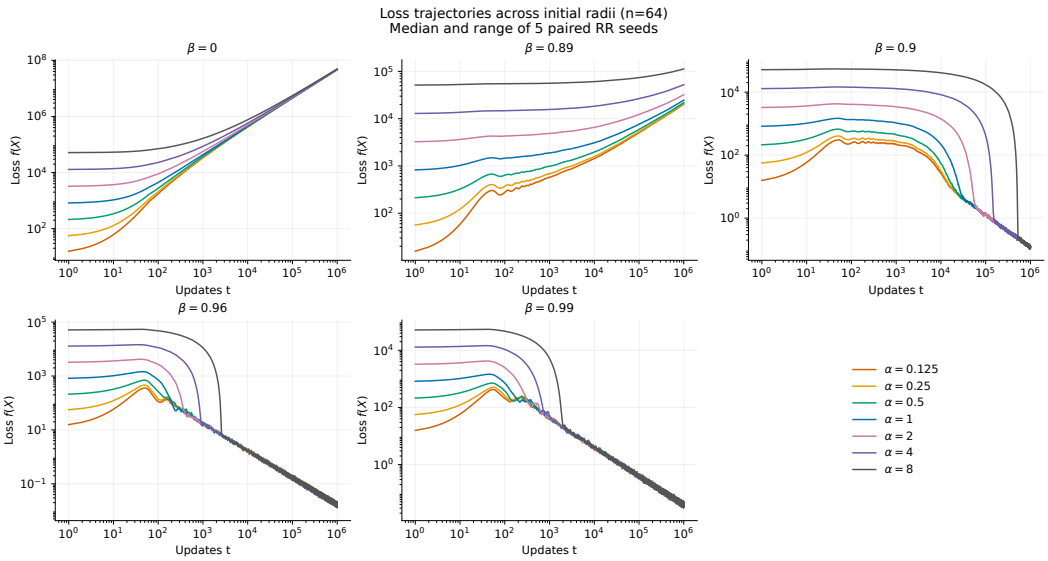

Figure 16: Trajectories at different initial radii for $n = r = 64$.

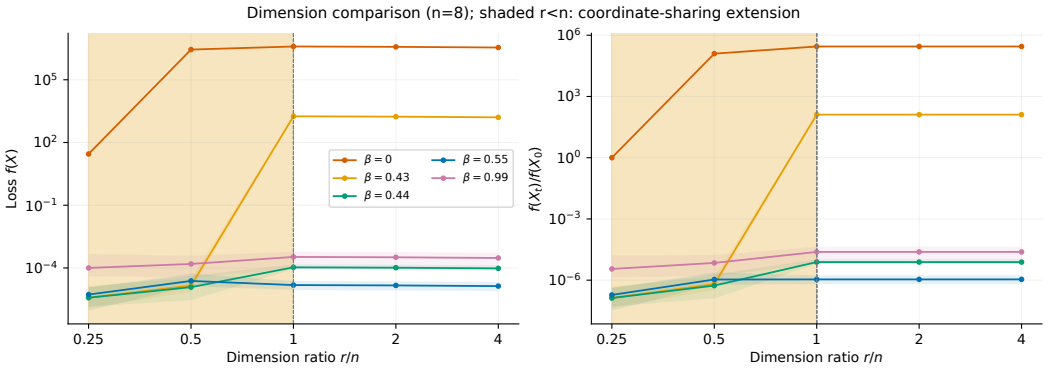

Figure 17: Terminal loss versus r/n for $n = 8$.

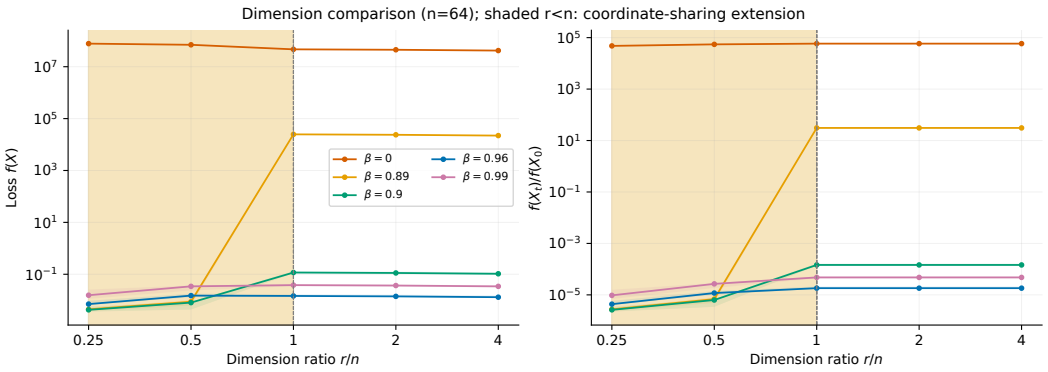

Figure 18: Terminal loss versus r/n for $n = 64$.

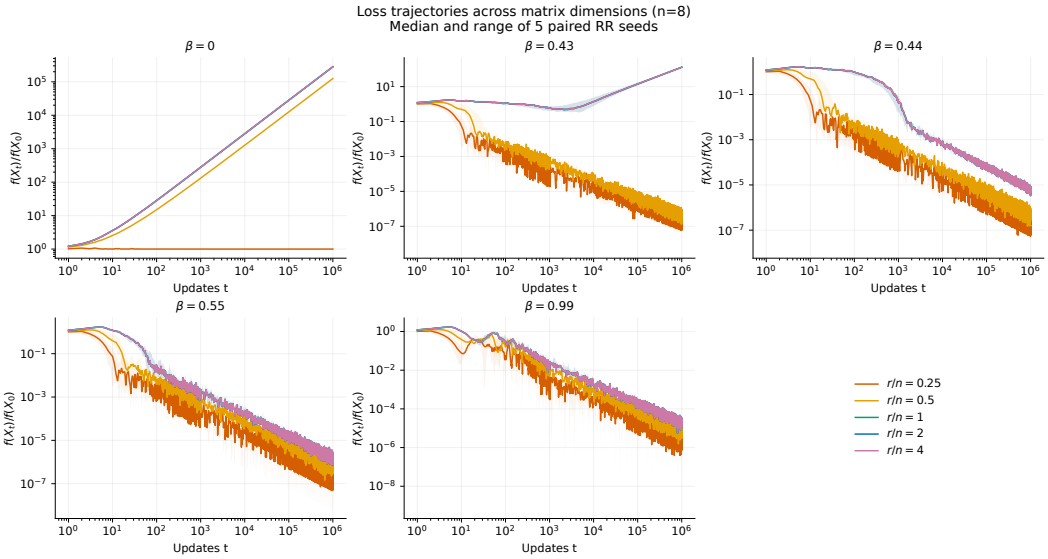

Figure 19: Trajectories at different dimension ratios for $n = 8$.

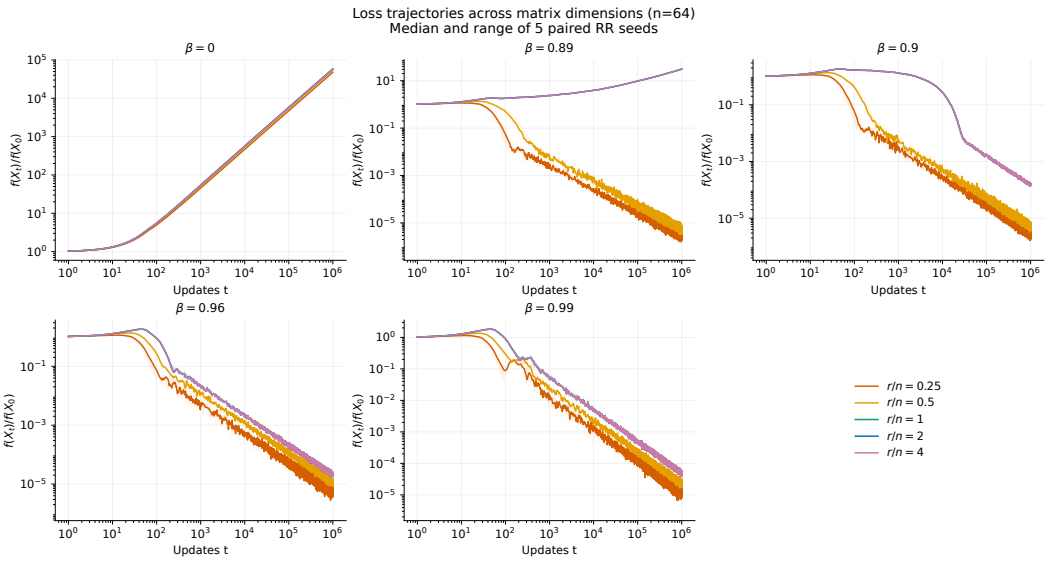

Figure 20: Trajectories at different dimension ratios for $n = 64$.

K.4 Diagnostics of the divergence mechanism

These diagnostics follow the proof's short-memory, weighted-counting, and invariant-region arguments. The low-momentum trajectories use the prescribed initializations and the sufficient low-momentum interval for each problem size.

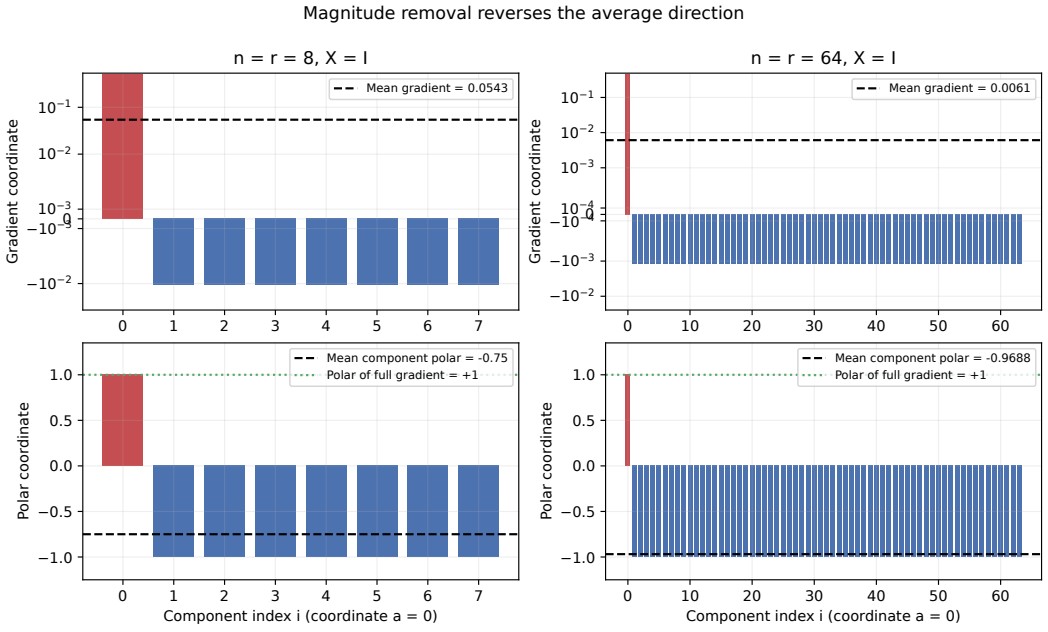

Figure 21: Component-gradient magnitudes and their polar directions at an equal-coordinate point.

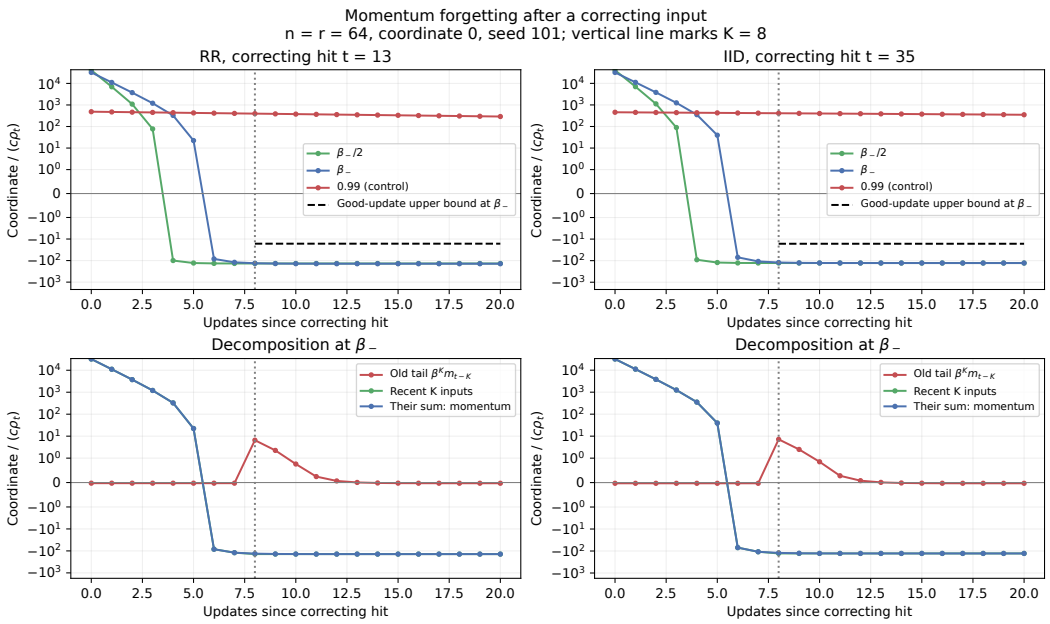

Figure 22: The old momentum contribution and recent inputs after one selected correcting event under each sampling scheme, with $n = r = 64$, seed 101, and coordinate 0. The vertical line marks the window length $K = 8$.

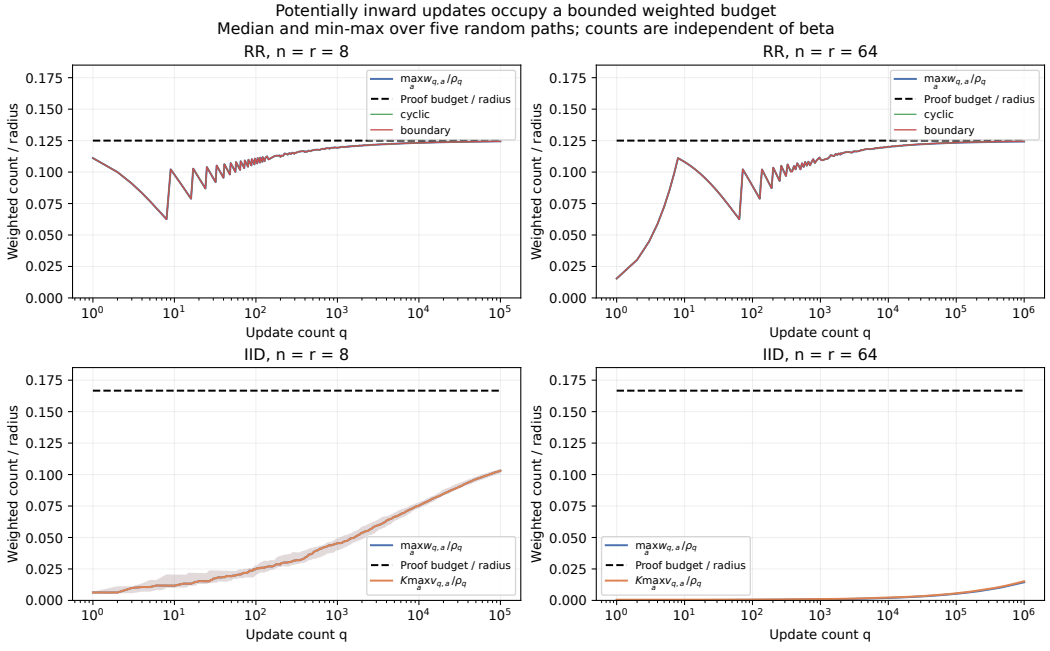

Figure 23: Cumulative stepsize weights of potentially inward updates and the corresponding proof bounds.

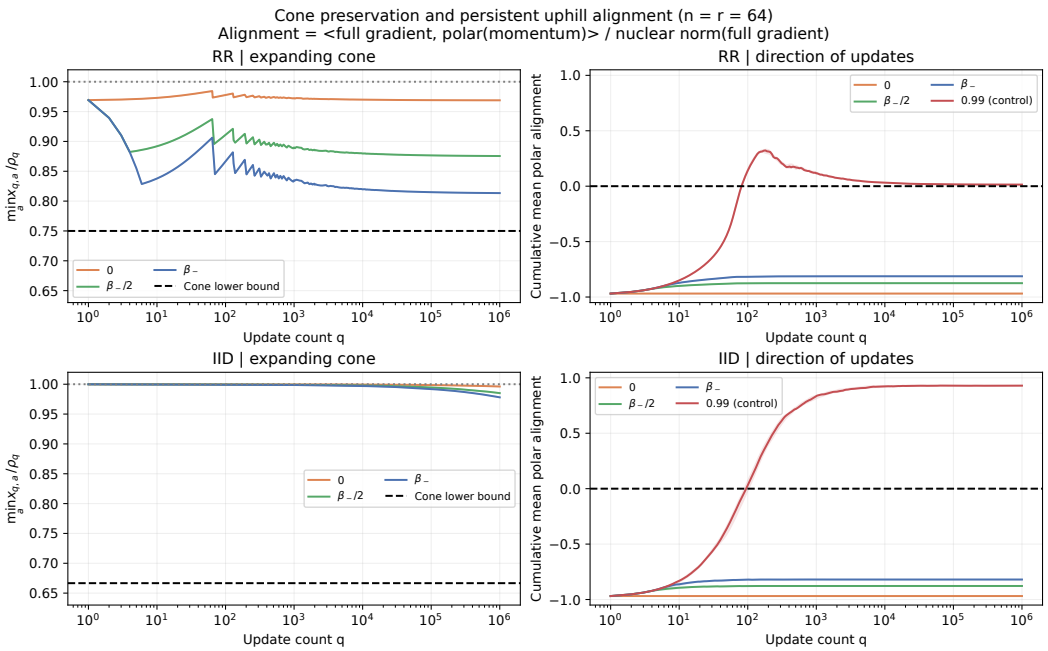

Figure 24: The expansion-region bounds and cumulative average alignment along the recorded trajectories.

K.5 ADDITIONAL NEURAL-NETWORK COMPARISONS

The MNIST scan retains the tanh architecture, balanced 2,560-image subset, $B = 32$, and class-ordered batches, but uses a constant learning rate 0.03. It evaluates 51 momentum values with seed 4060 over 120 epochs. The CIFAR-10 plots extend the ResNet-18 comparisons across learning rates and batch sizes using the epochwise square-root schedule and one seed (20260924) per configuration; the small-batch study covers 90 hyperparameter configurations.

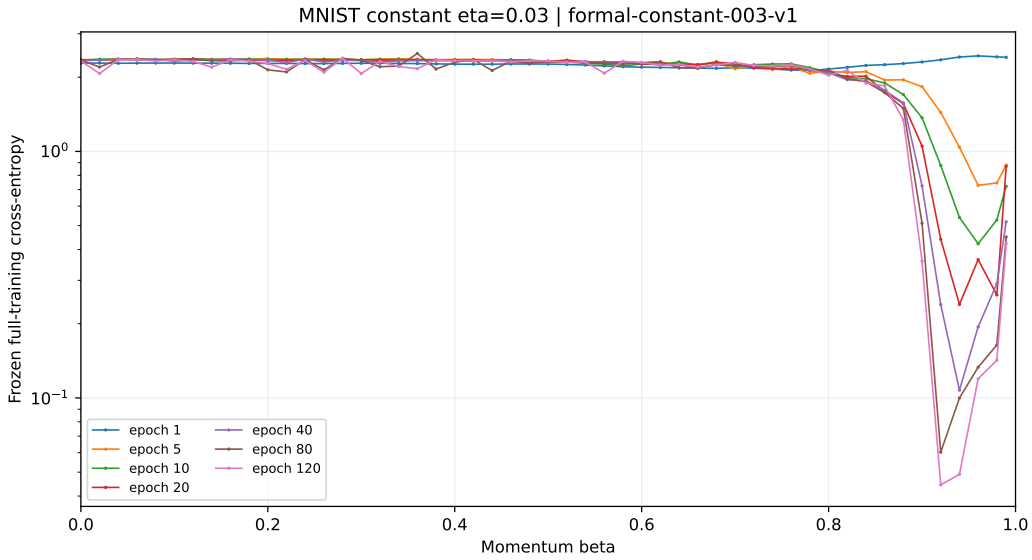

Figure 25: The dense MNIST momentum scan with a constant learning rate. All 51 measured momentum values are shown at each recorded epoch listed in the legend.

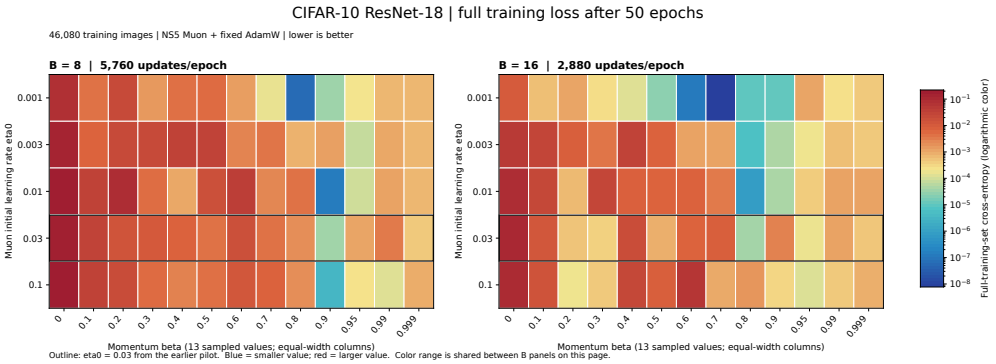

Figure 26: ResNet-18 training loss after 50 epochs, for $B = 8, 16$ and five initial Muon learning rates. AdamW's initial learning rate is fixed at 0.0003. Each row compares momentum values at a shared learning rate.

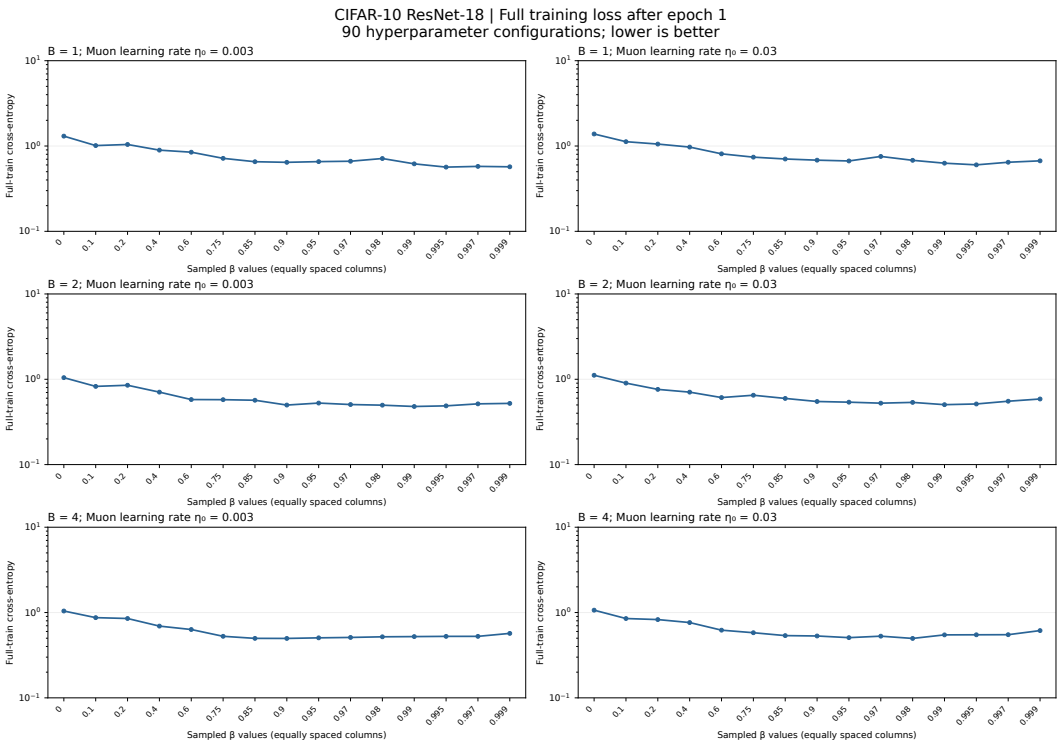

Figure 27: ResNet-18 with $B = 1, 2, 4$ at epoch 1, across 15 momentum values and initial Muon learning rates 0.003 and 0.03.

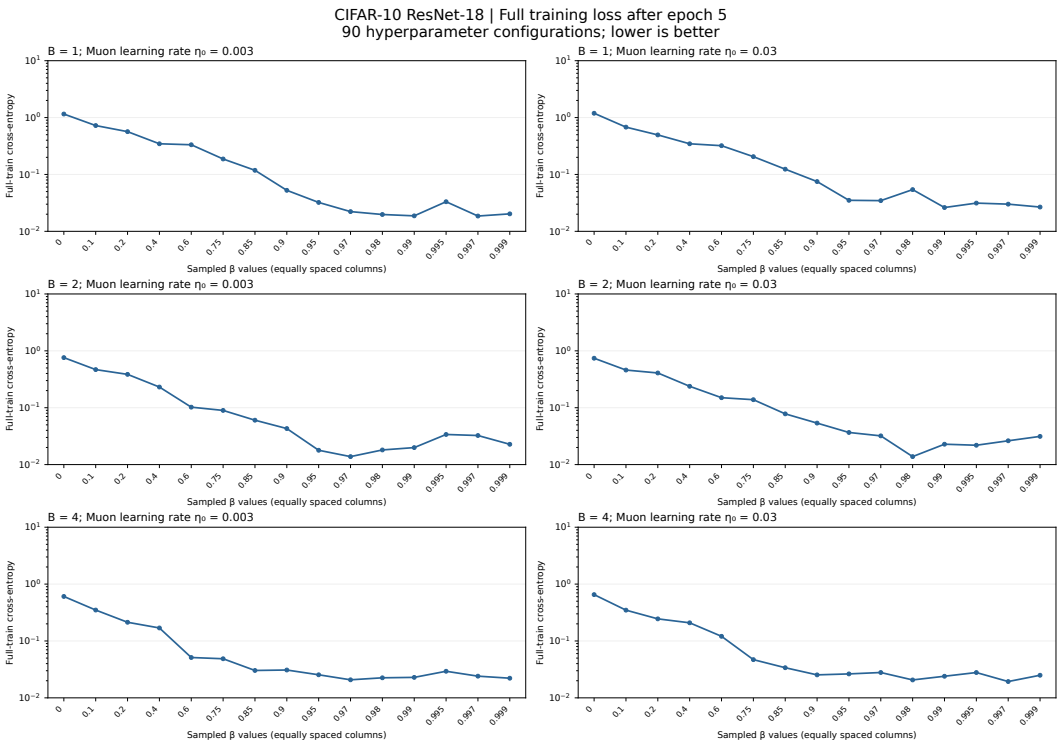

Figure 28: The same ResNet-18 small-batch comparisons at epoch 5.

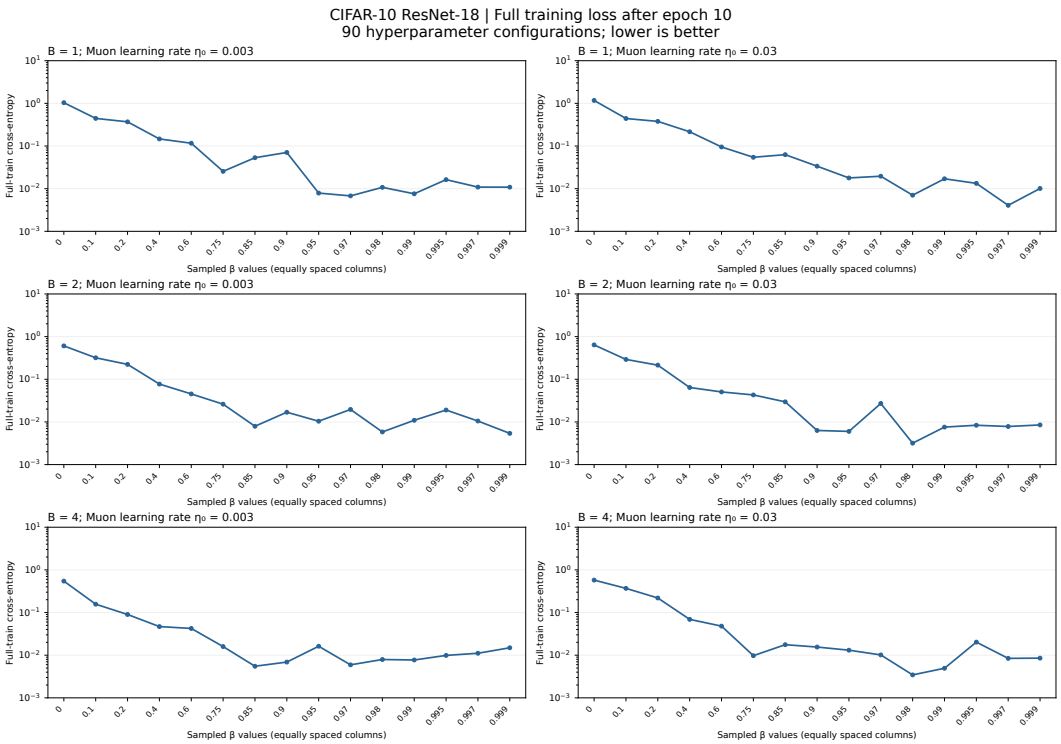

Figure 29: The same ResNet-18 small-batch comparisons at epoch 10.