PRECONDITIONING BENEFITS OF SPECTRAL ORTHOGONALIZATION IN MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

The Muon optimizer has achieved strong empirical performance in large language model training, yet the role of spectral orthogonalization in its effectiveness remains poorly understood. We study its preconditioning effect through two case studies: matrix factorization and in-context linear regression in an infinite-prompt Gaussian softmax-attention model. For both problems, we prove that exact-polar Muon attains logarithmic iteration complexity independent of the relevant condition number, for every fixed momentum parameter $0 \le \mu < 1$ under suitable initialization and stepsize conditions. We also establish condition-number-dependent lower bounds for gradient descent and standard full-batch Adam on separate hard instances. On a common locally initialized matrix-factorization instance, we further separate Muon's logarithmic dependence on the condition number from stabilized Adam's square-root dependence, assuming bounded scalar learning rates. Our analysis reduces the convergence argument to scalar dynamics in the spectral domain, revealing how orthogonalization balances progress across modes. These results identify spectral orthogonalization as the source of the preconditioning effect, which arises even without momentum.

1 INTRODUCTION

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

1.1 THE MUON ALGORITHM AND PRIOR THEORY

Setting the stage, consider an unconstrained optimization problem:

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

where $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is a matrix variable. For Muon, we study the following recursion at each iteration $t \ge 0$:

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

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

Here $\boldsymbol{B}_{-1} = \boldsymbol{0}$. The buffer $\boldsymbol{B}_t$ aggregates the current gradient with exponentially discounted past gradients, $0 \le \mu < 1$ controls the momentum, $\eta_t > 0$ is the learning rate at iteration t, and $\mathsf{msign}(\cdot)$ denotes the partial polar factor. If $\boldsymbol{Z} = \boldsymbol{U}_Z \boldsymbol{\Sigma}_Z \boldsymbol{V}_Z^\top$ is a thin SVD and $p = \min\{m, n\}$, we use the convention

$$\mathsf{msign}(\boldsymbol{Z}) := \boldsymbol{U}_Z \mathsf{diag}\{\boldsymbol{1}_{\{\sigma_i(\boldsymbol{Z})>0\}} : 1 \le i \le p\} \boldsymbol{V}_Z^\top = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{\dagger/2}, \tag{3}$$

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.

where † denotes the Moore–Penrose pseudoinverse; in particular, $\mathsf{msign}(\mathbf{0}) = \mathbf{0}$. In practice this map is approximated using Newton–Schulz iterations (Jordan et al., 2024; Higham, 2008). We analyze the exact-polar, ordinary-momentum recursion in Equation (2); finite orthogonalization is studied empirically in Appendix L.5. At $\mu = 0$, the buffer equals the current gradient and Equation (2) becomes the spectral gradient method. This specialization has been analyzed (An et al., 2025; Shen et al., 2025; Davis & Drusvyatskiy, 2025; Su, 2025) and can perform comparably to momentum-based variants in nonstochastic settings (Shen et al., 2025). In contrast to standard optimizers like `Adam` (Kingma & Ba, 2015) and `AdamW` (Loshchilov & Hutter, 2019), which apply independent per-coordinate transformations, `Muon` spectrally orthogonalizes a matrix-valued buffer.

Motivated by `Muon`'s remarkable empirical success, the past year has witnessed a surge of theoretical efforts aimed at elucidating the mechanisms behind its effectiveness from diverse perspectives. From an optimization standpoint, Li & Hong (2025); Shen et al. (2025) established convergence guarantees of `Muon` on smooth objectives. In particular, Shen et al. (2025) analyzed convergence under spectral-norm smoothness and identified improved guarantees for nuclear-norm stationarity under suitable low-rank, block-structured Hessians. Kovalev (2025) interpreted `Muon` as a trust region method with non-Euclidean trust regions and derived tighter convergence rates for certain function classes. Complementing this line of work, Chen et al. (2025) showed that `Muon` (with decoupled weight decay) implicitly optimizes under a spectral-norm constraint on weight matrices. Stepping beyond worst-case convergence guarantees, Davis & Drusvyatskiy (2025) compared the one-step progress of spectral updates (the $\mu = 0$ case of Equation (2)) with Euclidean gradient updates and showed that, with their prescribed stepsizes, spectral updates have a larger guaranteed *one-step* decrease than gradient descent (`GD`) when the gradient's nuclear rank exceeds the stable rank of the incoming activations. Su (2025) introduced an "isotropic curvature model"—proposed through heuristic arguments and validated empirically in transformer training—and derived gradient-orthogonalization iterations as optimal updates under certain assumptions.

Rhee et al. (2026) study momentum-free exact-polar dynamics in two-factor models, while Kang et al. (2026) analyze uniform spectral growth in LoRA-style factorization through spectral gradient flow. Li et al. (2026b) interpret momentum as spectral filtering under a structured signal-plus-perturbation model. Our discrete-time exact-polar guarantees cover symmetric matrix factorization in both wide and undercomplete regimes for every fixed $\mu < 1$ under the stated initialization conditions.

1.2 This paper: preconditioning with `Muon`

In this work, we investigate the preconditioning effect built into `Muon` through spectral orthogonalization, which may help the optimizer adapt to the geometry of neural networks (Jordan et al., 2024; Bernstein & Newhouse, 2024b; Vasudeva et al., 2025; Lau et al., 2025). We focus on two concrete matrix optimization problems: (a) matrix factorization and (b) in-context linear regression with a one-layer softmax transformer. These problems let us examine how the momentum buffer interacts with spectral orthogonalization and compare the resulting dynamics with coordinatewise adaptation. Our main contributions are summarized below.

- *Matrix factorization.* We show in Theorem 1 that `Muon` achieves a condition-number-free, logarithmic iteration complexity for every fixed momentum $\mu < 1$, both in the wide and undercomplete regimes under their respective initialization conditions. The undercomplete initialization is chosen for the prescribed accuracy, and its admissible scale may depend on κ and other problem parameters. Our high-accuracy lower bound in Theorem 2 shows that a broader class of coordinatewise methods, including the specified deterministic full-batch `Adam` recursion, can have polynomial dependence on the condition number. A separate local construction gives a matched-instance separation from stabilized `Adam` with bounded scalar learning rates (Proposition 6).
- *In-context learning with softmax attention.* The ICL quadratic arises exactly as the infinite-prompt population loss of a structured one-layer softmax transformer under Gaussian design. Akin to matrix factorization, `Muon` converges with a condition-number-free rate in relative operator error (Theorem 3). We also prove a condition-number-dependent lower bound for the same coordinatewise class (Theorem 4).

problem	algorithm	regime/result	iterations	paper
matrix factorization	Muon	$k = r$ (exact)	$\log \frac{1}{\varepsilon}$	this work (Theorem 1)
		$k > r$ (over-parameterized)	$\log \frac{1}{\varepsilon}$	this work (Theorem 1)
	GD	$k = r$ (exact)	$\kappa \log \frac{1}{\varepsilon}$	Chi et al. (2019)
		$k > r$ (over-parameterized)	$\kappa^3 \log \frac{1}{\varepsilon}$	Stöger & Soltanolkotabi (2021)
		lower bound	$\sqrt{\kappa} \log \frac{1}{\kappa\varepsilon}$	this work (Theorem 2)
	Adam	lower bound	$\sqrt{\kappa} \log \frac{1}{\kappa\varepsilon}$	this work (Theorem 2)
	Muon	matched local upper bound	$\log \kappa + \log \frac{1}{\varepsilon}$	this work (Proposition 6)
	Adam	matched local lower bound	$\sqrt{\kappa} \log \frac{1}{\varepsilon}$	this work (Proposition 6)
in-context learning	Muon	quadratic	$\log \frac{1}{\varepsilon}$	this work (Theorem 3)
	GD	lower bound	$\sqrt{\kappa} \log \frac{1}{\varepsilon}$	this work (Theorem 4)
	Adam	lower bound	$\sqrt{\kappa} \log \frac{1}{\varepsilon}$	this work (Theorem 4)

Table 1: Muon upper bounds and coordinatewise history lower bounds, alongside existing GD upper bounds. Theorem 1 separates $r \le k < d$ (undercomplete) from $k \ge d$ (wide). Coordinatewise history lower bounds include the specified deterministic full-batch Adam recursion and use different initializations and metrics from the main upper bounds. The matched local rows share one instance, initialization, and target $f(\boldsymbol{U}_T) \le (\varepsilon/64) f(\boldsymbol{U}_0)$; their Adam bound requires a positive stabilizer and bounded scalar rates. Iteration orders are for $\varepsilon \to 0$, suppressing constants depending on fixed μ and ρ. The remaining upper bounds use relative reconstruction error (matrix factorization) or relative operator error (ICL). The remaining lower bounds use $f(\boldsymbol{U}_T) \le \varepsilon$ with $0 < \varepsilon < c/\kappa$ (matrix factorization), and $f(\boldsymbol{Q}_T) \le (\varepsilon/8) f(\boldsymbol{Q}_0)$ (ICL), under the conditions of the cited theorems.

Our analysis uses scalar sequences in the target eigenbasis to explain how spectral orthogonalization normalizes update magnitudes. The preconditioning effect is already present at $\mu = 0$; positive momentum changes scalar trajectories and rate constants while preserving this normalization. The coordinatewise history lower bounds use separate hard instances, initializations, and error metrics from the main Muon upper bounds; Proposition 6 supplies a matched local comparison for stabilized Adam. The practical numerical comparisons illustrate condition-number and rank dependence under their stated protocols; they do not certify the theorems' sufficient initialization and stepsize conditions. See Table 1 for the theoretical rates.

Numerical comparisons support weak condition-number dependence of Muon at both zero and positive momentum. On a circulant ICL quadratic and factorization of its normalized Hessian, SignGD and GD slow at large κ even after separate-seed tuning, while tuned GD remains competitive in relative objective error for ICL; see Section 2.1.4, Section 2.2, and Appendix L.

Due to space limitations, we defer additional related work to Appendix A.

1.3 Notation

We use bold uppercase letters for matrices and bold lowercase letters for vectors. For any matrix $\boldsymbol{M}$, we denote by $\sigma_i(\boldsymbol{M})$ the i-th largest singular value of $\boldsymbol{M}$, let $\sigma_{\min}(\boldsymbol{M})$ be its smallest singular value, and we let $\|\boldsymbol{M}\|$ (resp. $\|\boldsymbol{M}\|_{\mathrm{F}}$) represent its spectral norm (resp. Frobenius norm). For any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_2$ denotes its Euclidean norm. For any $k \le d$, we let $\mathcal{O}_{d\times k}$ denote the set of orthonormal matrices in $\mathbb{R}^{d\times k}$. For any set of scalars $(a_1, \ldots, a_d)$, we denote by $\mathsf{diag}\{a_1, \ldots, a_d\}$ the diagonal matrix whose diagonal entries are $a_1, \ldots, a_d$. Finally, for any scalar $x \in \mathbb{R}$, we define the sign function as $\mathsf{sign}(x) = 1$ if $x > 0$, $\mathsf{sign}(x) = 0$ if $x = 0$, and $\mathsf{sign}(x) = -1$ if $x < 0$. The notation $O_\mu(\cdot)$ hides multiplicative constants depending only on the fixed momentum parameter μ.

2 Main Results: Two Case Studies

We study the Muon recursion in Equation (2) for matrix factorization and the exact infinite-prompt Gaussian softmax population objective, giving problem-specific convergence guarantees, optimizer comparisons, and synthetic numerical illustrations.

2.1 MATRIX FACTORIZATION

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

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

Here, $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ is a rank-r positive semidefinite matrix, and $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ is a factor with $k \geq r$ columns. Exact parameterization means $k = r$, while over-parameterization means $k > r$. The width distinction in Theorem 1 compares k with d: $k = r < d$ and $r < k < d$ are both undercomplete, whereas $k \geq d$ is wide. The boundary case $k = r = d$ is both exactly parameterized and wide. Throughout, we let $\boldsymbol{M}^\star = \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}$ be the eigen-decomposition of $\boldsymbol{M}^\star$, where $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_r^\star\}$ contains the nonzero eigenvalues $\lambda_1^\star \geq \cdots \geq \lambda_r^\star > 0$, and $\boldsymbol{V}^\star \in \mathbb{R}^{d\times r}$ is an orthonormal matrix whose columns correspond to the associated eigenvectors. We set $\lambda_{\max}^\star := \lambda_1^\star$, $\lambda_{\min}^\star := \lambda_r^\star$, and define the condition number by

$$\kappa := \lambda_1^\star/\lambda_r^\star.$$

2.1.1 CONVERGENCE GUARANTEES FOR MUON

When applied to the matrix factorization problem (4), Muon takes the form

$$\boldsymbol{B}_t = \mu\boldsymbol{B}_{t-1} + (\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t, \tag{5a}$$
$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\,\mathsf{msign}(\boldsymbol{B}_t). \tag{5b}$$

Here $\boldsymbol{B}_{-1} = \boldsymbol{0}$ and $t = 0, 1, \ldots$.

Theorem 1. *Fix $\mu \in [0, 1)$. There are constants $C_\mu \geq 1$ and $\underline{\rho}_\mu \in (0, 1)$ depending only on μ such that every $\rho \in [\underline{\rho}_\mu, 1)$ admits stepsize schedules with geometric decay factor ρ after a possible initialization step and fresh independent $\mathsf{Unif}[1, 2]$ amplitudes drawn independently of the initialization, for which:*

(a) Wide regime ($k \geq d$). Fix $\boldsymbol{O}$ with $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$, choose $0 < \alpha \leq \sqrt{\lambda_{\max}^\star}$, and set $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$. Draw the stepsize amplitudes independently. Then, with probability one, $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \leq C_\mu\lambda_{\max}^\star\rho^t$ for every $t \geq 0$. Consequently, for every $0 < \varepsilon < \lambda_{\max}^\star$, $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \leq \varepsilon$ for all $t \geq T_\varepsilon$.

(b) Undercomplete regime ($r \leq k < d$). Fix a prescribed $0 < \varepsilon < \lambda_{\max}^\star$ and choose a sufficiently small deterministic $\alpha_\varepsilon > 0$. Draw $\boldsymbol{O} \sim \mathsf{Unif}(\mathcal{O}_{d\times k})$ and set $\boldsymbol{U}_0 = \alpha_\varepsilon\boldsymbol{O}$. Draw the stepsize amplitudes independently of $\boldsymbol{O}$. Then, with probability at least 0.99, $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \leq \varepsilon$ for every $t \geq T_\varepsilon$.

In both regimes, the iteration bound for reconstruction error ε is

$$T_\varepsilon = O_\mu\left(\frac{1 + \log(\lambda_{\max}^\star/\varepsilon)}{1 - \rho}\right), \qquad \textit{independently of } \kappa. \tag{6}$$

Choosing ρ as a function of μ alone gives $T_\varepsilon = O_\mu(1 + \log(\lambda_{\max}^\star/\varepsilon))$.

The wide guarantee controls one trajectory at every time; the undercomplete guarantee uses an initialization chosen for the prescribed accuracy. Its admissible scale may depend on ε, κ, the dimensions, and ρ; Appendix B gives the precise conditions. A separate positive-momentum refinement uses one deterministic amplitude for each prescribed accuracy and a suitably selected decay factor; see Appendix B.1.

In the aligned scalar dynamics, spectral orthogonalization gives each active coordinate the same step magnitude. Momentum aggregates gradient history separately within each coordinate.

2.1.2 COMPARISONS WITH OTHER OPTIMIZERS

To better understand the preconditioning benefit of Muon, we compare it with GD and coordinatewise methods.

We begin with the extensively studied GD update rule:

$$\textbf{(GD)} \qquad \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\big((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\big), \qquad t = 0, 1, \cdots \tag{7}$$

As $\varepsilon \to 0$, representative guarantees under the respective initialization and stepsize conditions give $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \le \varepsilon$ in

$$\begin{cases} O\big(\kappa \log(\lambda_1^\star/\varepsilon)\big) \text{ iterations} & \text{if } k = r \text{ (exactly-parameterized)}, \\ O\big(\min\{\kappa^3 \log(\lambda_1^\star/\varepsilon),\, \lambda_1^\star/\varepsilon\}\big) \text{ iterations} & \text{if } k > r \text{ (over-parameterized)}; \end{cases} \tag{8}$$

see, e.g., Chi et al. (2019); Stöger & Soltanolkotabi (2021); Zhuo et al. (2024) for more details. The two over-parameterized bounds use different sufficient conditions; the logarithmic bound uses an accuracy-dependent small initialization. These are upper bounds for GD, and their dependence on κ contrasts with the rate in Theorem 1.

Next, consider SignGD, the simplest coordinatewise sign method:

$$\textbf{(SignGD)} \qquad \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \,\mathsf{sign}\,\big((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\big), \qquad t = 0, 1, \cdots \tag{9}$$

where $\mathsf{sign}(\cdot)$ is applied entrywise. We use SignGD as the coordinatewise baseline in the numerical experiments. The lower bound below goes further: it also covers methods that retain gradient history.

We call the shared-coordinate zero-respecting gradient-history class the *coordinatewise history class*: every entry uses the same function of its own gradient history, and an all-zero history produces no displacement. For deterministic full-batch Adam, we use shared hyperparameters, zero initial moments, and no weight decay; Appendix G gives the exact class definition, recursion, and inclusion proof.

Theorem 2. *There is a universal $c > 0$ such that, for every $\kappa \ge 5$ and $0 < \varepsilon < c/\kappa$, there exists a matrix-factorization instance of condition number κ and an initialization such that every possibly randomized method in the coordinatewise history class satisfies*

$$f(\boldsymbol{U}_T) \le \varepsilon \quad \Longrightarrow \quad T = \Omega\left(\sqrt{\kappa}\log\frac{1}{\kappa\varepsilon}\right). \tag{10}$$

This includes affine sign-span methods and the specified deterministic full-batch Adam recursion.

The high-accuracy regime corresponds to relative objective error of order $1/\kappa$ or smaller; the theorem does not establish an $\Omega(\sqrt{\kappa})$ barrier for a constant-factor reduction. The construction also obstructs polar-sign updates, but its initialization does not satisfy Theorem 1, so it does not separate the algorithms on a common instance. See Section G for the proof.

A matched local separation. For an aligned local initialization, Proposition 6 separates $O_\mu(\log\kappa + \log(1/\varepsilon))$ iterations for Muon from $\Omega(\sqrt{\kappa}\log(1/\varepsilon))$ for stabilized full-batch Adam, on the same instance, initialization, and relative-objective target. The result assumes a positive Adam stabilizer and bounded scalar learning rates; its initialization scale depends on the requested accuracy. Appendix J gives the precise statement, construction, deterministic Muon schedule, and proof.

2.1.3 Intuition

To build intuition, consider the exactly-parameterized regime $k = r$ and suppose the iterate and momentum buffer are aligned with the eigenvectors of the target matrix.

Decoupling of Muon dynamics into independent scalar sequences. Write this alignment as

$$\boldsymbol{U}_t = \boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top, \tag{11}$$

$$\boldsymbol{B}_{t-1} = \boldsymbol{V}^\star\boldsymbol{P}_{t-1}\boldsymbol{R}^\top, \qquad \boldsymbol{P}_{-1} = \boldsymbol{0}, \qquad \text{for all } t \ge 0, \tag{12}$$

where $\boldsymbol{\Sigma}_t = \mathsf{diag}\{\sigma_{1,t}, \ldots, \sigma_{r,t}\}$, $\boldsymbol{P}_{t-1} = \mathsf{diag}\{b_{1,t-1}, \ldots, b_{r,t-1}\}$, and $\boldsymbol{R} \in \mathcal{O}_{r\times r}$. The $\sigma_{i,t}$ are signed spectral coordinates and can take either sign. The momentum buffer shares the same frame, and Equation (5) becomes

$$\boldsymbol{P}_t = \mu\boldsymbol{P}_{t-1} + \boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star\boldsymbol{\Sigma}_t, \tag{13}$$

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \eta_t \,\mathsf{diag\text{-}sign}(\boldsymbol{P}_t), \tag{14}$$

where diag-sign applies the scalar sign to the diagonal entries. Thus the spectral coordinates evolve independently according to the two-state recursions

$$b_{i,t} = \mu b_{i,t-1} + (\sigma_{i,t}^2 - \lambda_i^\star)\sigma_{i,t}, \tag{15a}$$

$$\sigma_{i,t+1} = \sigma_{i,t} - \eta_t \,\mathsf{sign}(b_{i,t}), \qquad b_{i,-1} = 0. \tag{15b}$$

In this aligned eigenbasis, polar orthogonalization removes buffer magnitudes, so every active spectral coordinate moves by η_t, independently of its target eigenvalue. By comparison, `GD` displacements scale with the scalar gradient, retaining its target-eigenvalue dependence. This spectral normalization is the preconditioning mechanism behind the condition-number-free iteration rate. For $\mu > 0$, gradient history affects the buffer signs and scalar trajectories, while the equal-step normalization remains. Set $a_i = \sqrt{\lambda_i^\star}$ and $g_a(u) = (u^2 - a^2)u$. Write $\eta_t = c_t h_t$, where $h_t = A\rho^t$ is the geometric base step, $A > 0$, and $c_t \in [1, 2]$. For $\rho \in (0, 1)$ sufficiently close to one depending only on μ, the scalar analysis relates the normalized buffer to the current gradient through the local, scale-free estimate

$$\left| \frac{b_{i,t}}{\sum_{j=0}^{t} \mu^j} - g_{a_i}(\sigma_{i,t}) \right| \lesssim_\mu (\sigma_{i,t}^2 + a_i^2)h_t + h_t^3. \tag{16}$$

This estimate localizes any sign disagreement between the buffer and current gradient to neighborhoods of the target roots or zero. The relay argument shows that each scalar mode reaches a shrinking neighborhood of a target root at a rate uniform over the target eigenvalues for fixed μ. The undercomplete case must also control imperfect alignment; the details are in Section B.

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

Why can coordinatewise preconditioning be slower? The coordinatewise history lower bounds exploit a feature that SignGD and the stated deterministic full-batch `Adam` recursion share: neither can activate a coordinate whose entire gradient history is zero. Spectral orthogonalization generally acts on the matrix as a whole and mixes information across its entries. The leading-block MF construction also restricts polar updates. In the matched local construction, instead, Adam's stabilizer makes its normalized dynamics approach a quadratic zero chain, while polar updates retain their spectral normalization even for arbitrarily small gradients.

To see this more formally, denote by $\boldsymbol{u}_t = \mathsf{vec}(\boldsymbol{U}_t)$ the flattened iterate, where $\mathsf{vec}(\boldsymbol{Z})$ stacks the rows of a matrix $\boldsymbol{Z}$ into a single column vector. Invoking the identities $\mathsf{vec}(\boldsymbol{AXB}) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\mathsf{vec}(\boldsymbol{X})$ and $\mathsf{msign}(\boldsymbol{Z}) = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{\dagger/2}$, we can express the `Muon` update in Equation (5) as

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

The matrix multiplying the buffer is generally non-diagonal.

In contrast, `SignGD` employs a diagonal, coordinatewise rescaling. For instance, the `SignGD` update (9) can be expressed as

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

where $|\boldsymbol{z}|^\dagger$ is the entrywise reciprocal of each nonzero magnitude and is zero elsewhere. This diagonal rescaling does not encode cross-coordinate curvature. The deterministic full-batch `Adam` recursion in Proposition 4 uses momentum and a second-moment state through coordinatewise functions of gradient history.

2.1.4 Numerical Experiments

We factor the normalized Hessian of the circulant quadratic in Equation (355): $\boldsymbol{M}^\star = \boldsymbol{A}/\kappa$, with $d = r = k = 4096$, $\lambda_{\max}(\boldsymbol{M}^\star) = 1$, and $\lambda_{\min}(\boldsymbol{M}^\star) = 1/\kappa$. We initialize $\boldsymbol{U}_0 = 0.1\boldsymbol{O}$ with an independent random circulant orthogonal matrix and measure $e_t^{\mathrm{MF}} = \|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\|$. Figure 1 has one column per algorithm: `Muon`, `SignGD`, and `GD` (Momentum `GD` for $\mu > 0$). The top

row uses $\mu = 0$; the bottom row uses $\mu = 0.9$ in all three methods. With $\boldsymbol{B}_{-1} = \boldsymbol{0}$, we set $\boldsymbol{B}_t = \mu \boldsymbol{B}_{t-1} + \nabla f(\boldsymbol{X}_t)$ and use directions $\mathsf{msign}(\boldsymbol{B}_t)$, $\mathsf{sign}(\boldsymbol{B}_t)$, and $\boldsymbol{B}_t$, respectively. `Muon` uses geometric decay, `SignGD` plateau decay, and `GD` constant steps. For each method and momentum, a normalized base rate is selected on two pilot seeds, then fixed across κ; ten separate evaluation seeds are paired across methods. Appendix L.1 gives the protocol.

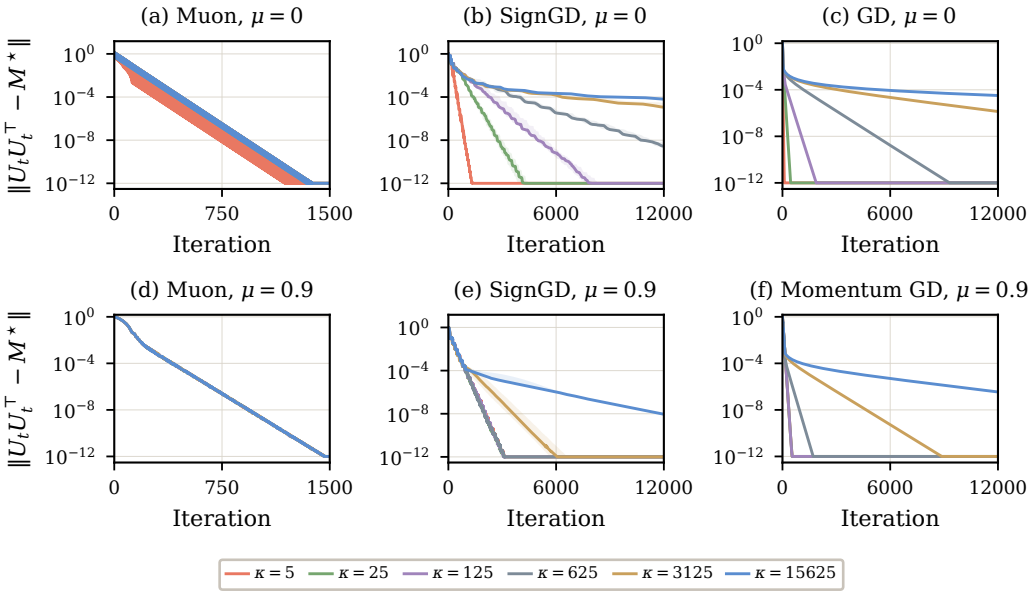


Figure 1: Matrix factorization: $\boldsymbol{M}^\star = \boldsymbol{A}/\kappa$, $d = r = k = 4096$, $\|\boldsymbol{M}^\star\| = 1$, $\boldsymbol{U}_0 = 0.1\boldsymbol{O}$ with random circulant orthogonal $\boldsymbol{O}$. Columns: `Muon`, `SignGD`, and `GD` (Momentum `GD` in the bottom row). Top: $\mu = 0$; bottom: $\mu = 0.9$ for all three algorithms. Curves show $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\|$ (median and 20th–80th percentiles over ten seeds). Display horizons are 1,500 for `Muon` and 12,000 for both `SignGD` and `GD` rows. Rates follow the fixed normalized protocol; errors below 10^{-12} are displayed at that floor.

At $\mu = 0.9$, `Muon` reaches error 10^{-6} in 665–670 iterations throughout the grid. `SignGD`'s median rises from about 1,500 on the smaller grid to 6,035.5 at $\kappa = 15625$; the slowdown persists after separate-seed tuning. Appendix L.1 gives the tuning comparison, and Appendix L.3 examines the search-rank regimes of Theorem 1.

2.2 In-context learning with infinite-prompt softmax attention

Our second case study is in-context linear regression with a structured one-layer softmax-attention transformer (Vaswani et al., 2017). Each task is indexed by $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, with labels $y = \boldsymbol{w}^\top \boldsymbol{x}$. Independently of $\boldsymbol{w}$, the context inputs and query $\boldsymbol{x}_{\mathrm{q}}$ are independent draws from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{S})$, where $\boldsymbol{S} \succ \boldsymbol{0}$. Given labeled context examples and the query, the model must predict $\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{q}}$ without updating its parameters (Brown et al., 2020).

The matrix $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ is the structured merged score parameter: the softmax layer scores each context input against the query and forms a softmax-weighted average of the corresponding labels. We do not train arbitrary separate key, query, and value matrices. In the infinite-prompt limit, Gaussian exponential tilting gives

$$\widehat{y}_{\mathrm{q}}^{\mathsf{sm},\infty} = \boldsymbol{w}^\top \boldsymbol{S}\boldsymbol{Q}\boldsymbol{x}_{\mathrm{q}}.$$

Appendix H gives the temperature and masking conventions and the limit derivation.

Averaging half the squared prediction error over tasks and queries yields the population training problem

$$\underset{\boldsymbol{Q} \in \mathbb{R}^{d \times d}}{\text{minimize}} \quad f(\boldsymbol{Q}) = f_\infty^{\mathsf{sm}}(\boldsymbol{Q}) := \frac{1}{2} \operatorname{tr}\big((\boldsymbol{S}\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{S}(\boldsymbol{S}\boldsymbol{Q} - \boldsymbol{I})^\top\big). \tag{19}$$

Thus this quadratic is exactly the infinite-prompt population training loss of the structured one-layer softmax-attention model. The finite-prompt objective is not covered by Theorem 3; Appendix H explains the additional control needed to transfer exact-polar dynamics.

Its minimizer is $\boldsymbol{Q}^\star = \boldsymbol{S}^{-1}$, and its Hessian condition number in Frobenius geometry is

$$\kappa := \kappa(\boldsymbol{S})^3. \tag{20}$$

Indeed, $\nabla^2 f(\boldsymbol{Q})[\boldsymbol{H}] = \boldsymbol{S}^2\boldsymbol{H}\boldsymbol{S}$, whose eigenvalues are $\lambda_i(\boldsymbol{S})^2\lambda_j(\boldsymbol{S})$. We also write

$$q^\star_{\max} := \|\boldsymbol{Q}^\star\| = \|\boldsymbol{S}^{-1}\| = \frac{1}{\sigma_{\min}(\boldsymbol{S})}. \tag{21}$$

Convergence guarantees for `Muon`. Applied to Equation (19), `Muon` follows

$$\boldsymbol{B}_t = \mu\boldsymbol{B}_{t-1} + \boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2, \tag{22a}$$

$$\boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t - \eta_t\,\mathsf{msign}(\boldsymbol{B}_t). \tag{22b}$$

Here $\boldsymbol{B}_{-1} = \boldsymbol{0}$ and $t = 0, 1, \ldots$. The next theorem shows that this recursion converges at a rate independent of the covariance condition number.

Theorem 3. *Fix $\mu \in [0, 1)$ and any $\rho \in [(2+\mu)/3, 1)$. Use $\eta_t = \frac{1}{2}C_{\eta,t}(\rho-\mu)\|\boldsymbol{S}^{-1}\|\rho^t$ with $C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1, 2]$, and initialize $\boldsymbol{B}_{-1} = \boldsymbol{Q}_0 = \boldsymbol{0}$. Then, for every realization and every $t \geq 0$,*

$$\frac{\|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\|}{\|\boldsymbol{S}^{-1}\|} \leq \rho^t. \tag{23}$$

Consequently, for every $\varepsilon \in (0, 1)$, relative operator error at most ε is reached within

$$T_\varepsilon = O\left(\frac{1 + \log(1/\varepsilon)}{1-\rho}\right), \qquad \textit{independently of } \kappa(\boldsymbol{S}). \tag{24}$$

Choosing ρ as a function of μ alone gives $T_\varepsilon = O_\mu(1 + \log(1/\varepsilon))$.

Every decay factor permitted in Theorem 1 is also admissible here. The operator bound also gives a geometric objective bound, whose conversion to relative objective error introduces a spectrum-dependent prefactor and an additive $O(\log\kappa/(1-\rho))$ term; the exact schedules and conversion are in Appendix C.

Akin to matrix factorization, the ICL dynamics separate into scalar sequences in the eigenbasis of $\boldsymbol{S}$. The ICL gradient is linear in each coordinate, whereas the matrix-factorization gradient is cubic. This explains why ICL allows $\boldsymbol{Q}_0 = \boldsymbol{0}$: at zero, $\nabla f(\boldsymbol{0}) = -\boldsymbol{S}^2$ already has the covariance eigenvectors. Matrix factorization instead has a nonoptimal stationary point at zero.

Comparisons with other optimizers. For this quadratic, `GD` and `SignGD` use directions $\boldsymbol{G}_t = \nabla f(\boldsymbol{Q}_t) = \boldsymbol{S}^2\boldsymbol{Q}_t\boldsymbol{S} - \boldsymbol{S}^2$ and $\mathsf{sign}(\boldsymbol{G}_t)$, respectively, with the sign applied entrywise. The next theorem gives a lower bound within the ICL objective for the coordinatewise history class, including the specified deterministic full-batch `Adam` recursion.

Theorem 4. *For every $\kappa \geq 4$ and $\varepsilon \in (0, 1)$, there exists a softmax-ICL instance with $\kappa(\boldsymbol{S})^3 = \kappa$ and an initialization arbitrarily close to $\boldsymbol{Q}^\star = \boldsymbol{S}^{-1}$ such that every possibly randomized method in the coordinatewise history class satisfies*

$$f(\boldsymbol{Q}_T) \leq \frac{\varepsilon}{8} f(\boldsymbol{Q}_0) \quad \Longrightarrow \quad T = \Omega\left(\sqrt{\kappa}\log\frac{1}{\varepsilon}\right). \tag{25}$$

This includes affine sign-span methods and the specified deterministic full-batch `Adam` recursion.

The construction and proof are in Section I.

Numerical experiments. We test nonlocal starts with $d = 4096$ and the circulant covariance $\boldsymbol{S} = \boldsymbol{A}^{1/3}$ from Equation (355), where $\lambda_{\min}(\boldsymbol{A}) = 1$ and $\lambda_{\max}(\boldsymbol{A}) = \kappa$. Thus $\kappa(\boldsymbol{S})^3 = \kappa$ and $\|\boldsymbol{S}^{-1}\| = 1$. We draw a symmetric circulant $\boldsymbol{Q}_0$ from Gaussian cyclic-diagonal coefficients, independently of the target, and normalize $\|\boldsymbol{Q}_0\| = 2$. The same draw is paired across κ and

algorithms. Figure 2 plots $e_t^{\mathrm{ICL}} = \|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\|$; Appendix L.2 gives the construction and tuning protocol.

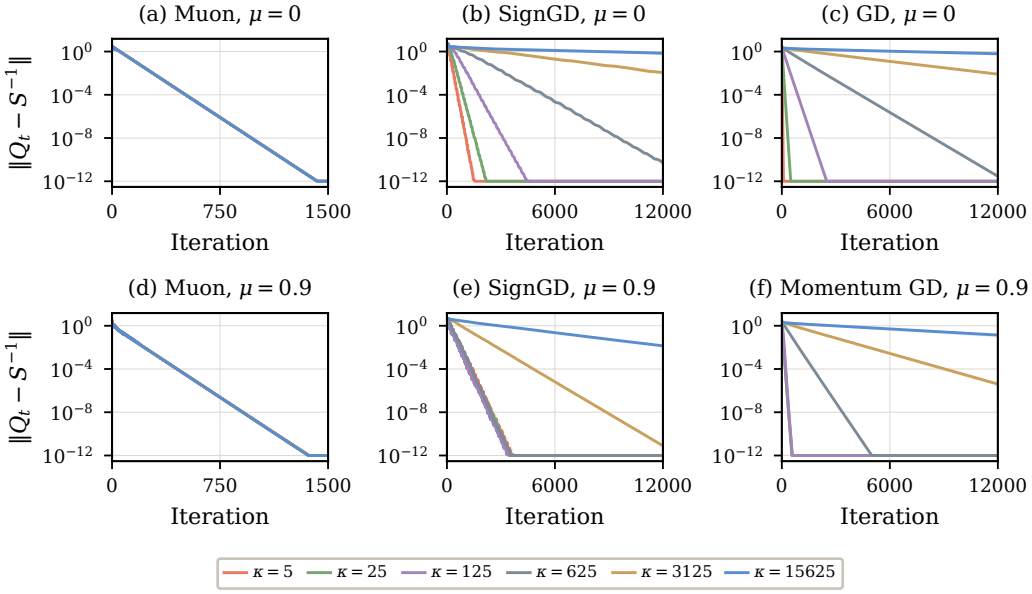


Figure 2: Softmax ICL with $d = 4096$, random initialization $\|\boldsymbol{Q}_0\| = 2$, and $\|\boldsymbol{S}^{-1}\| = 1$. Columns: `Muon`, `SignGD`, and `GD` (Momentum `GD` in the bottom row). Top: $\mu = 0$; bottom: $\mu = 0.9$ in every method. Curves show $\|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\|$ (median and 20th–80th percentiles over ten seeds). `Muon` is untuned; baseline multipliers are fixed after separate-seed pilot selection. Display horizons are 1,500 for `Muon` and 12,000 for both `SignGD` and `GD` rows. The display floor is 10^{-12}.

Momentum `Muon` retains nearly coincident error trajectories, whereas the baselines slow at large κ. At $\mu = 0.9$, `Muon` reaches 10^{-6} in median 680 iterations at every tested κ. Even after per-κ tuning, `SignGD` reaches this tolerance on none of ten seeds at $\kappa = 15625$ within 12,000 updates. These nonlocal starts extend beyond the $\boldsymbol{Q}_0 = \boldsymbol{0}$ initialization of Theorem 3. Tuned `GD` remains competitive in relative objective error; Appendix L.2 compares fixed and per-κ tuning in both metrics, and Appendix L.4 checks the zero-initialized operator envelope.

3 Discussion

Spectral orthogonalization balances progress across spectral coordinates, yielding condition-number-free iteration rates for every fixed $\mu < 1$ in the two analyzed problems under the stated conditions. The lower bounds identify a structural limitation of the specified coordinatewise history class. The matched local MF construction further separates `Muon` from stabilized `Adam` on the same instance, initialization, and relative objective, under a finite scalar learning-rate cap. Figure 9 (Appendix L.6) tests this mechanism with actual full-batch `Adam` and exact-polar `Muon` across common initialization scales, reporting relative objective and denominator dominance. Extending this separation to a regime where Adam's second moments dominate its stabilizer, or to the isotropic initializations of Theorem 1, remains open.

The guarantees concern exact partial polar updates and, for ICL, the infinite-prompt objective; the simulations also expose schedule and tuning sensitivity. Appendix L.5 examines finite Newton–Schulz orthogonalization empirically. Extensions to broader matrix and transformer models require conditions that preserve this effect along optimization trajectories, including under approximate orthogonalization (Davis & Drusvyatskiy, 2025; Su, 2025; Shen et al., 2025). Neural-network structure, such as block Hessians (Zhang et al., 2024), may help identify such conditions.

AI USE STATEMENT

We used GPT Pro to assist with the theoretical analysis extending the convergence rates of `Muon` to the momentum setting. We used Codex to improve the exposition and organization of the manuscript, assist with literature searches and related-work summaries, and edit the LaTeX source. Codex also assisted with checking the proofs and extending the admissible stepsize schedules and decay factors, as well as implementing the simulation suite, analyzing its saved results, and generating the figures. The authors take responsibility for the correctness, originality, and integrity of the final manuscript, including all AI-assisted content.

ETHICS STATEMENT

This work studies optimization theory through matrix factorization and a Gaussian softmax ICL model. The numerical illustrations use synthetic instances; the study does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

The detailed assumptions, stepsize schedules, initialization conditions, and proofs for the `Muon` upper bounds appear in Appendices B and C. The lower-bound constructions and the precise coordinatewise history class, including the full-batch `Adam` recursion, are given in Appendices G and I. Appendix J gives the matched local construction, its deterministic `Muon` schedule, and the uniform transfer argument for stabilized `Adam`. Appendix H derives the softmax ICL population objective. Sections 2.1.4 and 2.2 report normalized numerical comparisons. Appendix L specifies the seeds, algorithms, learning-rate selection, schedules, stopping rules, and censoring. Per-iteration metrics and configurations for these numerical comparisons are saved, and their figures can be regenerated without rerunning optimization. The metric-factor illustration in Appendix D is a heuristic diagnostic with the protocol limitations stated there.

REFERENCES

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024a.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024b.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: A measure-based perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2512.11784.

Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. *arXiv preprint arXiv:2601.22652*, 2026. URL https://arxiv.org/abs/2601.22652.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted Boltzmann machines. In *Artificial intelligence and statistics*, pp. 111–119. PMLR, 2015a.

David Carlson, Ya-Ping Hsieh, Edo Collins, Lawrence Carin, and Volkan Cevher. Stochastic spectral descent for discrete graphical models. *IEEE Journal of Selected Topics in Signal Processing*, 10 (2):296–311, 2016.

David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015b.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025. URL https://arxiv.org/abs/2501.18322.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

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

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

Sara Dragutinović, Yedi Zhang, and Rajesh Ranganath. To use or not to use Muon: How simplicity bias in optimizers matters. *arXiv preprint arXiv:2603.00742*, 2026. URL https://arxiv.org/abs/2603.00742.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on Muon from simple quadratics. *arXiv preprint arXiv:2602.11948*, 2026. URL https://arxiv.org/abs/2602.11948.

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

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19660–19722. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24d.html.

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/.

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform spectral growth and convergence of Muon in LoRA-style matrix factorization. *arXiv preprint arXiv:2602.06385*, 2026. URL https://arxiv.org/abs/2602.06385.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.19156.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

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

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

Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2602.05725.

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

Wen Li and Weiwei Sun. Some remarks on the perturbation of polar decompositions for rectangular matrices. *Numerical Linear Algebra with Applications*, 13(4):327–338, 2006.

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. URL https://arxiv.org/abs/2606.03899.

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. In *International Conference on Learning Representations*, 2019.

Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon eclipses SignSGD. *arXiv preprint arXiv:2605.09552*, 2026. URL https://arxiv.org/abs/2605.09552.

Ali Parviz, Gal Mishne, and Alex Cloninger. Reassessing Muon for matrix factorization. In *Proceedings of the 2nd Conference on Topology, Algebra, and Geometry in Data Science*, volume 334 of *Proceedings of Machine Learning Research*, pp. 59–86. PMLR, 2026. URL https://proceedings.mlr.press/v334/parviz26a.html.

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.

Mark Rhee, Jamie Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. *arXiv preprint arXiv:2606.30509*, 2026. URL https://arxiv.org/abs/2606.30509.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010*, pp. 1576–1602. World Scientific, 2010.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Tianqi Shen, Jinji Yang, Runze Shi, Jianhao Ma, Jiaye Teng, and Ziye Ma. Towards understanding the power and limits of the Muon optimizer: A river-valley perspective. *arXiv preprint arXiv:2606.21514*, 2026. URL https://arxiv.org/abs/2606.21514.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025.

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

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

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

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

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

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pp. 38611–38654. PMLR, 2023.

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

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

Thomas T Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni. On the concurrence of layer-wise preconditioning methods and provable feature learning. *arXiv preprint arXiv:2502.01763*, 2025.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need Adam: A Hessian perspective. *Advances in neural information processing systems*, 37: 131786–131823, 2024.

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

APPENDIX CONTENTS

A Additional related work **15**

B Analysis for matrix factorization (proof of Theorem 1) **16**
B.1 Detailed conditions for Theorem 1 . . . 16
B.2 Absolute nearest-root control at zero momentum . . . 18
B.3 Scale-free fixed-momentum relay . . . 19
B.4 The condition-number-free wide regime at zero momentum . . . 22
B.5 The fixed-positive-momentum wide regime . . . 23
B.6 Shared tools for undercomplete matrix factorization . . . 23
B.7 Positive momentum with one random stepsize amplitude . . . 28
B.8 The condition-number-free undercomplete regime at zero momentum . . . 37
B.9 Arbitrary admissible decay with independent amplitudes . . . 42

C Analysis of the infinite-prompt softmax ICL population objective (proof of Theorem 3) **44**
C.1 Detailed conditions for Theorem 3 . . . 44
C.2 Scalar decomposition and tracking . . . 44

D The zero-momentum connection between `Muon` and `ScaledGD` for matrix factorization **47**

E Exact aligned momentum reduction for matrix factorization **49**

F Additional auxiliary lemmas for matrix factorization **49**
F.1 Proof of Lemma 9 . . . 49
F.2 Proof of Lemma 25 . . . 50

G Coordinatewise history lower bound for matrix factorization **50**

H Infinite-prompt Gaussian softmax formulation of the ICL objective **55**

I Coordinatewise history lower bound for softmax ICL **56**

J A matched local separation between stabilized `Adam` and `Muon` **59**
J.1 A common local initialization . . . 60
J.2 Deterministic local contraction for `Muon` . . . 60
J.3 A uniform quadratic limit for stabilized `Adam` . . . 61
J.4 The reference barrier and transfer to actual `Adam` . . . 62

K Technical lemmas **63**

L Additional numerical experiments **65**
L.1 Factorization of the hard quadratic Hessian . . . 65
L.2 Circulant ICL with nonlocal random initialization . . . 66
L.3 Search rank and overparameterization . . . 67
L.4 An untuned check of the ICL operator envelope . . . 68
L.5 Finite Newton–Schulz orthogonalization . . . 70
L.6 Actual `Adam` on the matched local factorization instance . . . 72

A ADDITIONAL RELATED WORK

The norm-constrained interpretation of simplified Adam, Shampoo, and Prodigy in Bernstein & Newhouse (2024b) helped motivate convergence analyses of Muon (Li & Hong, 2025; Shen et al., 2025; Chen et al., 2025). Spectral normalization or orthogonalization of gradients also appeared in earlier optimizer designs (Carlson et al., 2015a; 2016; 2015b; Tuddenham et al., 2022). Other perspectives connect Muon's update to an approximation of Shampoo (Jordan et al., 2024; Shah et al., 2025; Gupta et al., 2018) or derive related methods through norm-constrained optimization and matrix preconditioning (Pethick et al., 2025; Lau et al., 2025; Bernstein & Newhouse, 2024a; An et al., 2025). Approximation errors in practical orthogonalization have also been studied (Shulgin et al., 2025). For smooth objectives with stochastic gradients, Kim & Oh (2026) establish stationarity bounds for momentum Muon with finitely many Newton–Schulz steps, matching the exact-polar rate up to an approximation-dependent factor. Complementing these convergence analyses, Li et al. (2026b) explain momentum as a spectral filter under a model with a persistent signal in fixed singular directions and mean-zero perturbations uncorrelated in the Frobenius inner product. Our upper bounds establish the preconditioning effect of exact spectral orthogonalization in two structured objectives for every fixed $\mu \in [0, 1)$.

Another line of work studies implicit bias and generalization. For separable multiclass linear classification, Fan et al. (2025) characterize spectral-norm max-margin bias, complementing the Euclidean or coordinatewise norm biases of GD and Adam. More uniform learning of spectral components has been linked to gains on imbalanced or long-tailed data and to more isotropic learned spectra (Vasudeva et al., 2025; Wang et al., 2025). In a linear associative-memory model with softmax retrieval and heterogeneous task frequencies, Li et al. (2026a) establish optimization and scaling advantages over GD for momentum-free Muon under orthogonal task embeddings and zero initialization. Related statistical benefits of layer-wise preconditioning were established by Zhang et al. (2025), and Tveit et al. (2025) reported faster grokking. These benefits depend on the learning task: Dragutinović et al. (2026) analyze spectral gradient flow in deep linear networks and structured linear attention, showing simultaneous learning of modes or heads under alignment assumptions, and give experiments in which the resulting loss of sequential simplicity bias harms generalization.

Preconditioning also accelerates nonconvex matrix factorization. Tong et al. (2021a;b) introduced ScaledGD and a nonsmooth counterpart, establishing condition-number-free linear convergence in the exactly-parameterized regime with spectral initialization. Subsequent work extended such guarantees to over-parameterization (Zhang et al., 2021; 2023) and small random initialization (Xu et al., 2023). For orthogonalized updates, Kang et al. (2026) analyze uniform spectral growth in LoRA-style two-factor models via spectral gradient flow; their global convergence result uses a smoothed polar map and bounded factor trajectories, with regularization ensuring boundedness. Rhee et al. (2026) study momentum-free, exact-polar dynamics in two-factor models, including alignment, balanced solutions, and geometric annealing after alignment. Our matrix-factorization theorem gives discrete-time guarantees for every fixed momentum $\mu < 1$ under the stated initializations.

The benefits of spectral updates also depend on the objective geometry and stepsize schedule. In population phase retrieval with anisotropic Gaussian inputs, Braun et al. (2026) show how momentum-free spectral descent suppresses misalignment caused by a covariance spike orthogonal to the signal. In a mixed-spiked matrix-sensing model, Shen et al. (2026) analyze early progress and late-stage limitations of momentum-free Muon through river-valley dynamics, and extend their comparisons to momentum updates at points on a spectral river. On simple quadratics, Gonon et al. (2026) demonstrate fixed-step obstructions to minimizer convergence for exact-polar updates and show that one-step improvement need not predict performance over a trajectory. In controlled matrix-factorization experiments, Parviz et al. (2026) find that Muon's ranking relative to tuned AdamW varies with the problem and hyperparameters. These studies motivate keeping the objective, initialization, and schedule explicit in comparisons of optimizers.

Coordinatewise-sign optimization has also been studied as a gradient-compression primitive, with convergence analyses and convex nonconvergence counterexamples for SignGD (Bernstein et al., 2018; Karimireddy et al., 2019). More directly comparing spectral and coordinatewise updates, Paquette et al. (2026) derive high-dimensional dynamics for stochastic matrix least squares, identifying batch-size-dependent preconditioning and regimes favoring either SignSVD or SignSGD, with the latter serving as an Adam proxy. Our lower bounds apply to the coordinatewise history

class containing affine sign-span methods and the specified full-batch `Adam` recursion, within the matrix-factorization and softmax-ICL objectives. These bounds do not follow directly from general first-order or sign-compression results.

A growing theoretical ICL literature studies genuine softmax attention. For example, Huang et al. (2024) analyze the training dynamics of a one-layer softmax transformer. Most directly relevant here, Boursier & Boyer (2026) derive an exact infinite-prompt formula for centered Gaussian prompts and establish finite-prompt concentration together with a gradient-flow transfer result. We use their Gaussian bridge for the population-risk identification; their finite-prompt result does not cover `Muon`.

B Analysis for matrix factorization (proof of Theorem 1)

Proof route. Sections B.2–B.3 establish the scalar bounds, and Sections B.4–B.5 prove the wide regimes. For undercomplete factors, Section B.6 supplies alignment and stability estimates; Section B.8 proves the zero-momentum case and develops the probability and bootstrap tools reused for positive momentum in Section B.9. These sections prove Theorem 1 with fresh independent amplitudes. Section B.7 establishes the additional single-amplitude refinement, including a deterministic amplitude for each prescribed accuracy. Its decay-ratio screening is required only for this refinement, not for Theorem 1.

B.1 Detailed conditions for Theorem 1

The main theorem uses independent amplitudes at every iteration in both regimes. We first record the canonical schedules and initialization conditions, including a separate one-amplitude refinement, then give the independent-amplitude schedules for every admissible decay factor. Throughout, $\boldsymbol{B}_{-1} = \mathbf{0}$.

Fix any momentum $\mu \in [0, 1)$. There are constants $\rho_\mu \in (0, 1)$ and $C_\mu \geq 1$, depending only on μ and not on the target spectrum or its condition number, for which the following statements hold. At zero momentum we take $\rho_0 = 3/4$. Write $a_{\max} = \sqrt{\lambda^\star_{\max}}$. For $0 < \mu < 1$, define

$$\gamma_\mu = \frac{(1-\mu)^4}{68(1-\mu)^4 + 3840(1+\mu)(\mu^2 + 10\mu + 37)}. \tag{26}$$

There is also a ratio $\widehat{\rho}_\mu \in [1 - \gamma_\mu, 1 - \gamma_\mu/2]$, selected as a function of μ alone, for the positive-momentum undercomplete statement below.

(a) Suppose $k \geq d$. Fix the target, $\boldsymbol{O}$, and α before drawing $C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1, 2]$, where $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$ and $0 < \alpha \leq a_{\max}$. Let $\boldsymbol{U}_0 = \alpha \boldsymbol{O}$ and set

$$\eta_t = 2(1-\rho_\mu) C_{\eta,t} a_{\max} \rho_\mu^t, \qquad t \geq 0. \tag{27}$$

Then

$$\mathbb{P}\big(\forall T \in \mathbb{Z}_{\geq 0} : \quad \|\boldsymbol{U}_T \boldsymbol{U}_T^\top - \boldsymbol{M}^\star\| \leq C_\mu \lambda^\star_{\max} \rho_\mu^T\big) = 1. \tag{28}$$

Consequently, for $0 < \varepsilon < \lambda^\star_{\max}$, reconstruction error at most ε is reached in

$$T = O_\mu\left(1 + \log \frac{\lambda^\star_{\max}}{\varepsilon}\right) \tag{29}$$

iterations.

(b) If $r \leq k < d$, then there are $c_{\text{init}} > 0$ and $q_{\text{init}} \in (0, \infty)$ with

$$c_{\text{init}} = c_{\text{init}}(d, r, k, \kappa, \mu), \qquad q_{\text{init}} = q_{\text{init}}(d, k, \mu), \tag{30}$$

such that the following holds for every prescribed $0 < \varepsilon < \lambda^\star_{\max}$. Choose any deterministic scale satisfying

$$0 < \frac{\alpha_\varepsilon}{\sqrt{\lambda^\star_{\max}}} \leq c_{\text{init}} \left(\frac{\varepsilon}{\lambda^\star_{\max}}\right)^{q_{\text{init}}} \tag{31}$$

and draw $\boldsymbol{O} \sim \mathsf{Unif}(\mathcal{O}_{d\times k})$. Set $\boldsymbol{U}_0 = \alpha_\varepsilon \boldsymbol{O}$. For $0 < \mu < 1$, draw one scalar $C_\eta \sim \mathsf{Unif}[1,2]$ independently of $\boldsymbol{O}$ and use

$$\eta_t = 4\gamma_\mu C_\eta a_{\max} \widehat{\rho}_\mu^t, \qquad t \geq 0. \tag{32}$$

This same amplitude is used from the first step onward. With $x = \varepsilon/\lambda^\star_{\max}$, one may take

$$T_\varepsilon = 1 + \left\lceil \frac{\log 2}{-\log \widehat{\rho}_\mu} \right\rceil + \left\lceil \frac{\log(320/x)}{-\log \widehat{\rho}_\mu} \right\rceil, \tag{33}$$

$$T_\varepsilon \leq 3 + \frac{2}{\gamma_\mu} \log \frac{640}{x} = O\left(\frac{1 + \log(1/x)}{(1-\mu)^4} \right). \tag{34}$$

At $\mu = 0$, independently draw $C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1,2]$ and use the schedule

$$\eta_0 = C_{\eta,0} a_{\max}, \qquad \eta_t = 4(1-\rho_0) C_{\eta,t} a_{\max} \rho_0^{t-1}, \quad t \geq 1. \tag{35}$$

In either case there is a deterministic integer T_ε satisfying

$$T_\varepsilon \leq C_\mu \left(1 + \log \frac{\lambda^\star_{\max}}{\varepsilon} \right) \tag{36}$$

such that

$$\mathbb{P}\left(\sup_{t \geq T_\varepsilon} \|\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \leq \varepsilon \right) \geq 0.99 \tag{37}$$

Choice of constants and initialization. For positive momentum, the canonical wide-regime ratio ρ_μ is constructed in Equations (50) and (51); the undercomplete ratio $\widehat{\rho}_\mu$ is selected in Equation (147). A sufficient positive-momentum initialization prefactor is

$$c_{\text{init}} = \frac{c_{\mu,d,k}}{\kappa\sqrt{dr}} \exp\left[-C_{\mu,k}(1 + \log\kappa)^2 \right], \tag{38}$$

with q_{init} as in Equation (184), independent of κ. For zero momentum, the exact horizon and a sufficient exponent are given in Equations (190) and (242); the prefactor is chosen using Equations (192) and (244). Any fixed $\rho \in [2/3, 1)$ may replace $\rho_0 = 3/4$, with constants depending on ρ. The explicit constants are conservative, especially as μ approaches one.

Deterministic amplitude for each prescribed accuracy. Fix $0 < \mu < 1$ and $r \leq k < d$. For each normalized target spectrum and prescribed $x \in (0,1)$, there is a deterministic $c_\eta \in [1,2]$ such that the schedule in Equation (32), with $C_\eta = c_\eta$, satisfies the same initialization and eventual-error bounds and has success probability at least 0.995 over $\boldsymbol{O}$ alone. The amplitude is chosen before $\boldsymbol{O}$ and works for every deterministic α_ε satisfying Equation (31). It may depend on the normalized spectrum, x, k, and μ, but not on $\boldsymbol{O}$ or the particular admissible initialization scale.

Arbitrary admissible decay factors. The canonical choices above can be replaced by any fixed $\rho \in [\underline{\rho}_\mu, 1)$, where

$$\underline{\rho}_0 = 2/3, \qquad \underline{\rho}_\mu = \rho_\mu \quad (0 < \mu < 1),$$

and the positive-momentum threshold is defined in Equation (51). In the wide regime, replace ρ_μ by ρ in Equations (27) and (28); the envelope constant C_μ can be chosen uniformly over this entire interval. In the undercomplete regime, use Equation (31) with $c_{\text{init}} = c_{\text{iid}}(d, r, k, \kappa, \mu, \rho) > 0$ and $q_{\text{init}} = q_{\text{iid}}(d, k, \mu, \rho) < \infty$, draw independent $C_{\eta,t} \sim \mathsf{Unif}[1,2]$ independently of $\boldsymbol{O}$, and set

$$\eta_0 = C_{\eta,0} a_{\max}, \qquad \eta_t = 4(1-\rho) C_{\eta,t} a_{\max} \rho^{t-1}, \quad t \geq 1. \tag{39}$$

The eventual-error guarantee again has probability at least 0.99. For either regime the iteration bound is

$$T_\varepsilon = O_\mu\left(\frac{1 + \log(\lambda^\star_{\max}/\varepsilon)}{1-\rho} \right).$$

The zero-momentum proof already allows every $\rho \in [2/3, 1)$. For positive momentum, Appendix B.9 gives the exact horizon and initialization conditions for this independent-amplitude extension. The one-amplitude refinement above retains its screened ratio $\widehat{\rho}_\mu$: its proof uses the nonvanishing and finite-word separation conditions in Equations (144) and (147), which are not consequences of a lower threshold on ρ alone.

Consequences and schedule interpretation. The detailed statement and its arbitrary-decay extension give Theorem 1: the wide bound holds along one trajectory for all times, whereas the undercomplete initialization is chosen for a prescribed accuracy and gives the stated eventual-error guarantee. The latter does not assert an all-time geometric envelope from one fixed initialization.

The theorem measures operator reconstruction error. If $\boldsymbol{E}_T = \boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star$, then

$$f(\boldsymbol{U}_T) = \frac{1}{4}\|\boldsymbol{E}_T\|_{\mathrm{F}}^2 \le \frac{d}{4}\|\boldsymbol{E}_T\|^2. \tag{40}$$

Thus the wide objective bound has factor ρ^2, with a dimension-dependent prefactor. In the undercomplete regime, $f(\boldsymbol{U}_t) \le d\varepsilon^2/4$ for every $t \ge T_\varepsilon$ on the success event.

We also briefly explain the rationale for using exponentially decaying learning rates. In contrast to `GD`—where the update norm depends on both the gradient norm and the learning rate—a nonstationary `Muon` step has fixed spectral norm: whenever $\boldsymbol{B}_t \neq \boldsymbol{0}$, $\|\eta_t\,\mathsf{msign}(\boldsymbol{B}_t)\| = \eta_t$. Consequently, the geometrically decaying schedule makes the spectral-norm step length decrease geometrically.

B.2 Absolute nearest-root control at zero momentum

For $a \ge 0$, define

$$g_a(u) \coloneqq (u^2 - a^2)u, \qquad D_a(u) \coloneqq \big||u| - a\big|. \tag{41}$$

Lemma 1 (Nearest-root tail bound). *Fix $a \ge 0$. Let $A > 0$, $2/3 \le \rho < 1$, and $\{c_t\}_{t\ge 0} \subset [1,2]$. Set*

$$\eta_t = c_t A\rho^t, \qquad S_t = \sum_{s=t}^{\infty} \eta_s, \tag{42}$$

and consider

$$u_{t+1} = u_t - \eta_t\,\mathsf{sign}(g_a(u_t)). \tag{43}$$

If $D_a(u_0) \le S_0$ and the trajectory does not hit the nonoptimal stationary point $u_t = 0$ when $a > 0$, then

$$D_a(u_t) \le S_t \le \frac{2A}{1-\rho}\rho^t, \qquad t \ge 0. \tag{44}$$

If $u_0 \neq 0$ is fixed before the prefactors are drawn independently from an atomless distribution on $[1,2]$, the required non-hitting event has probability one.

Proof. Write $x = |u_t|$ and $D = D_a(u_t)$. Whenever u_t is not the nonoptimal stationary point 0 for a positive target, a direct case split gives

$$D_a\big(u_t - \eta_t\,\mathsf{sign}(g_a(u_t))\big) \le |D - \eta_t|. \tag{45}$$

Indeed, if $0 < x < a$, the update increases x toward a and equality holds. If $x > a$ and the update does not cross zero, equality again holds. If it does cross zero, then

$$D_a(u_{t+1}) = |\eta_t - x - a| \le \eta_t - x + a = |\eta_t - (x-a)|. \tag{46}$$

At an exact root the update is zero and Equation (45) is immediate. The only excluded case is $x = 0 < a$, where the scalar gradient vanishes away from the target.

The prefactor bounds and $\rho \ge 2/3$ imply

$$\eta_t \le 2A\rho^t \le \frac{A\rho^{t+1}}{1-\rho} \le S_{t+1}. \tag{47}$$

Since $S_t = \eta_t + S_{t+1}$, the inductive hypothesis $D_a(u_t) \le S_t$ and Equation (45) give

$$D_a(u_{t+1}) \le \max\{\eta_t, S_t - \eta_t\} = S_{t+1}. \tag{48}$$

The geometric upper bound in Equation (44) follows from $c_s \le 2$.

For the null-event claim, there is nothing to prove when $a = 0$. Suppose $a > 0$ and condition on the prefactor history. Unless the coordinate is already at a nonzero root, the equation $u_{t+1} = 0$ determines at most one value of the fresh atomless prefactor. Induction and a countable union over time prove the assertion. Exact nonzero roots are harmless because the coordinate then remains optimal. □

The elementary conversion used repeatedly below is

$$|u^2 - a^2| = D_a(u)(|u| + a) \le D_a(u)(D_a(u) + 2a). \tag{49}$$

B.3 SCALE-FREE FIXED-MOMENTUM RELAY

We first bound the discrepancy between the momentum buffer and the current cubic gradient. This confines sign disagreement to neighborhoods of zero and the target roots. A displacement argument then brings the trajectory back to a target-root neighborhood after each excursion. The tracking and band constants depend only on momentum; the renewal constants may also depend on the chosen decay factor, but none depend on the target spectrum. Fix $0 < \mu < 1$ and put

$$
\begin{aligned}
q_\mu &:= \frac{1+\mu}{2}, & r_\mu &:= \left(\frac{2\mu}{1+\mu}\right)^{1/3}, \\
B_1(\mu) &:= \frac{2q_\mu}{(1-q_\mu)^2}, & B_2(\mu) &:= \frac{4q_\mu(1+q_\mu)}{(1-q_\mu)^3}, \\
B_3(\mu) &:= \frac{8q_\mu(1+4q_\mu+q_\mu^2)}{(1-q_\mu)^4}, & C_{\mathrm{tr}}(\mu) &:= 4B_1(\mu)+3B_2(\mu)+B_3(\mu), \\
L_\mu &:= 2+\max\left\{4, 20C_{\mathrm{tr}}(\mu), \left(4C_{\mathrm{tr}}(\mu)\right)^{1/3}\right\}, & P_\mu &:= 3L_\mu+2.
\end{aligned} \tag{50}
$$

Define canonically

$$
\delta_\mu := \min\left\{\frac{1}{4}, \frac{1-r_\mu}{2}, \frac{1}{4P_\mu+2}\right\}, \qquad \rho_\mu := 1-\frac{\delta_\mu}{2}. \tag{51}
$$

The arguments below work for every ρ satisfying the first three inequalities in the following display. In particular, every $\rho \in [\rho_\mu, 1)$ satisfies them:

$$
\begin{aligned}
&\rho > \frac{3}{4}, \qquad \frac{\mu}{\rho^3} < q_\mu, \qquad \frac{\rho}{1-\rho} > 2P_\mu, \\
&\rho_\mu \geq \frac{3+r_\mu}{4} \geq \frac{1+\mu^{1/3}}{2} \geq \frac{1+\mu}{2}.
\end{aligned} \tag{52}
$$

The last chain uses $\delta_\mu \leq (1-r_\mu)/2$, $r_\mu \geq \mu^{1/3}$, and $\mu^{1/3} \geq \mu$. Finally, set

$$
N_{\mu,\rho} := \left\lceil \frac{\log 2}{-\log \rho} \right\rceil, \qquad K_{\mu,\rho} := P_\mu \rho^{-(N_{\mu,\rho}+1)}. \tag{53}
$$

Since $\rho^{N_{\mu,\rho}} \leq 1/2$,

$$
\sum_{j=1}^{N_{\mu,\rho}} \rho^j = \frac{\rho(1-\rho^{N_{\mu,\rho}})}{1-\rho} > P_\mu. \tag{54}
$$

Consider the scalar recursion

$$
\begin{aligned}
b_t &= \mu b_{t-1} + g_a(u_t), & u_{t+1} &= u_t - c_t h_t \,\mathsf{sign}(b_t), & b_{-1} &= 0, \\
g_a(u) &= (u^2-a^2)u, & h_t &= A\rho^t, & c_t &\in [1,2].
\end{aligned} \tag{55}
$$

Lemma 2 (Scale-free tracking of the current cubic)**.** *Let*

$$
W_t = \sum_{j=0}^{t} \mu^j, \qquad y_t = \frac{b_t}{W_t}. \tag{56}
$$

Every path of Equation (55) *satisfies*

$$
|y_t - g_a(u_t)| \leq C_{\mathrm{tr}}(\mu)\left[(u_t^2+a^2)h_t + h_t^3\right], \qquad t \geq 0. \tag{57}
$$

Proof. For $j \geq 0$, put

$$
Q_j = 2\sum_{\ell=1}^{j} \rho^{-\ell}, \qquad Q_0 = 0, \qquad M_m = \sum_{j=0}^{\infty} \mu^j Q_j^m. \tag{58}
$$

For $0 \le j \le t$, the path-length bound gives

$$|u_{t-j} - u_t| \le 2 \sum_{s=t-j}^{t-1} h_s = h_t Q_j. \tag{59}$$

The sums arising from y_t run only over $0 \le j \le t$ and are bounded by the corresponding infinite sums M_m. Moreover, $Q_j \le 2j\rho^{-j}$ and Equation (52) imply, for $m = 1, 2, 3$,

$$M_m \le 2^m \sum_{j \ge 0} j^m \left(\frac{\mu}{\rho^m}\right)^j \le B_m(\mu). \tag{60}$$

Using

$$g_a(u + \Delta) - g_a(u) = (3u^2 - a^2)\Delta + 3u\Delta^2 + \Delta^3 \tag{61}$$

and $3|u|h_t^2 \le \frac{3}{2}(u^2 h_t + h_t^3)$, we obtain

$$\begin{aligned} |y_t - g_a(u_t)| &\le \left(3M_1 + \frac{3}{2}M_2\right)(u_t^2 + a^2)h_t + \left(\frac{3}{2}M_2 + M_3\right)h_t^3 \\ &\le (4M_1 + 3M_2 + M_3)\left[(u_t^2 + a^2)h_t + h_t^3\right]. \end{aligned} \tag{62}$$

Now use Equation (60). □

Lemma 3 (Two local cubic lower bounds). *For $x = |u|$, $d = |x - a|$, and $m = \min\{x, d\}$,*

$$|g_a(u)| = xd(x + a) \ge \frac{m}{5}(x^2 + a^2), \qquad |g_a(u)| \ge m^3. \tag{63}$$

Proof. The second inequality follows from $x, d, x + a \ge m$. If $x \le d$ and $a > 0$, then $a \ge 2x$ and $d \ge a/2$, so $xd(x + a) \ge xa^2/2 \ge x(x^2 + a^2)/5$; the case $a = 0$ is immediate. If $d \le x$ and $x \ge a$, then $xd(x + a) \ge dx^2 \ge d(x^2 + a^2)/2$. If $d \le x < a$, then $a \le 2x$ and $xd(x + a) \ge dx^2 \ge d(x^2 + a^2)/5$. □

Define the moving root and zero bands

$$\mathcal{R}_t = \{D_a(u_t) \le L_\mu h_t\}, \qquad \mathcal{Z}_t = \{|u_t| \le L_\mu h_t\}. \tag{64}$$

Lemma 4 (Localization of sign disagreement). *If $\mathsf{sign}(b_t) \ne \mathsf{sign}(g_a(u_t))$, then $\mathcal{R}_t$ or $\mathcal{Z}_t$ holds.*

Proof. Since $W_t > 0$, a sign mismatch implies $|g_a(u_t)| \le |y_t - g_a(u_t)|$. If both $|u_t| > L_\mu h_t$ and $D_a(u_t) > L_\mu h_t$, then Lemma 3 gives

$$\begin{aligned} |g_a(u_t)| &> \max\left\{\frac{L_\mu}{5}(u_t^2 + a^2)h_t, L_\mu^3 h_t^3\right\} \\ &\ge \frac{L_\mu}{10}(u_t^2 + a^2)h_t + \frac{L_\mu^3}{2}h_t^3 \\ &> C_{\mathrm{tr}}(\mu)\left[(u_t^2 + a^2)h_t + h_t^3\right], \end{aligned} \tag{65}$$

contradicting Lemma 2. □

Outside $\mathcal{R}_t \cup \mathcal{Z}_t$, the update points toward the nearest signed root. If its update segment crosses a root, then

$$D_a(u_{t+1}) \le c_t h_t \le 2h_t < L_\mu h_{t+1}, \tag{66}$$

where $L_\mu > 2/\rho$.

Lemma 5 (Central momentum relay). *Fix q with $|u_q| < a$ and suppose the buffer is nonzero at every nonoptimal state before the next signed-root crossing. For every $n \ge q$, if*

$$\sum_{t=q}^{n} c_t h_t \ge a + |u_q|, \tag{67}$$

then some update segment with index in $\{q, \dots, n\}$ crosses $+a$ or $-a$.

Proof. By odd symmetry, take $u_q \geq 0$. If $u_q = 0$, the nonzero buffer selects one direction; on the corresponding interval between zero and a signed root, every new gradient reinforces that buffer sign. Thus the required geometric displacement is a.

Now suppose $0 < u_q < a$. If $b_q < 0$, the path moves right. While it remains in $(0, a)$, all gradients are negative, so the buffer stays negative and the geometric displacement to $+a$ is $a - u_q$. If $b_q > 0$, the path initially moves left. Either the buffer becomes negative before or upon reaching zero, or the path crosses zero while the buffer remains positive. In the first case, if the inward travel is $r \leq u_q$, negative gradients keep the reversed buffer negative and the total geometric route to $+a$ is

$$r + (a - u_q + r) \leq a + u_q. \tag{68}$$

In the second case, positive gradients on $(-a, 0)$ reinforce the positive buffer, and the geometric route through zero to $-a$ is $a + u_q$. Consequently a root-crossing update must occur once the cumulative step length reaches the budget in Equation (67). A terminal discrete step may overshoot the root; the lemma bounds the required geometric displacement, not the full length of that overshooting step. □

Lemma 6 (Root-band renewal)**.** *Assume that the buffer is nonzero at every nonoptimal state. If $\mathcal{R}_\tau$ holds at some time τ, then*

$$D_a(u_t) \leq K_{\mu,\rho} h_t, \qquad t \geq \tau. \tag{69}$$

Proof. Take an anchor time s at which $\mathcal{R}_s$ holds. If $\mathcal{R}_{s+1}$ also holds, advance the anchor. Otherwise,

$$D_a(u_{s+1}) \leq D_a(u_s) + c_s h_s \leq (L_\mu + 2) h_s. \tag{70}$$

If $\mathcal{Z}_{s+1}$ does not hold, then the state is outside both bands. By Lemma 4, every subsequent update points toward the nearest root until the path returns to a root band. An update toward the nearest root cannot newly enter the zero band without either crossing a root or entering the root band: in the central interval it increases $|u|$, while on an outer branch reaching the zero band requires crossing the intervening root. Hence the required displacement budget is at most $(L_\mu + 2)h_s < P_\mu h_s$.

If $\mathcal{Z}_{s+1}$ holds but $\mathcal{R}_{s+1}$ does not, then $|u_{s+1}| < a$ and

$$\begin{aligned} a &\leq D_a(u_s) + |u_s| \\ &\leq L_\mu h_s + |u_{s+1}| + c_s h_s \leq \big(L_\mu(1+\rho) + 2\big) h_s. \end{aligned} \tag{71}$$

The relay lemma bounds the required displacement by

$$|u_{s+1}| + a \leq \big(L_\mu(1 + 2\rho) + 2\big) h_s \leq P_\mu h_s. \tag{72}$$

The lower step lengths at updates $s+1, \ldots, s+N_{\mu,\rho}$ sum to more than $P_\mu h_s$ by Equation (54). In the direct case, either the path reaches a root band earlier, giving the next anchor, or every update through $s + N_{\mu,\rho}$ remains outside both bands and points toward the same nearest root; the displacement budget then forces a root-crossing update. In the central case, the relay lemma directly forces such a crossing. Therefore a new root-band anchor exists no later than time $s + N_{\mu,\rho} + 1$, with Equation (66) covering either crossing case.

During the excursion the root distance is at most $P_\mu h_s$: in the direct case it decreases from Equation (70); in the central case it is at most a, which is bounded by Equation (71); and after the terminal crossing it is at most one step. Repeating the construction shows that every $t \geq \tau$ has an anchor $s \in [t - (N_{\mu,\rho} + 1), t]$, and therefore

$$D_a(u_t) \leq P_\mu h_s \leq P_\mu \rho^{-(N_{\mu,\rho}+1)} h_t = K_{\mu,\rho} h_t. \tag{73}$$

□

Lemma 7 (Finite scale-free capture)**.** *Suppose $u_0 > 0$,*

$$D_a(u_0) \leq D, \qquad \frac{A}{1-\rho} \geq 2D. \tag{74}$$

Then a root band is reached by time $N_{\mu,\rho}$ and

$$D_a(u_t) \leq K_{\mu,\rho} A \rho^t, \qquad t \geq N_{\mu,\rho}, \tag{75}$$

provided no nonoptimal buffer zero occurs.

Proof. If $u_0 = a$, the initial point is already in the root band. Otherwise, before the first crossing of the root on the initial side, all gradients have one sign, so the momentum buffer has that sign and the path is monotone toward the root. The first $N_{\mu,\rho}$ lower step lengths obey

$$\sum_{t=0}^{N_{\mu,\rho}-1} h_t = \frac{A(1-\rho^{N_{\mu,\rho}})}{1-\rho} \geq \frac{A}{2(1-\rho)} \geq D. \tag{76}$$

Thus a root-crossing update occurs by index $N_{\mu,\rho} - 1$, and Equation (66) gives a root-band state by time $N_{\mu,\rho}$. Apply Lemma 6. □

Lemma 8 (Almost-sure exclusion of nonoptimal buffer zeros)**.** *Fix $a \geq 0$ and suppose u_0 is deterministic or independent of the entire prefactor sequence. Suppose the prefactors are independent and atomless on $[1, 2]$, and $u_0 \neq 0$ when $a > 0$. Then, almost surely, no finite-time buffer zero occurs at a nonoptimal state. An exact target root with zero buffer is harmless.*

Proof. At time zero, $b_0 = g_a(u_0)$. If it vanishes, the state is either an exact target root or the excluded nonoptimal stationary point. An exact root with zero buffer remains fixed under all later updates. Otherwise, inductively condition on a history for which $b_{t-1} \neq 0$. The fresh prefactor enters u_t affinely with nonzero slope, so $b_t = \mu b_{t-1} + g_a(u_t)$ is a nonzero cubic in that atomless variable. Its zero set has conditional probability zero. A countable union over time proves the claim. □

B.4 The condition-number-free wide regime at zero momentum

For the theorem take $\rho = \rho_0 = 3/4$. The argument below proves the stronger statement for every fixed $\rho \in [2/3, 1)$. Let $k \geq d$, let $\boldsymbol{O} \in \mathbb{R}^{d\times k}$ satisfy $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$, and set $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ with $0 < \alpha \leq a_{\max}$. Complete the target eigenvectors to $\boldsymbol{V} \in \mathcal{O}_{d\times d}$, put $\boldsymbol{R} = \boldsymbol{O}^\top\boldsymbol{V} \in \mathcal{O}_{k\times d}$, and write

$$\boldsymbol{M}^\star = \boldsymbol{V}\mathsf{diag}\{a_1^2, \ldots, a_d^2\}\boldsymbol{V}^\top, \qquad 0 \leq a_i \leq a_{\max}. \tag{77}$$

Then $\boldsymbol{U}_0 = \boldsymbol{V}(\alpha\boldsymbol{I}_d)\boldsymbol{R}^\top$, so the exact aligned reduction in Proposition 3 applies with $\mu = 0$.

Draw $C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1, 2]$ and set

$$A_{\mathrm{w}}^{(0)} = 2(1-\rho)a_{\max}, \qquad \eta_t = C_{\eta,t}A_{\mathrm{w}}^{(0)}\rho^t. \tag{78}$$

Every scalar coordinate begins at $u_{i,0} = \alpha \neq 0$. The null-event argument in Lemma 1, followed by a finite union over coordinates, shows that almost surely no positive-target mode ever hits the nonoptimal stationary point zero. Exact target roots and a zero-target coordinate at zero are harmless.

We use a slightly sharper induction than the generic tail estimate. Put $D_{i,t} = D_{a_i}(u_{i,t})$. Initially $D_{i,0} = |\alpha - a_i| \leq a_{\max}$. If $D_{i,t} \leq 2a_{\max}\rho^t$, then Equation (45) and

$$2(1-\rho)a_{\max}\rho^t \leq \eta_t \leq 4(1-\rho)a_{\max}\rho^t \tag{79}$$

give

$$D_{i,t+1} \leq \max\left\{4(1-\rho)a_{\max}\rho^t, 2a_{\max}\rho^t - 2(1-\rho)a_{\max}\rho^t\right\} \tag{80}$$

$$\leq 2a_{\max}\rho^{t+1}, \tag{81}$$

where the last inequality is exactly $\rho \geq 2/3$. Thus this bound holds for every $t \geq 0$. By Equation (49) and $\rho^t \leq 1$,

$$|u_{i,t}^2 - a_i^2| \leq 2a_{\max}\rho^t\left(2a_{\max}\rho^t + 2a_i\right) \leq 8\lambda_{\max}^\star\rho^t. \tag{82}$$

Finally, $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}_d$, and hence the reconstruction norm is the largest scalar squared-root error. This proves the sharper version of Equation (28) with wide-envelope constant 8. In particular, the theorem-level constant may be enlarged so that $C_0 \geq 8$, and the stated iteration bound follows.

B.5 THE FIXED-POSITIVE-MOMENTUM WIDE REGIME

Fix any $\rho \in [\rho_\mu, 1)$, with the threshold from Equation (51). Let $k \geq d$, fix $\boldsymbol{O} \in \mathbb{R}^{d\times k}$ with $\boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d$, and take $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ with $0 < \alpha \leq a_{\max}$. Use

$$A_{\mathrm{w}} = 2(1-\rho)a_{\max}, \qquad \eta_t = C_{\eta,t}A_{\mathrm{w}}\rho^t. \tag{83}$$

Complete the target eigenvectors to $\boldsymbol{V} \in \mathcal{O}_{d\times d}$, put $\boldsymbol{R} = \boldsymbol{O}^\top\boldsymbol{V} \in \mathcal{O}_{k\times d}$, and write

$$\boldsymbol{U}_0 = \boldsymbol{V}(\alpha\boldsymbol{I}_d)\boldsymbol{R}^\top, \qquad \boldsymbol{M}^\star = \boldsymbol{V}\mathsf{diag}\{a_1^2,\dots,a_d^2\}\boldsymbol{V}^\top, \qquad 0 \leq a_i \leq a_{\max}. \tag{84}$$

The exact aligned reduction in Proposition 3 applies. Every coordinate begins at $u_{i,0} = \alpha > 0$, has $D_{a_i}(u_{i,0}) \leq a_{\max}$, and satisfies

$$\frac{A_{\mathrm{w}}}{1-\rho} = 2a_{\max}. \tag{85}$$

Thus Lemma 7 applies simultaneously to all positive and padded zero modes. The no-buffer-zero event in Lemma 8, followed by a finite union over coordinates, has probability one. Exact target roots with zero buffer are absorbing and harmless. Consequently, almost surely,

$$D_{a_i}(u_{i,t}) \leq 2K_{\mu,\rho}(1-\rho)a_{\max}\rho^t, \qquad t \geq N_{\mu,\rho}. \tag{86}$$

It follows that, for $t \geq N_{\mu,\rho}$,

$$|u_{i,t}^2 - a_i^2| \leq \left[4K_{\mu,\rho}(1-\rho) + 4K_{\mu,\rho}^2(1-\rho)^2\right]\lambda^\star_{\max}\rho^t. \tag{87}$$

For the finite prefix, the total upper path length is at most $2A_{\mathrm{w}}/(1-\rho) = 4a_{\max}$, so $|u_{i,t}| \leq 5a_{\max}$ and $|u_{i,t}^2 - a_i^2| \leq 26\lambda^\star_{\max}$. The definition of $N_{\mu,\rho}$ gives $\rho^{-N_{\mu,\rho}} \leq 2/\rho$ and $K_{\mu,\rho} \leq 2P_\mu/\rho^2$. Since $1-\rho \leq 1-\rho_\mu \leq (8P_\mu+4)^{-1}$, the envelope constant can be bounded uniformly over all admissible ρ:

$$\max\left\{4K_{\mu,\rho}(1-\rho) + 4K_{\mu,\rho}^2(1-\rho)^2,\ 26\rho^{-N_{\mu,\rho}}\right\} \leq \max\left\{\frac{1}{\rho^2} + \frac{1}{4\rho^4}, \frac{52}{\rho}\right\} < 80.$$

Thus the same constant $C_\mu \geq 80$ suffices for the wide envelope for every fixed decay factor in $[\rho_\mu, 1)$. The zero-momentum envelope has constant 8 as proved above. Taking $T \geq \lceil\log(C_\mu\lambda^\star_{\max}/\varepsilon)/(-\log\rho)\rceil$ and using $-\log\rho \geq 1-\rho$ proves the arbitrary-decay iteration bound. The canonical result follows by setting $\rho = \rho_\mu$.

B.6 SHARED TOOLS FOR UNDERCOMPLETE MATRIX FACTORIZATION

The following alignment construction applies to both undercomplete branches in Appendix B.1. Fix $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$, with $\boldsymbol{O}$ uniform on $\mathcal{O}_{d\times k}$, and any first step $0 < \eta_0 \leq 2a_{\max}$. The later schedule is specified separately for each branch.

B.6.1 INITIAL SUBSPACE ALIGNMENT

A key property is that after the first `Muon` iteration, $\boldsymbol{U}_1$ is already well aligned with the eigenspace $\boldsymbol{V}^\star$. Note that when initialized at $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$, the gradient takes the following form

$$\boldsymbol{G}_0 := \nabla f(\boldsymbol{U}_0) = (\boldsymbol{U}_0\boldsymbol{U}_0^\top - \boldsymbol{M}^\star)\boldsymbol{U}_0 = \underbrace{-\alpha\boldsymbol{M}^\star\boldsymbol{O}}_{=:\boldsymbol{Q}} + \alpha^3\boldsymbol{O}, \tag{88}$$

where $\boldsymbol{Q}$ denotes the leading term for small enough α. Fix a deterministic Borel convention for ordered singular-value decompositions, including a fixed tie-breaking rule. Let $\boldsymbol{G}_{0,\leq r}$ be the resulting leading rank-r truncation and put $\boldsymbol{G}_{0,>r} = \boldsymbol{G}_0 - \boldsymbol{G}_{0,\leq r}$. The two groups of singular components have orthogonal initial and final spaces, and therefore

$$\mathsf{msign}(\boldsymbol{G}_0) = \mathsf{msign}(\boldsymbol{G}_{0,\leq r}) + \mathsf{msign}(\boldsymbol{G}_{0,>r}). \tag{89}$$

For small enough α, $\mathsf{msign}(\boldsymbol{G}_{0,\leq r})$ and $\mathsf{msign}(\boldsymbol{Q})$ are close in spectral norm on one high-probability event. The proof is postponed to Section F.1.

Lemma 9 (Initial-alignment event)**.** *There exists a universal constant $c_0 \in (0, 1]$ such that the event*

$$\mathcal{E}_{\text{init}} := \left\{ \sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) \geq \frac{c_0 \lambda_r^\star}{\sqrt{dr}} \right\} \tag{90}$$

satisfies $\mathbb{P}(\mathcal{E}_{\text{init}}) \geq 0.995$. On $\mathcal{E}_{\text{init}}$,

$$\left\| \operatorname{msign}(\boldsymbol{G}_{0,\leq r}) - \operatorname{msign}(\boldsymbol{Q}) \right\| \leq \frac{16\alpha^2\sqrt{dr}}{c_0 \lambda_r^\star} \tag{91}$$

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

Work on the single event $\mathcal{E}_{\text{init}}$ throughout the alignment and stability arguments. The standing condition $4\alpha^2 \leq c_0\lambda_r^\star/\sqrt{dr}$ then implies $\alpha^2 < \lambda_r^\star$. Hence $\boldsymbol{G}_0 = \alpha(\alpha^2 \boldsymbol{I} - \boldsymbol{M}^\star)\boldsymbol{O}$ has full column rank, and $\operatorname{msign}(\boldsymbol{G}_0) = \operatorname{msign}(\boldsymbol{G}_{0,\leq r}) + \operatorname{msign}(\boldsymbol{G}_{0,>r})$ is a $d \times k$ column isometry. Apply Lemma 30 with

$$\boldsymbol{A} = \operatorname{msign}(\boldsymbol{G}_{0,\leq r}), \qquad \boldsymbol{C} = \operatorname{msign}(\boldsymbol{G}_{0,>r}), \qquad \boldsymbol{B} = \operatorname{msign}(\boldsymbol{Q}).$$

Under the second initialization restriction stated below, $\alpha \leq c_0\lambda_r^\star/(800\sqrt{\lambda_{\max}^\star dr})$, the preceding perturbation bound gives $\|\boldsymbol{A} - \boldsymbol{B}\| \leq 1/40000 < 1/12$. We therefore use the explicit small-perturbation construction from the proof of that lemma,

$$\boldsymbol{D} := \operatorname{msign}\big((\boldsymbol{I}_d - \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{C}(\boldsymbol{I}_k - \boldsymbol{B}^\top\boldsymbol{B})\big)\,.$$

This is a measurable function of the initialization and is a rank-$(k-r)$ partial isometry, orthogonal to $\boldsymbol{B}$ on both sides, such that $\widehat{\boldsymbol{F}} = \boldsymbol{B} + \boldsymbol{D}$ has orthonormal columns and

$$\|\operatorname{msign}(\boldsymbol{G}_0) - \widehat{\boldsymbol{F}}\| \leq 25 \|\operatorname{msign}(\boldsymbol{G}_{0,\leq r}) - \operatorname{msign}(\boldsymbol{Q})\|.$$

Set $\gamma := \sigma_r(\boldsymbol{Q})/2$ and define $\widetilde{\boldsymbol{G}}_0 := \boldsymbol{Q} + \gamma\boldsymbol{D}$. Because the two summands have orthogonal initial and final spaces, the leading rank-r truncation of $\widetilde{\boldsymbol{G}}_0$ is exactly $\boldsymbol{Q}$ and $\operatorname{msign}(\widetilde{\boldsymbol{G}}_0) = \widehat{\boldsymbol{F}}$. Consequently,

$$\widetilde{\boldsymbol{G}}_{0,\leq r} = \boldsymbol{Q} \qquad \text{and} \tag{92a}$$

$$\left\| \operatorname{msign}(\boldsymbol{G}_0) - \operatorname{msign}(\widetilde{\boldsymbol{G}}_0) \right\| \leq 25 \left\| \operatorname{msign}(\boldsymbol{G}_{0,\leq r}) - \operatorname{msign}(\boldsymbol{Q}) \right\| \leq \frac{400\alpha^2\sqrt{dr}}{c_0\lambda_r^\star}. \tag{92b}$$

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

$$\boldsymbol{U}_1 = \alpha\boldsymbol{O} - \eta_0\big(\operatorname{msign}(\widetilde{\boldsymbol{G}}_0) + \boldsymbol{R}_0\big) = -\eta_0 \operatorname{msign}(\widetilde{\boldsymbol{G}}_0) + \boldsymbol{R}_1, \tag{93a}$$

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

$$\|\boldsymbol{R}_0\| \leq \frac{400\alpha^2\sqrt{dr}}{c_0\lambda_r^\star} \qquad \text{and} \qquad \|\boldsymbol{R}_1\| \leq \alpha + \frac{400\eta_0\alpha^2\sqrt{dr}}{c_0\lambda_r^\star} \leq 2\alpha, \tag{93b}$$

provided that $\eta_0 \leq 2a_{\max}$ and $\alpha \leq c_0\lambda_r^\star/(800\sqrt{\lambda_{\max}^\star dr})$.

B.6.2 Construction of an aligned auxiliary trajectory

To facilitate analysis, construct an auxiliary factor–buffer trajectory $\{(\widetilde{\boldsymbol{U}}_t, \widetilde{\boldsymbol{B}}_{t-1})\}_{t\geq 1}$ as follows:

$$\widetilde{\boldsymbol{U}}_1 = -\eta_0 \operatorname{msign}(\widetilde{\boldsymbol{G}}_0), \tag{94a}$$

$$\widetilde{\boldsymbol{B}}_0 = \boldsymbol{0}, \tag{94b}$$

$$\widetilde{\boldsymbol{B}}_t = \mu\widetilde{\boldsymbol{B}}_{t-1} + \big(\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top - \boldsymbol{M}^\star\big)\widetilde{\boldsymbol{U}}_t, \qquad t \geq 1, \tag{94c}$$

$$\widetilde{\boldsymbol{U}}_{t+1} = \widetilde{\boldsymbol{U}}_t - \eta_t \operatorname{msign}(\widetilde{\boldsymbol{B}}_t), \qquad t \geq 1. \tag{94d}$$

The auxiliary factor discards $\boldsymbol{R}_1$, and its previous buffer is reset to zero at time one. The actual algorithm retains its true buffer $\boldsymbol{B}_0$. In particular, Equation (93b) gives

$$\left\|\widetilde{\boldsymbol{U}}_1 - \boldsymbol{U}_1\right\| = \|\boldsymbol{R}_1\| \leq 2\alpha. \tag{95}$$

Next, we show that the auxiliary trajectory $\{\widetilde{\boldsymbol{U}}_t\}_{t\geq 1}$ decomposes into scalar coordinate dynamics. The coordinates share the same scalar stepsize but are dynamically decoupled. On $\mathcal{E}_{\text{init}}$, $\operatorname{msign}(\boldsymbol{Q})$ can be decomposed as

$$\operatorname{msign}(\boldsymbol{Q}) = \boldsymbol{V}^\star \boldsymbol{O}'^\top \tag{96}$$

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

Proof of property (96). Observe that

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

where $\boldsymbol{W} \coloneqq \boldsymbol{V}^{\star\top} \boldsymbol{O}$. On $\mathcal{E}_{\text{init}}$, $\sigma_r(\boldsymbol{\Lambda}^\star \boldsymbol{W}) > 0$. Define explicitly $\boldsymbol{O}' \coloneqq -\operatorname{msign}(\boldsymbol{\Lambda}^\star \boldsymbol{W})^\top \in \mathcal{O}_{k\times r}$; this gives Equation (96). This completes the proof. □

Fix once and for all deterministic Borel completion maps $\operatorname{Comp}_{m,s}$ that augment any element of $\mathcal{O}_{m\times s}$ to an $m \times m$ orthogonal matrix, for example lexicographically pivoted Gram–Schmidt applied to the standard basis with a fixed sign convention. On $\mathcal{E}_{\text{init}}$, set $\boldsymbol{R} \coloneqq \operatorname{Comp}_{k,r}(\boldsymbol{O}')$ and $\boldsymbol{V} \coloneqq \operatorname{msign}(\widetilde{\boldsymbol{G}}_0)\boldsymbol{R}$. Then $\boldsymbol{V} \in \mathcal{O}_{d\times k}$, $\boldsymbol{R} \in \mathcal{O}_{k\times k}$, and

$$\boldsymbol{V}_{:,1:r} = \boldsymbol{V}^\star, \qquad \boldsymbol{R}_{:,1:r} = \boldsymbol{O}', \qquad \text{and} \qquad \operatorname{msign}(\widetilde{\boldsymbol{G}}_0) = \boldsymbol{V}\boldsymbol{R}^\top, \tag{97}$$

Here $\boldsymbol{A}_{:,1:r}$ denotes the first r columns of a generic matrix $\boldsymbol{A}$. The first identity follows from $\operatorname{msign}(\widetilde{\boldsymbol{G}}_0)\boldsymbol{O}' = \operatorname{msign}(\boldsymbol{Q})\boldsymbol{O}' = \boldsymbol{V}^\star$, while the completion component annihilates $\boldsymbol{O}'$. Assign fixed arbitrary values to all auxiliary objects off $\mathcal{E}_{\text{init}}$. This makes the construction globally measurable as a function of $\boldsymbol{O}$ alone. Moreover, Condition (97) allows us to express the ground truth as

$$\boldsymbol{M}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top} = \boldsymbol{V} \boldsymbol{\Lambda}_{\mathsf{aug}} \boldsymbol{V}^\top, \tag{98}$$

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

Recall that $\widetilde{\boldsymbol{U}}_1 = -\eta_0 \operatorname{msign}(\widetilde{\boldsymbol{G}}_0) = -\eta_0 \boldsymbol{V}\boldsymbol{R}^\top$ (cf. (94a)). Combining this with Equation (98), we apply Proposition 3, with $s = k$, to the time-shifted auxiliary path. Thus, for each $t \geq 1$, $\widetilde{\boldsymbol{U}}_t$ can be decomposed as

$$\widetilde{\boldsymbol{U}}_t = \boldsymbol{V} \widetilde{\boldsymbol{\Sigma}}_t \boldsymbol{R}^\top \qquad \text{for some } \widetilde{\boldsymbol{\Sigma}}_t = \operatorname{diag}\{\widetilde{\sigma}_{1,t}, \dots, \widetilde{\sigma}_{k,t}\} \in \mathbb{R}^{k\times k}, \tag{99a}$$

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

$$\widetilde{b}_{i,t} = \mu \widetilde{b}_{i,t-1} + (\widetilde{\sigma}_{i,t}^2 - \lambda_i^\star)\widetilde{\sigma}_{i,t}, \qquad \widetilde{b}_{i,0} = 0, \tag{99b}$$

$$\widetilde{\sigma}_{i,t+1} = \widetilde{\sigma}_{i,t} - \eta_t \operatorname{sign}(\widetilde{b}_{i,t}). \tag{99c}$$

B.6.3 Fixed-frame differential and Taylor estimates

Normalize factors, buffers, the target, and stepsizes by $a_{\max}$, $a_{\max}^3$, $a_{\max}^2$, and $a_{\max}$, respectively, and suppress the normalization in the notation. Let $\ell_t > 0$ satisfy $\ell_{t+1} = \rho \ell_t$ and $\ell_t \leq \eta_t \leq 2\ell_t$. For the single-amplitude proof we may take $\ell_t = \eta_t$; the proofs with independent amplitudes use their deterministic base steps.

Fix the measurable frames $(\boldsymbol{V}, \boldsymbol{R})$ from Equation (97) once and for all and set

$$\boldsymbol{V}_{\text{full}} \coloneqq \operatorname{Comp}_{d,k}(\boldsymbol{V}) \in \mathcal{O}_{d\times d}. \tag{100}$$

For every normalized factor–buffer pair, define its fixed-frame representation by

$$\boldsymbol{U}^\circ \coloneqq \boldsymbol{V}_{\text{full}}^\top \boldsymbol{U}\boldsymbol{R}, \qquad \boldsymbol{H}^\circ \coloneqq \boldsymbol{V}_{\text{full}}^\top \boldsymbol{H}\boldsymbol{R}. \tag{101}$$

We henceforth work in these coordinates and continue to write $(\boldsymbol{U}, \boldsymbol{H})$ for a generic represented pair. In particular,

$$\overline{\boldsymbol{U}}_t \coloneqq \boldsymbol{V}_{\text{full}}^\top \widetilde{\boldsymbol{U}}_t \boldsymbol{R} = \begin{bmatrix} \operatorname{diag}\{\bar{u}_{i,t}\} \\ \boldsymbol{0} \end{bmatrix}, \qquad \overline{\boldsymbol{H}}_{t-1} \coloneqq \boldsymbol{V}_{\text{full}}^\top \widetilde{\boldsymbol{B}}_{t-1} \boldsymbol{R} = \begin{bmatrix} \operatorname{diag}\{\bar{h}_{i,t-1}\} \\ \boldsymbol{0} \end{bmatrix}, \tag{102}$$

and the target in the left frame is

$$\boldsymbol{V}_{\text{full}}^\top \boldsymbol{M}^\star \boldsymbol{V}_{\text{full}} = \operatorname{diag}\{a_1^2, \dots, a_k^2, 0, \dots, 0\}. \tag{103}$$

We also suppress the circle on this represented target and continue to denote it by $\boldsymbol{M}^\star$ in the coordinate state map below. The aligned scalar buffer coordinates obey

$$\bar{b}_{i,t} = \mu \bar{h}_{i,t-1} + (\bar{u}_{i,t}^2 - a_i^2)\bar{u}_{i,t}. \tag{104}$$

Accordingly,

$$\overline{\boldsymbol{B}}_t \coloneqq \mu\overline{\boldsymbol{H}}_{t-1} + \mathcal{G}(\overline{\boldsymbol{U}}_t) = \begin{bmatrix} \mathsf{diag}\{\bar{b}_{i,t}\} \\ \mathbf{0} \end{bmatrix}. \tag{105}$$

Here $\boldsymbol{H}$ denotes the previous buffer, so the one-step state map is

$$\mathcal{T}_t(\boldsymbol{U}, \boldsymbol{H}) = (\boldsymbol{U} - \eta_t\, \mathsf{msign}(\boldsymbol{B}),\, \boldsymbol{B}), \qquad \boldsymbol{B} = \mu\boldsymbol{H} + \mathcal{G}(\boldsymbol{U}), \tag{106}$$

$$\mathcal{G}(\boldsymbol{U}) = (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star)\boldsymbol{U}. \tag{107}$$

The represented actual state at time one is $(\boldsymbol{U}_1^\circ, \boldsymbol{B}_0^\circ)$; the represented auxiliary state is $(\overline{\boldsymbol{U}}_1, \mathbf{0})$.

A fixed-frame weighted norm. Define

$$w_{i,t} = \bar{u}_{i,t}^2 + a_i^2 + \ell_t^2, \qquad w_{ij,t} = w_{i,t} + w_{j,t}. \tag{108}$$

For a factor perturbation $\boldsymbol{X}$ and previous-buffer perturbation $\boldsymbol{Y}$, write $\boldsymbol{Y}_\parallel \in \mathbb{R}^{k\times k}$ and $\boldsymbol{Y}_\perp \in \mathbb{R}^{(d-k)\times k}$ for the top and bottom blocks in the fixed frames, and set

$$\mathcal{N}_t(\boldsymbol{Y})^2 = \sum_{i,j=1}^{k} \frac{|(\boldsymbol{Y}_\parallel)_{ij}|^2}{w_{ij,t}^2} + \sum_{\gamma=1}^{d-k}\sum_{i=1}^{k} \frac{|(\boldsymbol{Y}_\perp)_{\gamma i}|^2}{w_{i,t}^2}, \tag{109}$$

$$\|(\boldsymbol{X}, \boldsymbol{Y})\|_{\star,t} = \|\boldsymbol{X}\|_{\mathrm{F}} + \mathcal{N}_t(\boldsymbol{Y}). \tag{110}$$

This is an ordinary norm on the factor–buffer pair, not a moving quotient metric. In particular, vertical right-rotation errors are included rather than regauged.

Lemma 10 (Full-column polar differential)**.** *Let $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{H} \in \mathbb{R}^{d\times k}$ have full column rank, where $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_k$ and $\boldsymbol{H} = (\boldsymbol{B}^\top\boldsymbol{B})^{1/2} \succ 0$. For every $\boldsymbol{Z} \in \mathbb{R}^{d\times k}$,*

$$\mathrm{D}\,\mathsf{msign}(\boldsymbol{B})[\boldsymbol{Z}] = (\boldsymbol{I}_d - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{Z}\boldsymbol{H}^{-1} + \boldsymbol{Q}\boldsymbol{\Omega}, \tag{111}$$

$$\boldsymbol{H}\boldsymbol{\Omega} + \boldsymbol{\Omega}\boldsymbol{H} = \boldsymbol{Q}^\top\boldsymbol{Z} - \boldsymbol{Z}^\top\boldsymbol{Q}, \qquad \boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}. \tag{112}$$

At an aligned buffer $\boldsymbol{B} = [\mathsf{diag}\{b_i\}; \mathbf{0}]$, put $v_i = |b_i|$ and $q_i = \mathsf{sign}(b_i)$. Then

$$\big(\mathrm{D}\,\mathsf{msign}(\boldsymbol{B})[\boldsymbol{Z}]\big)_{k+\gamma,i} = \frac{\boldsymbol{Z}_{k+\gamma,i}}{v_i}, \qquad 1 \le \gamma \le d-k, \tag{113}$$

$$\big(\mathrm{D}\,\mathsf{msign}(\boldsymbol{B})[\boldsymbol{Z}]\big)_{ij} = \frac{\boldsymbol{Z}_{ij} - q_i q_j \boldsymbol{Z}_{ji}}{v_i + v_j}, \qquad 1 \le i, j \le k, \tag{114}$$

$$|\boldsymbol{\Omega}_{ij}| \le \frac{|\boldsymbol{Z}_{ij}| + |\boldsymbol{Z}_{ji}|}{v_i + v_j}, \qquad 1 \le i, j \le k. \tag{115}$$

Proof. Differentiate $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{H}$ and put $\boldsymbol{\Omega} = \boldsymbol{Q}^\top\mathrm{D}\boldsymbol{Q}[\boldsymbol{Z}]$. Differentiating $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_k$ shows that $\boldsymbol{\Omega}$ is skew-symmetric. The component orthogonal to the columns of $\boldsymbol{Q}$ gives the first term in Equation (111). Subtracting the transpose of $\boldsymbol{Q}^\top\boldsymbol{Z} = \boldsymbol{\Omega}\boldsymbol{H} + \mathrm{D}\boldsymbol{H}[\boldsymbol{Z}]$ gives Equation (112); its solution is unique because the eigenvalues of $\boldsymbol{H}$ are positive. At the aligned buffer, $\boldsymbol{Q} = [\mathsf{diag}\{q_i\}; \mathbf{0}]$ and $\boldsymbol{H} = \mathsf{diag}\{v_i\}$. Thus the orthogonal component gives Equation (113), while

$$(v_i + v_j)\boldsymbol{\Omega}_{ij} = q_i\boldsymbol{Z}_{ij} - q_j\boldsymbol{Z}_{ji}, \tag{116}$$

which proves Equations (114) and (115). □

Lemma 11 (Weighted factor–buffer differential)**.** *At the aligned state in Equation (102), assume $\bar{b}_{i,t} \neq 0$ for all i. Define*

$$\overline{R}_t \coloneqq 2\ell_t \max_{1\le i\le k} \frac{w_{i,t}}{|\bar{b}_{i,t}|}. \tag{117}$$

There is a constant $C_{\mathrm{lin}} = C_{\mathrm{lin}}(d, k, \rho) \ge 1$ such that

$$\|\mathrm{D}\mathcal{T}_t(\overline{\boldsymbol{U}}_t, \overline{\boldsymbol{H}}_{t-1})[\boldsymbol{X}, \boldsymbol{Y}]\|_{\star,t+1} \le C_{\mathrm{lin}}(1 + \overline{R}_t)\|(\boldsymbol{X}, \boldsymbol{Y})\|_{\star,t}. \tag{118}$$

Proof. Put $\boldsymbol{Z} = \mu\boldsymbol{Y} + \mathrm{D}\mathcal{G}(\overline{\boldsymbol{U}}_t)[\boldsymbol{X}]$. In the aligned coordinates, direct differentiation gives

$$(\mathrm{D}\mathcal{G}[\boldsymbol{X}])_{ij} = (\bar{u}_{i,t}^2 + \bar{u}_{j,t}^2 - a_i^2)\boldsymbol{X}_{ij} + \bar{u}_{i,t}\bar{u}_{j,t}\boldsymbol{X}_{ji}, \tag{119}$$

$$(\mathrm{D}\mathcal{G}[\boldsymbol{X}])_{k+\gamma,i} = \bar{u}_{i,t}^2\boldsymbol{X}_{k+\gamma,i}, \qquad 1 \le \gamma \le d-k. \tag{120}$$

Every coefficient in the first line is bounded by a fixed multiple of $w_{ij,t}$, and the coefficient in the second line is at most $w_{i,t}$. Consequently,

$$\mathcal{N}_t(\boldsymbol{Z}) \le C\big(\|\boldsymbol{X}\|_{\mathrm{F}} + \mathcal{N}_t(\boldsymbol{Y})\big). \tag{121}$$

Let $\bar{v}_i = |\bar{b}_{i,t}|$ and $q_i = \mathsf{sign}(\bar{b}_{i,t})$. By Lemma 10, the differential at $[\mathsf{diag}\{\bar{b}_{i,t}\}; \boldsymbol{0}]$ has bottom entries $\boldsymbol{Z}_{k+\gamma,i}/\bar{v}_i$ and top skew coordinates bounded by

$$\frac{|\boldsymbol{Z}_{ij}| + |\boldsymbol{Z}_{ji}|}{\bar{v}_i + \bar{v}_j}. \tag{122}$$

Because

$$\frac{\eta_t(w_{i,t} + w_{j,t})}{\bar{v}_i + \bar{v}_j} \le \max_{\ell} \frac{\eta_t w_{\ell,t}}{\bar{v}_\ell} \le \overline{R}_t, \tag{123}$$

we obtain

$$\eta_t \|\mathrm{D}\,\mathsf{msign}(\overline{\boldsymbol{B}}_t)[\boldsymbol{Z}]\|_{\mathrm{F}} \le C\overline{R}_t\mathcal{N}_t(\boldsymbol{Z}). \tag{124}$$

This calculation includes vertical directions. For example, if $\boldsymbol{X} = \overline{\boldsymbol{U}}_t\boldsymbol{\Omega}$ and $\boldsymbol{Y} = \overline{\boldsymbol{H}}_{t-1}\boldsymbol{\Omega}$ with $\boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}$, then $\boldsymbol{Z}_{ij} = \bar{b}_{i,t}\boldsymbol{\Omega}_{ij}$ and $\boldsymbol{Z}_{ji} = -\bar{b}_{j,t}\boldsymbol{\Omega}_{ij}$, so the apparent polar denominator cancels exactly.

It remains to compare consecutive weights. Since $|\bar{u}_{i,t+1} - \bar{u}_{i,t}| \le 2\ell_t$,

$$w_{i,t} \le 2\bar{u}_{i,t+1}^2 + a_i^2 + 9\ell_t^2 \le \frac{9}{\rho^2} w_{i,t+1}. \tag{125}$$

The same bound holds for $w_{ij,t}/w_{ij,t+1}$. Thus $\mathcal{N}_{t+1}(\boldsymbol{Z}) \le (9/\rho^2)\mathcal{N}_t(\boldsymbol{Z})$. Combining this fact with Equations (121) and (124) proves the lemma. □

Lemma 12 (Quantitative polar Taylor remainder)**.** *There is a universal constant $C > 0$ with the following property. If $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ has full column rank, $s = \sigma_{\min}(\boldsymbol{B})$, and $\|\boldsymbol{\Delta}\|_{\mathrm{F}} \le s/2$, then*

$$\big\| \mathsf{msign}(\boldsymbol{B} + \boldsymbol{\Delta}) - \mathsf{msign}(\boldsymbol{B}) - \mathrm{D}\,\mathsf{msign}(\boldsymbol{B})[\boldsymbol{\Delta}] \big\|_{\mathrm{F}} \le \frac{C}{s^2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2. \tag{126}$$

Along the segment $\boldsymbol{B}_\tau = \boldsymbol{B} + \tau\boldsymbol{\Delta}$, one also has $\|\mathrm{D}\,\mathsf{msign}(\boldsymbol{B}_\tau)\|_{\mathrm{F}\leftarrow\mathrm{F}} \le C/s$.

Proof. The full-column polar map is smooth on the open set of full-column-rank matrices. Write $\boldsymbol{B}_\tau = \boldsymbol{Q}_\tau\boldsymbol{H}_\tau$. Weyl's inequality gives $\sigma_{\min}(\boldsymbol{B}_\tau) \ge s/2$. Apply Lemma 10 with the fixed direction $\boldsymbol{\Delta}$. The orthogonal term in Equation (111) is at most $2\|\boldsymbol{\Delta}\|_{\mathrm{F}}/s$. In an eigenbasis of $\boldsymbol{H}_\tau$, the Sylvester equation divides each entry by a sum of two eigenvalues; hence

$$\|\boldsymbol{\Omega}_\tau\|_{\mathrm{F}} \le \frac{2}{s}\|\boldsymbol{\Delta}\|_{\mathrm{F}}, \qquad \|\boldsymbol{Q}_\tau'\|_{\mathrm{F}} \le \frac{4}{s}\|\boldsymbol{\Delta}\|_{\mathrm{F}}. \tag{127}$$

Moreover, $\boldsymbol{Q}_\tau^\top\boldsymbol{\Delta} = \boldsymbol{\Omega}_\tau\boldsymbol{H}_\tau + \boldsymbol{H}_\tau'$. Entrywise in the same eigenbasis, $\|\boldsymbol{\Omega}_\tau\boldsymbol{H}_\tau\|_{\mathrm{F}} \le 2\|\boldsymbol{\Delta}\|_{\mathrm{F}}$, because $h_j/(h_i + h_j) \le 1$. Thus $\|\boldsymbol{H}_\tau'\|_{\mathrm{F}} \le 3\|\boldsymbol{\Delta}\|_{\mathrm{F}}$.

Differentiating Equation (112) gives

$$\boldsymbol{H}_\tau\boldsymbol{\Omega}_\tau' + \boldsymbol{\Omega}_\tau'\boldsymbol{H}_\tau = \boldsymbol{Q}_\tau'^\top\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top\boldsymbol{Q}_\tau' - \boldsymbol{H}_\tau'\boldsymbol{\Omega}_\tau - \boldsymbol{\Omega}_\tau\boldsymbol{H}_\tau'. \tag{128}$$

The preceding bounds and the inverse Sylvester bound therefore yield $\|\boldsymbol{\Omega}_\tau'\|_{\mathrm{F}} \le C\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2/s^2$. Differentiating Equation (111), using $(\boldsymbol{H}_\tau^{-1})' = -\boldsymbol{H}_\tau^{-1}\boldsymbol{H}_\tau'\boldsymbol{H}_\tau^{-1}$, now gives

$$\|\boldsymbol{Q}_\tau''\|_{\mathrm{F}} \le \frac{C}{s^2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2. \tag{129}$$

The integral form of Taylor's theorem proves Equation (126), and Equation (127) gives the derivative bound. □

Lemma 13 (Ambient weighted Taylor bound)**.** *Assume the normalized auxiliary factors and step scales are uniformly bounded, and put* $L_t^{\mathrm{lin}} = C_{\mathrm{lin}}(1 + \overline{R}_t)$*. There are constants* $c_{\mathrm{Tay}}, C_{\mathrm{Tay}} > 0$*, depending only on* d, k, ρ *and these uniform bounds, such that, at every auxiliary time for which* $\min_{i\le k} |\bar{b}_{i,t}| \ge \beta_t, \quad 0 < \beta_t \le 1$*, if*

$$e = \|(\boldsymbol{U} - \overline{\boldsymbol{U}}_t, \boldsymbol{H} - \overline{\boldsymbol{H}}_{t-1})\|_{\star,t} \le c_{\mathrm{Tay}}\beta_t, \tag{130}$$

then

$$\|\mathcal{T}_t(\boldsymbol{U}, \boldsymbol{H}) - \mathcal{T}_t(\overline{\boldsymbol{U}}_t, \overline{\boldsymbol{H}}_{t-1})\|_{\star,t+1} \le L_t^{\mathrm{lin}} e + C_{\mathrm{Tay}}(\ell_{t+1}^{-2} + \beta_t^{-3})e^2. \tag{131}$$

Proof. The assumed factor and step bounds control all weights by constants independent of κ. Thus the weighted norm controls the ordinary factor–buffer Frobenius norm. Put

$$\boldsymbol{X} = \boldsymbol{U} - \overline{\boldsymbol{U}}_t, \qquad \boldsymbol{Y} = \boldsymbol{H} - \overline{\boldsymbol{H}}_{t-1}, \tag{132}$$

$$\boldsymbol{Z} = \mu\boldsymbol{Y} + \mathrm{D}\mathcal{G}(\overline{\boldsymbol{U}}_t)[\boldsymbol{X}], \qquad \boldsymbol{R}_{\mathcal{G}} = \mathcal{G}(\boldsymbol{U}) - \mathcal{G}(\overline{\boldsymbol{U}}_t) - \mathrm{D}\mathcal{G}(\overline{\boldsymbol{U}}_t)[\boldsymbol{X}]. \tag{133}$$

The auxiliary factors are uniformly bounded, and $\mathcal{G}$ is cubic, so on the segment under consideration

$$\|\boldsymbol{X}\|_{\mathrm{F}} + \|\boldsymbol{Y}\|_{\mathrm{F}} \le Ce, \qquad \|\boldsymbol{Z}\|_{\mathrm{F}} \le Ce, \qquad \|\boldsymbol{R}_{\mathcal{G}}\|_{\mathrm{F}} \le Ce^2. \tag{134}$$

The current-buffer perturbation is $\boldsymbol{\Delta} = \boldsymbol{Z} + \boldsymbol{R}_{\mathcal{G}}$. By the displayed buffer lower bound, $\sigma_{\min}(\overline{\boldsymbol{B}}_t) \ge \beta_t$. Reducing c_{Tay} makes $\|\boldsymbol{\Delta}\|_{\mathrm{F}} \le \beta_t/2$, so the whole segment has smallest singular value at least $\beta_t/2$.

Apply Lemma 12 at $\overline{\boldsymbol{B}}_t$. After subtracting the derivative of the state map, the factor-component remainder is

$$-\eta_t\big(\mathrm{D}\,\mathsf{msign}(\overline{\boldsymbol{B}}_t)[\boldsymbol{R}_{\mathcal{G}}] + \boldsymbol{R}_{\mathrm{polar}}\big), \qquad \|\boldsymbol{R}_{\mathrm{polar}}\|_{\mathrm{F}} \le C\beta_t^{-2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2. \tag{135}$$

By the derivative bound in the same lemma and Equation (134), its Frobenius norm is at most $C\eta_t(\beta_t^{-1} + \beta_t^{-2})e^2$. The buffer-component remainder is exactly $\boldsymbol{R}_{\mathcal{G}}$ and has Frobenius norm at most Ce^2. Since $w_{i,t+1} \ge \ell_{t+1}^2$, converting this last term to $\mathcal{N}_{t+1}$ costs at most a fixed dimensional constant times ℓ_{t+1}^{-2}. Finally, η_t is bounded and we may take $\beta_t \le 1$, so the total quadratic remainder is bounded by the second term in Equation (131). The linear term is controlled by Lemma 11; in particular, no global β_t^{-1} bound enters the linear product. □

B.7 Positive momentum with one random stepsize amplitude

Fix $0 < \mu < 1$ and $r \le k < d$. We prove part (b) of the detailed statement in Appendix B.1 with one draw $C_\eta \sim \mathsf{Unif}[1, 2]$, independent of $\boldsymbol{O}$. The proof first screens a decay ratio, then counts scalar sign-history intervals to control the logarithmic inverse-buffer cost of positive modes. An exact cancellation handles the pure zero modes. Together these estimates bound every ambient linear propagator and close a nonlinear comparison.

Normalize $a_{\max} = 1$, let α denote $\alpha_\varepsilon/a_{\max}$, write $x = \varepsilon/\lambda_{\max}^\star$, and let $a_i \in [\kappa^{-1/2}, 1]$ for $i \le r$, with $a_i = 0$ for $r < i \le k$. Retain $C_{\mathrm{tr}}, L_\mu, P_\mu$ from Equation (50), and put

$$C_{\mathrm{tr}} = \frac{8(1+\mu)(\mu^2 + 10\mu + 37)}{(1-\mu)^4}, \qquad L_\mu = 2 + 20C_{\mathrm{tr}}, \qquad P_\mu = 3L_\mu + 2,$$

$$\gamma = (8P_\mu + 4)^{-1}, \qquad \rho_- = 1 - \gamma, \qquad \rho_+ = 1 - \gamma/2,$$

$$I = [\rho_-, \rho_+], \qquad A_0 = 4\gamma, \qquad a_+ = -\log\rho_+. \tag{136}$$

These agree with γ_μ in Equation (26). Every $\rho \in I$ satisfies $\rho > 0.99$, $\mu/\rho^3 < (1+\mu)/2$, and the tracking and renewal conditions in Equations (52) and (54). Here ρ denotes the new ratio $\widehat{\rho}_\mu$ to be selected below, not the canonical wide-regime ratio ρ_μ.

For a candidate amplitude $c \in [1, 2]$, put $A = A_0 c$ and $\eta_t = A\rho^t$ for every $t \ge 0$. The alignment construction in Section B.6 applies because $A \le 1$. On $\mathcal{E}_{\mathrm{init}}$ it gives

$$\|\boldsymbol{U}_1 - \widetilde{\boldsymbol{U}}_1\|_{\mathrm{F}} \le 2\sqrt{k}\alpha, \qquad \|\boldsymbol{B}_0\|_{\mathrm{F}} \le \sqrt{k}\alpha(1 + \alpha^2), \tag{137}$$

provided

$$\alpha \le \frac{c_0}{800\kappa\sqrt{dr}}. \tag{138}$$

The actual buffer remains $\boldsymbol{B}_0$; the auxiliary previous buffer is zero. For this subsection, replace the right frame $\boldsymbol{R}$ in Equation (97) by $-\boldsymbol{R}$, leaving the left frame unchanged. Represent both the actual and auxiliary states in these same fixed frames. The auxiliary scalar coordinates then satisfy

$$u_{i,1} = A, \quad b_{i,0} = 0, \qquad b_{i,t} = \mu b_{i,t-1} + (u_{i,t}^2 - a_i^2)u_{i,t}, \qquad u_{i,t+1} = u_{i,t} - A\rho^t \,\mathsf{sign}(b_{i,t}). \tag{139}$$

We will use the following pathwise consequence of the scalar relay. Set $N_\mu = \lceil \log 2/(-\log \rho)\rceil$ and $K_\mu = P_\mu \rho^{-(N_\mu+1)}$. On any trajectory without buffer zeros through the time being considered,

$$\big||u_{i,t}| - a_i\big| \le K_\mu A \rho^t \qquad (t \ge 1 + N_\mu). \tag{140}$$

Indeed $|A - a_i| \le 1$ and the first N_μ steps after time one have total length at least $A\rho/[2(1-\rho)] \ge 2\rho > 1$. Reindexing time one as zero therefore permits Lemmas 6 and 7. The screening below guarantees that the factors never vanish and excludes buffer zeros on the relevant horizon. For amplitude integrals, all finite-time buffer-zero parameters form a countable null set; the same pathwise argument applies to every remaining parameter. No capture statement at pauses is needed. Moreover $A/(1-\rho) \le 16$ and $K_\mu A \le 16 P_\mu \gamma/\rho^2 < 2/\rho^2$, so

$$\|\widetilde{\boldsymbol{U}}_t\| \le 16, \qquad \|\widetilde{\boldsymbol{U}}_t \widetilde{\boldsymbol{U}}_t^\top - \boldsymbol{M}^\star\| \le 80\rho^t \qquad (t \ge 1 + N_\mu). \tag{141}$$

At the theorem's horizon $T_x = T_\varepsilon$ in Equation (33), the last bound is at most $x/4$.

B.7.1 Decay-ratio screening and finite-block separation

We first fix ρ as a function of momentum alone.

For a word $\sigma \in \{-1, 0, 1\}^{t-1}$, put

$$z_s(\rho) = 1 - \sum_{j=1}^{s-1} \sigma_j \rho^j, \qquad P_{t,\sigma}(\rho) = \sum_{s=1}^{t} \mu^{t-s} z_s(\rho)^3, \qquad Q_{t,\sigma}(\rho) = \sum_{s=1}^{t} \mu^{t-s} z_s(\rho). \tag{142}$$

Along this word,

$$u_t = A z_t(\rho), \qquad b_t = A^3 P_{t,\sigma}(\rho) - a^2 A Q_{t,\sigma}(\rho). \tag{143}$$

Lemma 14 (All-history exponential lower bounds for the ratio)**.** *There are constants $m \in \mathbb{N}$, $D_\mu > 0$, and a compact subset $\mathcal{R} \subset I$ of positive Lebesgue measure, depending only on μ, such that every $\rho \in \mathcal{R}$ satisfies*

$$|z_t(\rho)| \ge \delta_t, \qquad |P_{t,\sigma}(\rho)| \ge \delta_t, \qquad \delta_t = D_\mu \mathrm{e}^{-d_\mu t}, \quad d_\mu = m \log 6, \tag{144}$$

for every t and every word.

Proof. We prove a uniform analytic sublevel estimate for the two polynomial families. Set $r_0 = \rho_+$, $R_1 = (1 + r_0)/2$, $R_2 = (1 + R_1)/2$, and

$$M_0 = [(1-\mu)(1-R_2)^3]^{-1}.$$

On $|\rho| \le R_2$, the two polynomial families have modulus at most M_0 and their values at zero have modulus at least one. A holomorphic function with these two properties satisfies

$$|\{s \in [0, r_0] : |f(s)| < z\}| \le C z^{1/m}$$

with m, C depending only on the displayed radii and M_0. To verify this, let $a_1, \dots, a_n$ be the zeros in $|z| < R_1$, counted with multiplicity, and write $f/M_0 = g \prod_{j=1}^n B_{a_j}$ using disk-R_2 Blaschke factors. The maximum principle gives $|g| \le 1$ on that disk. Since $|f(0)| \ge 1$ and $|B_{a_j}(0)| = |a_j|/R_2 < R_1/R_2$, we have

$$n \le \frac{\log M_0}{\log(R_2/R_1)}, \qquad |g(0)| \ge M_0^{-1}.$$

The function g is zero-free on $|z| < R_1$. Harnack's inequality applied to $-\log|g|$ gives $|g(s)| \ge M_0^{-H}$ on $[0, r_0]$, where $H = (R_1 + r_0)/(R_1 - r_0)$. On the same interval, $|B_{a_j}(s)| \ge |s - a_j|/(2R_2)$.

Thus, if $|f(s)| < z$, either $n = 0$ and z exceeds a fixed positive threshold, or s lies within $C'z^{1/n}$ of one of at most n zeros. Taking any integer $m \geq \max\{1, n\}$ uniformly above the displayed zero-count bound, and enlarging C for the remaining values of z, proves the claimed sublevel estimate.

Let $\ell = |I|$ and choose

$$\delta_t = \left(\frac{\ell}{32C6^t}\right)^m.$$

A union bound over the two functions and 3^{t-1} words excludes length at most $\ell/(48\,2^t)$ at time t. The sum is at most $\ell/48$. The complement of the union of the open exclusions is the required compact positive-measure set. □

Choose an integer $N \geq 2$, depending only on μ, large enough that

$$(N+1)\rho_+^N \leq \frac{1}{2}. \tag{145}$$

For $n \geq 1$ put $c_n = (1-\mu^n)/(1-\mu)$ and, for $q \in \{-1, 1\}^n$, define

$$D_{n,q}(\rho) = \sum_{j=0}^{n-1} q_j \rho^j c_{n-j}. \tag{146}$$

For two different words of the same length, $D_{n,q} - D_{n,q'}$ is a nonzero polynomial in ρ: its highest differing coefficient is $(q_j - q'_j)c_{n-j} \neq 0$. Consequently we can select, once and for all,

$$\widehat{\rho}_\mu \in \mathcal{R} \cap (\rho_-, \rho_+) \quad \text{such that} \quad D_{n,q}(\widehat{\rho}_\mu) \neq D_{n,q'}(\widehat{\rho}_\mu) \quad (1 \leq n \leq N,\ q \neq q'). \tag{147}$$

Only finitely many additional roots are excluded. A deterministic selection convention is obtained by deleting sufficiently small open neighborhoods of this finite exceptional set and of both endpoints of I, retaining positive measure, and taking the minimum of the surviving compact set. All choices depend only on μ.

From now on fix this ρ. Define the positive constant

$$\Delta = \min\left(\{1\} \cup \{|D_{n,q}(\rho) - D_{n,q'}(\rho)| : 1 \leq n \leq N,\ q \neq q'\}\right) > 0. \tag{148}$$

The smallness of Δ affects constants depending on momentum, not on the target spectrum.

Lemma 15 (A large set of nonsingular amplitudes)**.** *For each normalized spectrum and each $T \geq 2$, there is a compact set $S_T \subset [1, 2]$ of measure at least $399/400$ on which all reference buffers used before T obey*

$$|b_{i,t}(c)| \geq \beta_t := c_{\mathrm{buf}} \mathrm{e}^{-b_\mu t}, \quad c_{\mathrm{buf}} = \frac{A_0^3 D_\mu}{(800k)^3}, \quad b_\mu = (m+3)\log 6, \quad i \leq k,\ 1 \leq t < T. \tag{149}$$

Independently of c, all reference factors obey

$$|u_{i,t}| \geq A_0 D_\mu \mathrm{e}^{-d_\mu t} > 0. \tag{150}$$

Proof. As a polynomial in c, (143) has cubic leading coefficient $A_0^3 P_{t,\sigma}(\rho)$ of modulus at least $A_0^3\delta_t$. A cubic with leading coefficient $v \neq 0$ satisfies

$$|\{c \in [1, 2] : |p(c)| < \beta\}| \leq 6(\beta/|v|)^{1/3};$$

factor it over its three complex roots to obtain this inequality. Set $\beta_t = A_0^3\delta_t/(800k6^t)^3$. The union of the exclusions over all $k3^{t-1}$ coordinate-word choices has length at most $1/(400\,2^t)$. Summation proves the measure claim. Compactness follows either from the all-word closed inequalities, or recursively by imposing the actual-buffer inequalities: on the previously retained set all earlier signs are locally fixed, so the next buffer is continuous there. We use the latter definition of S_T, which contains the all-word screened set. Finally (150) follows from (144) and $A \geq A_0$. □

B.7.2 POSITIVE-TARGET PHASE CURVES AND THEIR SIGN-HISTORY INTERVALS

We now count the parameter intervals that realize scalar sign histories. All expectations below average the same amplitude shared by the whole trajectory.

Fix one positive root $a \in [\kappa^{-1/2}, 1]$ and put

$$\theta = \frac{a}{A} = \frac{a}{A_0 c}, \qquad \Theta_a = \left[\frac{a}{2A_0}, \frac{a}{A_0}\right]. \tag{151}$$

Writing $z_t = u_t/A$ and $h_t = b_t/A^3$ gives

$$z_1 = 1, \quad h_0 = 0, \qquad h_t = \mu h_{t-1} + (z_t^2 - \theta^2) z_t, \qquad z_{t+1} = z_t - \rho^t \,\mathsf{sign}(h_t). \tag{152}$$

On an interval of θ where the preceding sign history is fixed, z_t is constant and

$$h_t = P_t - \theta^2 Q_t, \tag{153}$$

where P_t, Q_t are constants on that interval. Since $P_t \neq 0$ by Lemma 14, each interval can split at no more than one positive zero of h_t. At time t, remove only parameters that hit a zero buffer before t, and let $\mathcal{P}_t$ be the resulting finite partition into connected intervals of a common nonzero sign history through time $t-1$. Its cardinality n_t is at most 2^{t-1}.

The union of buffer-zero parameters over all finite times is countable. We first prove the scalar tube and phase estimates outside this null set. For each fixed t, the estimates extend to each interval of $\mathcal{P}_t$ by continuity of its fixed-history formulas. We retain these intervals when integrating and counting cuts; removing the entire countable set could destroy their connectedness. Membership in S_T alone only excludes buffer zeros before T.

Lemma 16 (Uniform fine-regime phase coordinates)**.** *There is an integer $t_0 = t_0(\kappa, \mu)$ satisfying*

$$t_0 \le C_\mu (1 + \log \kappa) \tag{154}$$

such that, for each positive root and each history interval at every $t \ge t_0$, the reference factor has a sign $s \in \{-1, 1\}$ that remains fixed along every continuation without buffer zeros. Define

$$e_t = \frac{s z_t - \theta}{\rho^t}, \qquad v_t = \frac{s h_t}{2\theta^2 \rho^t}, \qquad q_t = \mathsf{sign}(v_t). \tag{155}$$

On each interval of $\mathcal{P}_t$, e_t is affine in θ with derivative $-\rho^{-t}$, and v_t is a function of e_t satisfying

$$|e_t| \le K_\mu, \quad |v_t| \le V_\mu, \qquad \frac{1}{2(1-\mu)} \le \frac{dv_t}{de_t} \le \frac{2}{1-\mu}. \tag{156}$$

We may also require $\eta_t/a \le 1/(4\max\{K_\mu, 1\})$ and any further fixed momentum-only upper bound on ρ^t/θ by enlarging the constant in (154).

Proof. The possible parameters obey $\theta \in [\theta_*, \theta^*]$, where $\theta_* = (2A_0\sqrt{\kappa})^{-1}$ and $\theta^* = A_0^{-1}$. The scalar tube gives $||z_t| - \theta| \le K_\mu \rho^t$ after a momentum-only time. Once $K_\mu \rho^t \le \theta_*/4$ and $\rho^t \le \theta_*/4$, the sign of z_t cannot change: its magnitude exceeds the length of the next step. Hence s is fixed on every later branch and $|e_t| \le K_\mu$. The tracking bound, or convolution of the root-tube gradient bound, gives $|h_t| \le C_\mu \theta^2 \rho^t$ in this regime, proving the bound on v_t.

On a fixed history interval, differentiation of (153) and (155) gives the exact identity

$$\frac{dv_t}{de_t} = \frac{sP_t}{\theta^3}. \tag{157}$$

This converges uniformly to $(1-\mu)^{-1}$ after a κ-dependent finite entrance. To check the quantitative dependence, let $t_f = O_\mu(1 + \log\kappa)$ be a time after which the sign is fixed. For $l \ge t_f$, write $sz_l = \theta + \epsilon_l$ with $|\epsilon_l| \le K_\mu \rho^l$. The uniform total path length bounds every z_l by a momentum-only constant Z_μ. Cubic expansion and geometric convolution give

$$\left| sP_t - \frac{\theta^3}{1-\mu} \right| \le C_\mu Z_\mu^3 \mu^{t-t_f} + C_\mu(\theta^2 \rho^t + \theta \rho^{2t} + \rho^{3t}). \tag{158}$$

We used $\mu < \rho^3$; missing stationary terms can be included in the first term. Dividing by $\theta^3 \ge \theta_*^3$ and taking $t \ge C_\mu(1 + \log\kappa)$ makes the relative error at most $1/[2(1-\mu)]$. This proves (156). Every additional restriction $\rho^t/\theta_* \le c_\mu$ costs only another $O_\mu(1 + \log\kappa)$ entrance time. □

The phase dynamics are exactly

$$e_{t+1} = \rho^{-1}(e_t - q_t), \tag{159a}$$

$$v_{t+1} = \frac{\mu}{\rho} v_t + e_{t+1} + \frac{3\rho^{t+1}}{2\theta} e_{t+1}^2 + \frac{\rho^{2(t+1)}}{2\theta^2} e_{t+1}^3. \tag{159b}$$

For a hypothetical future word $q = (q_0, \dots, q_{n-1})$ of length $n \le N$, iteration yields

$$\rho^n v_{t+n} = \mu^n v_t + c_n e_t - D_{n,q}(\rho) + E_{t,n,q}, \qquad |E_{t,n,q}| \le C_{\mu,N} \frac{\rho^t}{\theta}. \tag{160}$$

The error bound is uniform over all words and all starting phase states in (156). Indeed the first line of (159) bounds every hypothetical e_{t+j} by a fixed $K_{\mu,N}$, and the nonlinear terms in the second line can then be summed. We increase t_0 so that the error in (160) is at most $\Delta/16$.

Lemma 17 (At most N future cuts on a short phase interval)**.** *Let*

$$L_* := \frac{3}{1-\mu}, \qquad \delta := \min\left\{1, \frac{\Delta}{16 L_*}\right\} > 0. \tag{161}$$

For every $t \ge t_0$, a subinterval of a single history interval whose e_t-length is at most δ splits into at most $N+1$ intervals of the sign history through time $t+N-1$.

Proof. Potential cuts occur at $v_{t+n} = 0$ for $n = 0, \dots, N-1$. At $n = 0$ there is at most one zero by (153). Fix $1 \le n < N$. Along the given short phase curve, the quantity $\mu^n v_t + c_n e_t$ varies by at most $L_* \delta \le \Delta/16$, by (156). At a cut, (160) places this quantity within $\Delta/16$ of $D_{n,q}(\rho)$. Two different future words cannot both produce a cut on this short interval: their offsets are separated by at least Δ, whereas the total possible range, including both errors, has width at most $3\Delta/16$.

For the single possible word, the buffer after appending it has form $P - \theta^2 Q$, with $P \ne 0$ by (144). It has at most one positive zero, even if only a subset of its hypothetical branch is dynamically consistent. Thus each of the N times contributes at most one cut. Points that already lie on an earlier cut do not increase the count. Removing at most N points leaves at most $N+1$ intervals. □

Lemma 18 (Natural-scale interval count)**.** *Define*

$$Z_t := \frac{\rho^t n_t}{|\Theta_a|}. \tag{162}$$

For every $t \ge t_0$,

$$Z_{t+N} \le \zeta Z_t + \frac{\zeta}{\delta}, \qquad \zeta = (N+1)\rho^N \le \frac{1}{2}. \tag{163}$$

In particular, with a uniform $Z_* = Z_*(\kappa, \mu)$ such that $\log(1 + Z_*) \le C_\mu(1 + \log\kappa)$, for $0 \le j < N$ and $n \ge 0$,

$$Z_{t_0+j+nN} \le \delta^{-1} + 2^{-n} Z_*. \tag{164}$$

Proof. A current history interval of θ-length l has phase length $l\rho^{-t}$. Divide its phase image into at most $1 + l\rho^{-t}/\delta$ intervals of length at most δ, and apply Lemma 17. Summing over all current intervals gives

$$n_{t+N} \le (N+1) n_t + \frac{N+1}{\delta} \rho^{-t} |\Theta_a|.$$

Multiplying by $\rho^{t+N}/|\Theta_a|$ proves (163). The finite-time partition ignores only finitely many points, so its interval lengths sum to $|\Theta_a|$.

For each residue j, iterate the affine contraction. Its fixed-point bound is at most $1/\delta$. The elementary bound $n_t \le 2^{t-1}$ and $|\Theta_a| = a/(2A_0) \ge 1/(2A_0\sqrt{\kappa})$ allow

$$Z_* = 1 + 2A_0\sqrt{\kappa}\, 2^{t_0+N}.$$

Together with (154), this proves the asserted logarithmic bound on Z_*. □

B.7.3 A LINEAR POSITIVE-TARGET LOG BUDGET FOR ONE AMPLITUDE

For the normalized reference, set

$$w_{i,t} = u_{i,t}^2 + a_i^2 + \eta_t^2, \qquad R_{i,t} = \frac{2\eta_t w_{i,t}}{|b_{i,t}|}, \qquad \mathcal{H}_T^+ = \sum_{t=1}^{T-1} \sum_{i=1}^{r} \log(1 + R_{i,t}). \tag{165}$$

Only positive-target coordinates appear in this budget.

Lemma 19 (An averaged logarithmic budget)**.** *There is a constant $H_{\mu,k} < \infty$ such that, for every normalized spectrum, every $T \geq 2$, and normalized Lebesgue measure on S_T,*

$$\mathbb{E}_{c\sim \mathrm{Unif}(S_T)} \mathcal{H}_T^+ \leq \frac{8k}{3} C_* T + H_{\mu,k}(1 + \log\kappa)^2, \qquad C_* := 6 + \log(1 + \delta^{-1}). \tag{166}$$

This conditional averaging will yield a high-probability event for one uniform amplitude.

Proof. For $t < t_0$, the reference factors are bounded and (149) holds pointwise on S_T. Therefore

$$\log(1 + R_{i,t}) \leq C_{\mu,k} + b_\mu t.$$

The entire prefix costs at most $C_{\mu,k} t_0^2 \leq C_{\mu,k}(1 + \log\kappa)^2$.

For $t \geq t_0$, fix a positive root and first average uniformly over $\theta \in \Theta_a$, without restricting to S_T. By (156), the set $|v_t| < z$ has phase length at most $4(1-\mu)z$ on each current history interval. Change variables back to θ and sum over the n_t intervals:

$$\mathbb{P}_\theta(|v_t| < z) \leq \min\{1, 4Z_t z\}. \tag{167}$$

The fine-regime conditions give $|u_t| \leq 5a/4$ and $\eta_t/a \leq 1/4$, hence

$$R_{i,t} = \frac{u_t^2 + a^2 + \eta_t^2}{a^2|v_t|} \leq \frac{4}{|v_t|}.$$

The layer-cake formula and (167) imply

$$\mathbb{E}_\theta \log(1 + R_{i,t}) \leq 6 + \log(1 + Z_t). \tag{168}$$

For example, for $y \geq \log 2$ the tail of $\log(1 + 4/|v_t|)$ is at most $\min\{1, 32 Z_t \mathrm{e}^{-y}\}$. Splitting the integral at $\log 2 + \log(1 + 32Z_t)$ proves the stated loose constant.

The density induced by uniform $c \in [1,2]$ on Θ_a is $a/(A_0\theta^2)$, at most twice the uniform density $1/|\Theta_a|$. Thus the expectation over c is at most twice (168). By (164),

$$\log(1 + Z_{t_0+j+nN}) \leq \log(1 + \delta^{-1}) + \log(1 + Z_* 2^{-n}).$$

The elementary summation bound

$$\sum_{n\geq 0} \log(1 + z2^{-n}) \leq C\big(1 + \log(1+z)\big)^2$$

follows by splitting at $n = \lceil \log_2(1+z) \rceil$. Sum over the N residues and the $r \leq k$ positive roots. This gives, over the full amplitude interval,

$$\mathbb{E}_c \sum_{t_0 \leq t < T} \sum_{i \leq r} \log(1 + R_{i,t}) \leq 2kC_* T + C_{\mu,k}(1 + \log\kappa)^2.$$

Restriction to S_T increases the average of a nonnegative quantity by at most $1/|S_T| \leq 400/399 < 4/3$. Adding the pointwise prefix proves (166). At no point is a later amplitude treated as independent of its earlier sign history. □

Proposition 1 (One amplitude: probability and deterministic selection)**.** *For every normalized spectrum and $T \geq 2$, there is a nonempty compact set $\mathcal{A}_T \subset S_T$ of Lebesgue measure at least $(399/400)^2 > 0.995$ such that*

$$\mathcal{H}_T^+ \leq K_H T + \widehat{H}_{\mu,k}(1 + \log\kappa)^2, \qquad K_H = 1600 k C_*, \tag{169}$$

for every $c \in \mathcal{A}_T$. The set depends on the normalized spectrum and T, but not on the Stiefel matrix or the initialization scale.

Proof. By Lemma 19, Markov's inequality at 400 times the conditional mean excludes at most a $1/400$ fraction of S_T. The coefficient $1600kC_*$ exceeds $400(8k/3)C_*$; take $\widehat{H}_{\mu,k} \geq 400H_{\mu,k}$. Thus the subset of S_T satisfying (169) has measure at least $(399/400)^2$. The screened scalar recursion and its finite-horizon budget are continuous relative to S_T, so this subset is compact. Define it to be $\mathcal{A}_T$. □

For $C_\eta \sim \mathsf{Unif}[1,2]$, the event $\{C_\eta \in \mathcal{A}_T\}$ has probability greater than 0.995. This is an event used only in the proof: the algorithm draws once from the full interval and performs no screening or rejection sampling. Alternatively, $c_T = \min \mathcal{A}_T$ is a deterministic amplitude chosen before $\boldsymbol{O}$. Its selection may depend on the target spectrum and the prescribed horizon. Neither conclusion asserts that one good-amplitude event holds at every horizon.

B.7.4 From the positive budget to full matrix propagators

Fix $T = T_x$ and any $c \in \mathcal{A}_T$. All subsequent calculations are deterministic. Work in the fixed frames with the right-frame sign change specified above, so the reference factor and current buffer are

$$\overline{\boldsymbol{U}}_t = \begin{bmatrix} \mathrm{diag}(u_{i,t}) \\ 0 \end{bmatrix}, \qquad \overline{\boldsymbol{B}}_t = \begin{bmatrix} \mathrm{diag}(b_{i,t}) \\ 0 \end{bmatrix}.$$

Let $\boldsymbol{J}_t$ be the full derivative of the state map

$$\mathcal{T}_t(\boldsymbol{U}, \boldsymbol{H}) = (\boldsymbol{U} - \eta_t \,\mathsf{msign}(\mu\boldsymbol{H} + \mathcal{G}(\boldsymbol{U})),\ \mu\boldsymbol{H} + \mathcal{G}(\boldsymbol{U})), \qquad \mathcal{G}(\boldsymbol{U}) = (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star)\boldsymbol{U},$$

at $(\overline{\boldsymbol{U}}_t, \overline{\boldsymbol{B}}_{t-1})$.

Use the weighted Hilbert norm

$$\|(\boldsymbol{X}, \boldsymbol{Y})\|_{*,t}^2 = \|\boldsymbol{X}\|_{\mathrm{F}}^2 + \sum_{i,j \leq k} \frac{|\boldsymbol{Y}_{ij}|^2}{(w_{i,t} + w_{j,t})^2} + \sum_{\gamma > k,\ i \leq k} \frac{|\boldsymbol{Y}_{\gamma i}|^2}{w_{i,t}^2}. \tag{170}$$

With $\ell_t = \eta_t$, it is equivalent, within a factor $\sqrt{2}$, to the sum norm in Equation (110). We ultimately measure full-state perturbations in

$$\|(\boldsymbol{X}, \boldsymbol{Y})\|_{\mathsf{E}} = (\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)^{1/2}.$$

At an aligned point, differentiation gives

$$\mathrm{D}\mathcal{G}[\boldsymbol{X}]_{ij} = (u_i^2 + u_j^2 - a_i^2)\boldsymbol{X}_{ij} + u_i u_j \boldsymbol{X}_{ji}, \tag{171a}$$

$$\mathrm{D}\mathcal{G}[\boldsymbol{X}]_{\gamma i} = u_i^2 \boldsymbol{X}_{\gamma i}, \tag{171b}$$

$$\mathrm{D}\,\mathsf{msign}(\overline{\boldsymbol{B}})[\boldsymbol{Z}]_{ij} = \frac{\boldsymbol{Z}_{ij} - \mathsf{sign}(b_i)\,\mathsf{sign}(b_j)\boldsymbol{Z}_{ji}}{|b_i| + |b_j|}, \tag{171c}$$

$$\mathrm{D}\,\mathsf{msign}(\overline{\boldsymbol{B}})[\boldsymbol{Z}]_{\gamma i} = \boldsymbol{Z}_{\gamma i}/|b_i|. \tag{171d}$$

In the bottom-row formulas, $k < \gamma \leq d$; in the top-block formulas, $1 \leq i, j \leq k$. Thus the full variational equation is block diagonal: diagonal coordinate blocks, bottom-row blocks, and top off-diagonal coordinate pairs.

Blocks containing a positive target. For $t \geq t_0$, a zero-target weight satisfies

$$w_{j,t} \leq (K_\mu^2 + 1)\eta_t^2 \leq \tfrac{1}{8}a_i^2 \leq \tfrac{1}{8}w_{i,t} \quad \text{whenever } a_i > 0,\ a_j = 0,$$

by increasing K_μ to at least one and using $\eta_t \leq a_i/(4K_\mu)$. Therefore every top pair containing a positive target satisfies

$$\frac{2\eta_t(w_{i,t} + w_{j,t})}{|b_{i,t}| + |b_{j,t}|} \leq 2 \max_{\ell \in \{i,j\}:\, a_\ell > 0} R_{\ell,t}. \tag{172}$$

When both targets are positive, the factor two is unnecessary. Bottom positive-target blocks have the same control directly.

For completeness, the gradient coefficients in (171) give $\mathcal{N}_t(\mathrm{D}\mathcal{G}[\boldsymbol{X}]) \le (3/2)\|\boldsymbol{X}\|_{\mathrm{F}}$, where $\mathcal{N}_t$ denotes the buffer part of (170). Also $w_{i,t}/w_{i,t+1} \le 9/\rho^2 < 10$. Combining these bounds with (172) bounds the one-step weighted norm on every positive-containing block by

$$\|\boldsymbol{J}_t|_B\|_{*,t\to *,t+1} \le 100\left(1+\max_{i\le r} R_{i,t}\right) \qquad (t \ge t_0). \tag{173}$$

Here B denotes any invariant block containing a positive target, and the constant is deliberately loose. Before t_0, the same computation with the maximum over all modes, together with (149), bounds the logarithm of the one-step norm by $C_{\mu,k} + b_\mu t$. Consequently the entire prefix costs at most $C_{\mu,k}(1+\log\kappa)^2$.

Every subinterval product on a positive-containing block is therefore at most

$$C_{\mu,k}\exp\left[C_{\mu,k}(1+\log\kappa)^2 + (\log 100 + K_H)T\right] \tag{174}$$

in the endpoint weighted norms. We used nonnegativity of all summands in $\mathcal{H}_T^+$, so the budget for the full horizon bounds the budget of any subinterval.

Zero-target blocks: no product of inverse buffers. For a bottom-row zero-target block, write its scalar perturbations as (x_t, y_{t-1}). It obeys

$$y_t = \mu y_{t-1} + u_t^2 x_t, \qquad x_{t+1} = x_t - \eta_t y_t/|b_t|, \quad b_t = \mu b_{t-1} + u_t^3. \tag{175}$$

For $W_t = u_t y_{t-1} - b_{t-1}x_t$, the polar terms cancel exactly and

$$W_{t+1} = \mu W_t.$$

For arbitrary starting time $s < t$ this gives

$$\frac{x_t}{u_t} = \frac{x_s}{u_s} - W_s \sum_{j=s}^{t-1} \frac{\mu^{j-s+1}\eta_j}{|b_j| u_j u_{j+1}}, \tag{176}$$

$$y_{t-1} = \frac{\mu^{t-s} W_s + b_{t-1}x_t}{u_t}. \tag{177}$$

The divisions are valid by (149) and (150). The reference factors and buffers are uniformly bounded. Hence, for every $1 \le s \le t \le T$,

$$\|(x_t, y_{t-1})\|_2 \le C_{\mu,k}\exp[(b_\mu + 3d_\mu + 1)T]\|(x_s, y_{s-1})\|_2. \tag{178}$$

Indeed each summand in (176) is at most $C_{\mu,k}\exp[(b_\mu + 2d_\mu)j]$, and the last division in (177) costs at most another $\exp(d_\mu T)$. Thus inverse-buffer costs enter a sum.

All zero-target reference coordinates are identical. A top pair with two zero targets splits orthogonally into symmetric and antisymmetric components. Its symmetric factor perturbation is constant, with buffer update $y_t = \mu y_{t-1} + 3u_t^2 x_t$. Its antisymmetric component is exactly (175). Thus (178) controls these pairs as well. Diagonal blocks, of either target type, have constant factor perturbation because the diagonal polar derivative vanishes; their buffer perturbation is a geometrically discounted forcing with bounded coefficient.

Proposition 2 (Every ambient linear propagator is exponential in the horizon)**.** *Put $a_w = -\log\rho_-$, and, for example, define*

$$K_{\mathrm{coc}} := \log 100 + K_H + 2a_w + b_\mu + 3d_\mu + 2. \tag{179}$$

There are constants $C_{\mu,d,k}, H_{\mu,k} < \infty$ such that, simultaneously for every $1 \le s \le t \le T$,

$$\|\boldsymbol{J}_{t-1}\cdots\boldsymbol{J}_s\|_{\mathsf{E}\to\mathsf{E}} \le P_* := C_{\mu,d,k}\exp\left[H_{\mu,k}(1+\log\kappa)^2 + K_{\mathrm{coc}}T\right], \tag{180}$$

where the empty product is the identity and P_* may be enlarged to be at least one.

Proof. For positive-containing blocks use (174). All weights are bounded above by a numerical constant, so the endpoint ordinary norm is bounded by a constant times the endpoint weighted norm. At the input time s, conversion in the other direction costs at most $C_\mu\eta_s^{-2} \le C_\mu\exp(2a_wT)$, since $\eta_s \ge A_0\rho_-^T$. Zero-target blocks have the ordinary-norm estimate (178), and diagonal blocks have the bounded estimate just described. The decomposition into these blocks is orthogonal for the ordinary full-state Hilbert norm, so the full operator norm is the largest block norm. Taking a larger common prefactor gives (180) with the deliberately loose sum of slopes in (179). No symmetry constraint on the perturbation or moving right-frame quotient is needed. □

B.7.5 THE NONLINEAR TRANSFER AND THE POLYNOMIAL INITIALIZATION EXPONENT

The preceding proposition bounds arbitrary ambient linear perturbations and all subinterval propagators. This is enough to control nonlinear errors without asserting that those errors stay in one of the linearized blocks.

Let the paired matrix error and its scalar norm be

$$\boldsymbol{E}_t = (\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t,\ \boldsymbol{B}_{t-1} - \widetilde{\boldsymbol{B}}_{t-1}), \qquad e_t = \|\boldsymbol{E}_t\|_{\mathsf{E}}, \qquad \beta = c_{\mathrm{buf}} \mathrm{e}^{-b_\mu T}.$$

All quantities remain normalized by $a_{\max}$ or $a_{\max}^3$ as appropriate. In particular, $e_1 \le 4\sqrt{k}\,\alpha$ by (137).

Lemma 20 (Ordinary full-state Taylor bound)**.** *Fix an amplitude in $\mathcal{A}_T$. There are numerical constants $c, C > 0$ such that, for $1 \le t < T$, whenever $e_t \le c\beta$,*

$$\boldsymbol{E}_{t+1} = \boldsymbol{J}_t \boldsymbol{E}_t + \boldsymbol{R}_t, \qquad \|\boldsymbol{R}_t\|_{\mathsf{E}} \le M_T e_t^2, \qquad M_T = C\beta^{-2}. \tag{181}$$

On this neighborhood the actual current buffer and the segment used for Taylor expansion have full column rank.

Proof. The reference factor has operator norm at most 16. On $e_t \le 1$, the actual factor has operator norm at most 17. For $\mathcal{G}(\boldsymbol{U}) = (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star)\boldsymbol{U}$, with $\|\boldsymbol{M}^\star\| = 1$, write

$$\mathcal{G}(\widetilde{\boldsymbol{U}} + \boldsymbol{X}) - \mathcal{G}(\widetilde{\boldsymbol{U}}) = \mathrm{D}\mathcal{G}(\widetilde{\boldsymbol{U}})[\boldsymbol{X}] + \boldsymbol{R}_{\mathcal{G}}, \qquad \|\boldsymbol{R}_{\mathcal{G}}\|_{\mathrm{F}} \le C\|\boldsymbol{X}\|_{\mathrm{F}}^2.$$

The current-buffer perturbation is $\boldsymbol{D} = \mu\boldsymbol{Y} + \mathrm{D}\mathcal{G}[\boldsymbol{X}] + \boldsymbol{R}_{\mathcal{G}}$, of Frobenius norm at most Ce_t. Reducing c gives $\|\boldsymbol{D}\|_{\mathrm{F}} \le \beta/2$, so Weyl's inequality preserves a smallest singular value of at least $\beta/2$ along the segment. Lemmas 10 and 12 give bounds $C\beta^{-1}$ and $C\beta^{-2}\|\boldsymbol{D}\|_{\mathrm{F}}^2$, respectively. In the factor component the residual is $-\eta_t(\mathrm{D}\,\mathsf{msign}(\overline{\boldsymbol{B}}_t)[\boldsymbol{R}_{\mathcal{G}}] + \boldsymbol{R}_{\mathrm{polar}})$; in the buffer component it is $\boldsymbol{R}_{\mathcal{G}}$. Since $\eta_t \le 1$ and $\beta \le 1$, both are bounded by $C\beta^{-2}e_t^2$. □

Variation of constants and Proposition 2 imply, as long as the Taylor neighborhoods are respected,

$$e_t \le P_* e_1 + P_* M_T \sum_{s=1}^{t-1} e_s^2. \tag{182}$$

It suffices to impose

$$e_1 \le c_{\mathrm{boot}} \min\left\{ \frac{x}{P_*},\ \frac{\beta}{P_*},\ \frac{1}{TM_T P_*^2} \right\}. \tag{183}$$

Indeed, assuming $e_s \le 2P_* e_1$ through time $t-1$, the nonlinear sum in (182) is at most $4TM_T P_*^3 e_1^2 \le \frac{1}{2} P_* e_1$ after choosing the numerical constant sufficiently small. This closes $e_t \le 2P_* e_1$ and simultaneously keeps $e_t \le c\beta$. A further fixed reduction makes $e_T \le x/100$.

The arbitrary-subinterval propagator bound costs P_*^2 in the third denominator. Its dependence on accuracy is still polynomial.

Set

$$p = \frac{K_{\mathrm{coc}}}{a_+}, \qquad b = \frac{b_\mu}{a_+}, \qquad \boxed{q_{\mathrm{init}} = q = 2p + 2b + 2 = 2 + \frac{2(K_{\mathrm{coc}} + b_\mu)}{-\log \rho_+}}. \tag{184}$$

The horizon bound gives

$$P_* \le C_{\mu,d,k} \exp[H_{\mu,k}(1 + \log\kappa)^2] x^{-p}, \tag{185}$$

$$\beta \ge c_{\mu,k} x^b, \tag{186}$$

$$M_T \le C_{\mu,k} x^{-2b}. \tag{187}$$

The three powers needed in (183) are respectively $p+1$, $p+b$, and $2p+2b$, with one extra power absorbing $T = O_\mu(1 + \log(1/x))$. The choice (184) is sufficient for all $0 < x < 1$ after reducing the prefactor. In (38), choose the squared-log coefficient $C_{\mu,k}$ large enough to absorb twice the coefficient $H_{\mu,k}$ in (180), as required by P_*^2. Thus (38) and (31) imply (183) as well as (138).

At time T, the actual factor has norm at most 17 and the reference factor has norm at most 16. Therefore

$$\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \widetilde{\boldsymbol{U}}_T\widetilde{\boldsymbol{U}}_T^\top\| \le 33\|\boldsymbol{U}_T - \widetilde{\boldsymbol{U}}_T\|_{\mathrm{F}} \le 33x/100.$$

Add the reference error $x/4$ from (141); at time T the total is at most $29x/50$. On the same event this bound persists below x for all later times. Indeed, the partial polar factor has operator norm at most one, so, for every $t \ge T$,

$$\|\boldsymbol{U}_t - \boldsymbol{U}_T\| \le \sum_{s=T}^{\infty} \eta_s = \frac{A\rho^T}{1-\rho} \le 16\rho^T \le \frac{x}{20}. \tag{188}$$

Here Equation (33) gives $\rho^T \le x/320$. Since $\|\boldsymbol{M}^\star\| = 1$ and $x < 1$, the time-T bound yields $\|\boldsymbol{U}_T\| \le \sqrt{1 + 29x/50} < 1.26$. Consequently,

$$\begin{aligned}\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| &\le \frac{29x}{50} + 2\|\boldsymbol{U}_T\|\|\boldsymbol{U}_t - \boldsymbol{U}_T\| + \|\boldsymbol{U}_t - \boldsymbol{U}_T\|^2 \\ &\le \left(\frac{29}{50} + \frac{2(1.26)}{20} + \frac{1}{400}\right)x < x.\end{aligned} \tag{189}$$

This argument uses no buffer separation after T. The conclusion holds on $\mathcal{E}_{\mathrm{init}} \cap \{C_\eta \in \mathcal{A}_T\}$. The two amplitude exclusions and the alignment event give

$$\mathbb{P}\left(\sup_{t\ge T}\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \le x\right) \ge 1 - 0.005 - 0.0025 - 0.0025 = 0.99.$$

Rescale factors, buffers, and reconstruction error by $a_{\max}$, $a_{\max}^3$, and $a_{\max}^2$, respectively. This proves the positive-momentum case of part (b) in Appendix B.1. For the deterministic choice $c_T = \min \mathcal{A}_T$, only the alignment event is random, giving success probability at least 0.995 and proving the deterministic-amplitude refinement in Appendix B.1.

B.8 The condition-number-free undercomplete regime at zero momentum

For the theorem set $\rho = \rho_0 = 3/4$. As in the wide case, the same proof works for every fixed $\rho \in [2/3, 1)$. We now prove the $\mu = 0$ case of part (b) in Appendix B.1. Fix $0 < \varepsilon < \lambda_{\max}^\star$, put $x_\varepsilon = \varepsilon/\lambda_{\max}^\star$, and use

$$A_{\mathrm{u}}^{(0)} = 4(1-\rho)a_{\max}, \qquad T_\varepsilon = 1 + \left\lceil \frac{\log(320/x_\varepsilon)}{-\log\rho} \right\rceil. \tag{190}$$

Draw $\boldsymbol{O}$ uniformly from $\mathcal{O}_{d\times k}$ and independently draw $C_{\eta,t} \overset{\mathrm{iid}}{\sim} \mathsf{Unif}[1,2]$. Set $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ and

$$\eta_0 = C_{\eta,0}a_{\max}, \qquad \eta_t = C_{\eta,t}A_{\mathrm{u}}^{(0)}\rho^{t-1}, \quad t \ge 1. \tag{191}$$

Alignment and the auxiliary path. Put $z = \alpha/a_{\max}$ and initially impose

$$z^2 \le \frac{c_0}{384\kappa\sqrt{dr}}. \tag{192}$$

This implies both the rank condition in Lemma 9 and, on $\mathcal{E}_{\mathrm{init}}$,

$$\left\|\,\mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q})\right\| \le \frac{16\alpha^2\sqrt{dr}}{c_0\lambda_r^\star} \le \frac{1}{24} < \frac{1}{12}. \tag{193}$$

Thus the partial-isometry completion used in Equation (92) is valid under this weaker, quadratic restriction as well. Repeating that construction and retaining the unshortened time-one residual uses neither momentum nor the later step scale. Consequently, on the same event $\mathcal{E}_{\mathrm{init}}$ of probability at least 0.995, it produces the same frames $(\boldsymbol{V}, \boldsymbol{R})$ and an aligned auxiliary factor $\widetilde{\boldsymbol{U}}_1 = -\eta_0\,\mathsf{msign}(\widetilde{\boldsymbol{G}}_0)$. More explicitly, the completion lemma and $\eta_0 \le 2a_{\max}$ give

$$\left\|\,\mathsf{msign}(\boldsymbol{G}_0) - \mathsf{msign}(\widetilde{\boldsymbol{G}}_0)\right\| \le \frac{400\kappa\sqrt{dr}}{c_0}z^2, \tag{194}$$

$$\frac{\|\boldsymbol{U}_1 - \widetilde{\boldsymbol{U}}_1\|}{a_{\max}} \le z + \frac{800\kappa\sqrt{dr}}{c_0} z^2. \tag{195}$$

For this branch set $\widetilde{\boldsymbol{B}}_0 = \boldsymbol{0}$ and, for $t \ge 1$, use

$$\widetilde{\boldsymbol{B}}_t = \big(\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top - \boldsymbol{M}^\star\big)\widetilde{\boldsymbol{U}}_t, \qquad \widetilde{\boldsymbol{U}}_{t+1} = \widetilde{\boldsymbol{U}}_t - \eta_t\,\mathsf{msign}(\widetilde{\boldsymbol{B}}_t). \tag{196}$$

In the aligned frame, every scalar coordinate has $|\bar{u}_{i,1}| = C_{\eta,0}a_{\max} \in [a_{\max}, 2a_{\max}]$ and follows the zero-momentum recursion in Lemma 1. Conditional on $(\boldsymbol{O}, C_{\eta,0})$, these time-one coordinates are fixed and nonzero, while the prefactors $\{C_{\eta,t}\}_{t\ge1}$ remain independent and atomless, so the lemma's null-event statement applies to the time-shifted path.

Reindexing time one as zero, the lower prefactor bound gives

$$S_1 \ge \frac{A_{\mathrm{u}}^{(0)}}{1-\rho} = 4a_{\max}, \qquad D_{a_i}(\bar{u}_{i,1}) \le 2a_{\max}. \tag{197}$$

The null-event conclusion of Lemma 1, followed by a finite union over the k modes, therefore yields almost surely

$$D_{a_i}(\bar{u}_{i,t}) \le 8a_{\max}\rho^{t-1}, \qquad i \le k, \quad t \ge 1. \tag{198}$$

This statement treats positive and zero target modes uniformly. By Equation (49),

$$|\bar{u}_{i,t}^2 - a_i^2| \le 8\lambda_{\max}^\star\rho^{t-1}\big(8\rho^{t-1} + 2\big) \le 80\lambda_{\max}^\star\rho^{t-1}. \tag{199}$$

The proof-local choice in Equation (190) implies $\rho^{T_\varepsilon - 1} \le x_\varepsilon/320$, and hence

$$\big\|\widetilde{\boldsymbol{U}}_{T_\varepsilon}\widetilde{\boldsymbol{U}}_{T_\varepsilon}^\top - \boldsymbol{M}^\star\big\| \le \frac{\varepsilon}{4}. \tag{200}$$

A scale-free inverse-buffer moment. Normalize $a_{\max} = 1$ as in Section B.6.3, and, for $t \ge 1$, put

$$\ell_t = 4(1-\rho)\rho^{t-1}, \qquad w_{i,t} = \bar{u}_{i,t}^2 + a_i^2 + \ell_t^2. \tag{201}$$

The scalar envelope gives $|\bar{u}_{i,t}| \le a_i + 8\rho^{t-1} \le 9$. Define $\mathcal{F}_0 = \sigma(\boldsymbol{O})$ and

$$\mathcal{F}_t = \sigma(\boldsymbol{O}, C_{\eta,0}, \dots, C_{\eta,t-1}), \qquad t \ge 1. \tag{202}$$

In the remainder of this subsection, let all circles and bars refer to the present zero-momentum actual and auxiliary paths. Thus, with $\boldsymbol{V}_{\mathrm{full}} = \mathrm{Comp}_{d,k}(\boldsymbol{V})$,

$$\boldsymbol{U}_t^\circ = \boldsymbol{V}_{\mathrm{full}}^\top(\boldsymbol{U}_t/a_{\max})\boldsymbol{R}, \qquad \boldsymbol{B}_{t-1}^\circ = \boldsymbol{V}_{\mathrm{full}}^\top(\boldsymbol{B}_{t-1}/a_{\max}^3)\boldsymbol{R}, \tag{203}$$

$$\overline{\boldsymbol{U}}_t = \boldsymbol{V}_{\mathrm{full}}^\top(\widetilde{\boldsymbol{U}}_t/a_{\max})\boldsymbol{R}, \qquad \overline{\boldsymbol{H}}_{t-1} = \boldsymbol{V}_{\mathrm{full}}^\top(\widetilde{\boldsymbol{B}}_{t-1}/a_{\max}^3)\boldsymbol{R}. \tag{204}$$

Put $w_{ij,t} = w_{i,t} + w_{j,t}$ and use the fixed-frame norm Equation (110) with these weights. The proofs of the full-column polar differential, the weighted factor–buffer differential, and the ambient Taylor bound use only $\eta_t \le 2\ell_t$, $\ell_{t+1} = \rho\ell_t$, and the displayed weights. They therefore apply verbatim with $\mu = 0$ and the definitions above. It remains to establish a conditional inverse-buffer moment bound that is uniform over the target spectrum.

Lemma 21 (Uniform cubic and simple-root small-ball estimates)**.** *Let $C \sim \mathsf{Unif}[1,2]$ be independent of a sigma-field $\mathcal{F}$, let $s > 0$, and let q, x, a be finite $\mathcal{F}$-measurable random variables. With $g_a(u) = (u^2 - a^2)u$, define*

$$\pi_a(C) = q + g_a(x \pm sC). \tag{205}$$

Then, almost surely and uniformly in the lower-order coefficients,

$$\mathbb{P}\big(|\pi_a(C)| \le zs^3 \mid \mathcal{F}\big) \le \min\{1, 6z^{1/3}\}, \qquad z > 0, \tag{206}$$

$$\mathbb{E}\left[\left(\frac{s^3}{|\pi_a(C)|}\right)^p \middle| \mathcal{F}\right] \le C_p, \qquad 0 < p < \frac{1}{3}. \tag{207}$$

More generally, if an $\mathcal{F}$-measurable C^1 function $\phi : [1,2] \to \mathbb{R}$ satisfies $|\phi'(c)| \ge m > 0$ throughout $[1,2]$, then

$$\mathbb{P}\big(|\phi(C)| \le zm \mid \mathcal{F}\big) \le \min\{1, 2z\}, \tag{208}$$

$$\mathbb{E}\left[\left(\frac{m}{|\phi(C)|}\right)^p \middle| \mathcal{F}\right] \le C_p', \qquad 0 < p < 1. \tag{209}$$

Proof. Condition on $\mathcal{F}$. The polynomial $P(c) = s^{-3}\pi_a(c)$ has degree three and leading coefficient of absolute value one. Over $\mathbb{C}$, write $P(c) = \pm\prod_{j=1}^{3}(c - r_j)$. If $|P(c)| \le z$, at least one factor has modulus at most $z^{1/3}$. The intersection of each such disk with $[1, 2]$ has length at most $2z^{1/3}$, proving Equation (206). In particular, exact zeros have probability zero. For $Z = s^3/|\pi_a(C)|$, the layer-cake identity gives

$$\begin{aligned}\mathbb{E}[Z^p \mid \mathcal{F}] &= p\int_0^\infty t^{p-1}\mathbb{P}(Z > t \mid \mathcal{F})\,\mathrm{d}t \\ &\le 1 + 6p\int_1^\infty t^{p-1-1/3}\,\mathrm{d}t < \infty \end{aligned} \tag{210}$$

when $p < 1/3$. For the last two claims, continuity and the derivative lower bound make ϕ strictly monotone. The inverse image of $[-zm, zm]$ has length at most $2z$, and the same layer-cake calculation converges for $p < 1$. All estimates are pointwise after conditioning, so they also hold conditionally. □

Lemma 22 (Uniform zero-momentum inverse-buffer moment)**.** *For every $0 < p < 1/3$, there is $K_p = K_p(d, k, \rho) < \infty$, independent of $a_i \in [0, 1]$, such that, almost surely on the event of no previous exact auxiliary-buffer zero,*

$$\mathbb{E}\left[\left(\frac{\ell_t w_{i,t}}{|\bar{b}_{i,t}|}\right)^p \middle| \mathcal{F}_{t-1}\right] \le K_p, \qquad i \le k, \quad t \ge 1. \tag{211}$$

Proof. At $t = 1$, $\bar{u}_{i,1} = -C_{\eta,0}$ has affine scale one, $a_i \le 1$, and $\ell_1 \le 4/3$. Thus $\ell_1 w_{i,1}$ is uniformly bounded, whereas $\bar{b}_{i,1} = g_{a_i}(-C_{\eta,0})$ is a cubic with leading coefficient of absolute value one. The cubic part of Lemma 21 proves the claim.

Fix $t \ge 2$, condition on $\mathcal{F}_{t-1}$, and write

$$s = \ell_{t-1}, \qquad K_0 = \frac{2\rho}{1-\rho}, \qquad L_0 = 4K_0. \tag{212}$$

The iterate has the affine representation $\bar{u}_{i,t} = x \pm sC_{\eta,t-1}$. Moreover, Equation (198) gives, uniformly over the fresh prefactor,

$$D_{a_i}(\bar{u}_{i,t}) \le K_0 s. \tag{213}$$

This uniformity follows by applying the pathwise tail lemma to every possible value of the fresh prefactor. More directly, on the conditioning event $\bar{u}_{i,t-1} \ne 0$ for positive targets, the one-step inequality Equation (45) holds for every value of $C_{\eta,t-1}$, including the isolated value that could make $\bar{u}_{i,t} = 0$.

If $a_i \le L_0 s$, then $|\bar{u}_{i,t}| \le (L_0 + K_0)s$ and $\ell_t = \rho s$, so

$$\ell_t w_{i,t} \le \rho\big((L_0 + K_0)^2 + L_0^2 + \rho^2\big)s^3. \tag{214}$$

Since $g_{a_i}(x \pm sC_{\eta,t-1})$ is a cubic with leading scale s^3, the cubic fractional-moment estimate applies.

If $a_i > L_0 s$, then Equation (213) gives, throughout the fresh affine interval,

$$\frac{3}{4}a_i \le |\bar{u}_{i,t}| \le \frac{5}{4}a_i. \tag{215}$$

In particular, this interval cannot contain zero because $K_0 s < a_i/4$. Consequently,

$$|g'_{a_i}(\bar{u}_{i,t})| = |3\bar{u}_{i,t}^2 - a_i^2| \ge \frac{11}{16}a_i^2, \qquad w_{i,t} \le 3a_i^2, \qquad \ell_t w_{i,t} \le 3a_i^2 s. \tag{216}$$

As a function of $C_{\eta,t-1}$, the buffer derivative therefore has magnitude at least $(11/16)a_i^2 s$. The simple-root estimate in Lemma 21 proves the claim in this case for every $p < 1$, and hence for $p < 1/3$. The two bounds have constants depending only on ρ, not on the location of the coarse–fine transition or on a_i. This proves the lemma. □

At time one,

$$|\bar{b}_{i,1}| = C_{\eta,0}(C_{\eta,0} - a_i)(C_{\eta,0} + a_i) \ge C_{\eta,0} - 1. \tag{217}$$

Thus the buffer is zero only on the null endpoint event $C_{\eta,0} = a_i = 1$. At each later time it is a nonzero cubic in the fresh $C_{\eta,t-1}$. Hence every buffer in a prescribed finite horizon is nonzero almost surely, and the preceding lemma applies simultaneously on that horizon.

Condition-number-free stability exponent. For the remainder of the proof set $T = T_\varepsilon$. Set $\theta = 1/12$ and define

$$\overline{R}_t = 2\ell_t \max_{i\le k} \frac{w_{i,t}}{|\bar{b}_{i,t}|}, \qquad L_t^{\mathrm{lin}} = C_{\mathrm{lin}}(1+\overline{R}_t). \tag{218}$$

The preceding lemma yields

$$\mathbb{E}\big[(L_t^{\mathrm{lin}})^\theta \mid \mathcal{F}_{t-1}\big] \le K_\theta(d,k,\rho), \tag{219}$$

with no dependence on κ. Let

$$\overline{K}_\theta = \max\{K_\theta, 1\}, \qquad q_{\mathrm{amp}} = \frac{\log \overline{K}_\theta}{\theta(-\log\rho)}. \tag{220}$$

Almost surely on $\mathcal{E}_{\mathrm{init}}$, repeated conditioning gives

$$\mathbb{E}[P_T^\theta \mid \mathcal{F}_0] \le K_\theta^{T-1}, \qquad P_T = \prod_{t=1}^{T-1} L_t^{\mathrm{lin}}. \tag{221}$$

Markov's inequality therefore gives an event $\Omega_{\mathrm{amp}}^{(0)}$ such that, almost surely on $\mathcal{E}_{\mathrm{init}}$,

$$\mathbb{P}\big((\Omega_{\mathrm{amp}}^{(0)})^c \mid \mathcal{F}_0\big) \le \delta_{\mathrm{amp}}, \qquad \delta_{\mathrm{amp}} = 10^{-3}, \tag{222}$$

and on this event

$$P_T \le \delta_{\mathrm{amp}}^{-1/\theta} \overline{K}_\theta^{1/\theta} 320^{q_{\mathrm{amp}}} x_\varepsilon^{-q_{\mathrm{amp}}} =: C_{\mathrm{amp}} x_\varepsilon^{-q_{\mathrm{amp}}}. \tag{223}$$

For the quantitative buffer radius, again split $\delta_{\mathrm{buf}} = 10^{-3}$ into $\delta_{\mathrm{head}} = \delta_{\mathrm{fut}} = 5\cdot 10^{-4}$. Define

$$\Omega_{\mathrm{head}}^{(0)} = \{C_{\eta,0} \ge 1 + \delta_{\mathrm{head}}\}, \tag{224}$$

$$\Omega_{\mathrm{fut}}^{(0)} = \bigcap_{\substack{1\le i\le k\\ 2\le t<T}} \left\{|\bar{b}_{i,t}| \ge \beta_t\right\}, \tag{225}$$

$$\Omega_{\mathrm{buf}}^{(0)} = \Omega_{\mathrm{head}}^{(0)} \cap \Omega_{\mathrm{fut}}^{(0)}, \qquad \beta_1 = \delta_{\mathrm{head}}, \tag{226}$$

$$\beta_t = \frac{1}{512}\left(\frac{\delta_{\mathrm{fut}}}{kT}\right)^3 \ell_t^3, \qquad 2 \le t < T. \tag{227}$$

Uniformity of $C_{\eta,0}$ and Equation (217) give, respectively,

$$\mathbb{P}\big((\Omega_{\mathrm{head}}^{(0)})^c \mid \mathcal{F}_0\big) = \delta_{\mathrm{head}}, \qquad \min_i |\bar{b}_{i,1}| \ge \beta_1 \quad \text{on } \Omega_{\mathrm{head}}^{(0)}. \tag{228}$$

For $t \ge 2$, the uniform cubic sublevel estimate applies irrespective of the coarse–fine split. Its cubic scale is $s = \ell_{t-1} = \ell_t/\rho$, and the corresponding bound is $6\rho(1/512)^{1/3}\delta_{\mathrm{fut}}/(kT)$, which is at most $\delta_{\mathrm{fut}}/(kT)$. The union bound therefore gives, almost surely on $\mathcal{E}_{\mathrm{init}}$,

$$\mathbb{P}\big((\Omega_{\mathrm{buf}}^{(0)})^c \mid \mathcal{F}_0\big) \le \delta_{\mathrm{head}} + \delta_{\mathrm{fut}} = \delta_{\mathrm{buf}}. \tag{229}$$

All constants here are independent of κ.

Since $n = \lceil \log(320/x_\varepsilon)/(-\log\rho)\rceil$ satisfies $\rho^n \ge \rho x_\varepsilon/320$, one has

$$\ell_\star = \min_{1\le t\le T} \ell_t \ge c_\rho x_\varepsilon, \qquad \beta_\star = \min_{1\le t<T} \beta_t \ge c_{k,\rho} T^{-3} x_\varepsilon^3. \tag{230}$$

On $\Omega_{\mathrm{amp}}^{(0)} \cap \Omega_{\mathrm{buf}}^{(0)}$, define, for this subsection,

$$e_t = \|(\boldsymbol{U}_t^\circ - \overline{\boldsymbol{U}}_t, \boldsymbol{B}_{t-1}^\circ - \overline{\boldsymbol{H}}_{t-1})\|_{\star,t}, \tag{231}$$

$$P_t = \prod_{s=1}^{t-1} L_s^{\mathrm{lin}}, \qquad M_T = C_{\mathrm{Tay}}(\ell_\star^{-2} + \beta_\star^{-3}). \tag{232}$$

Apply Lemma 13 on $\Omega_{\mathrm{buf}}^{(0)}$. The preceding uniform factor bound ensures that its constants depend only on (d, k, ρ). For $1 \le t < T$, while $e_t \le c_{\mathrm{Tay}}\beta_\star \le c_{\mathrm{Tay}}\beta_t$,

$$M_T \le C_{d,k,\rho} T^9 x_\varepsilon^{-9}, \qquad e_{t+1} \le L_t^{\mathrm{lin}} e_t + M_T e_t^2. \tag{233}$$

Since $P_1 = 1$ and $P_{t+1} = P_t L_t^{\mathrm{lin}}$, a normalized quadratic bootstrap shows that if

$$e_1 \le c_{\mathrm{boot}} \min\left\{ \frac{x_\varepsilon}{P_T}, \frac{\beta_\star}{P_T}, \frac{1}{T M_T P_T} \right\}, \tag{234}$$

then, simultaneously for every $1 \le t \le T$,

$$e_t \le 2P_t e_1 \le c_{\mathrm{Tay}}\beta_\star, \qquad e_T \le c x_\varepsilon. \tag{235}$$

Indeed, for $z_t = e_t/P_t$, the identity $P_{t+1} = P_t L_t^{\mathrm{lin}}$ gives

$$z_{t+1} \le z_t + \frac{M_T P_t}{L_t^{\mathrm{lin}}} z_t^2 \le z_t + M_T P_t z_t^2, \tag{236}$$

where $L_t^{\mathrm{lin}} \ge 1$. Hence $P_t \le P_T$. If $z_s \le 2e_1$ for $s \le t$, then

$$z_{t+1} \le e_1 + 4t M_T P_T e_1^2 \le \frac{3}{2} e_1 < 2e_1 \tag{237}$$

provided $e_1 \le (8T M_T P_T)^{-1}$. This closes the induction. The other two terms in Equation (234), after reducing c_{boot}, give respectively $e_T \le c x_\varepsilon$ and $e_t \le c_{\mathrm{Tay}}\beta_\star$ for $1 \le t \le T$.

The three right-hand sides in Equation (234) satisfy

$$\frac{x_\varepsilon}{P_T} \ge c_{d,k,\rho} x_\varepsilon^{q_{\mathrm{amp}}+1}, \tag{238}$$

$$\frac{\beta_\star}{P_T} \ge c_{d,k,\rho} T^{-3} x_\varepsilon^{q_{\mathrm{amp}}+3}, \tag{239}$$

$$\frac{1}{T M_T P_T} \ge c_{d,k,\rho} T^{-10} x_\varepsilon^{q_{\mathrm{amp}}+9}. \tag{240}$$

Moreover, for every fixed $m \ge 0$ and $\zeta > 0$,

$$(1 + \log(1/x))^m \le C_{m,\zeta} x^{-\zeta}, \qquad 0 < x < 1. \tag{241}$$

Using this inequality, the deliberately loose choice

$$q_{\mathrm{init}} = q_{\mathrm{amp}} + 12 \tag{242}$$

shows that there is a constant $c_{\mathrm{stab}} = c_{\mathrm{stab}}(d, k, \rho) > 0$ such that $e_1 \le c_{\mathrm{stab}} x_\varepsilon^{q_{\mathrm{init}}}$ implies all three bootstrap restrictions in Equation (234). In particular, for the canonical numerical choice $\rho = \rho_0$, q_{init} depends only on (d, k).

For completeness, the prefactor c_{init} is not condition-number free. With $z = \alpha/a_{\max}$ as above, use the time-one estimate Equation (195). In addition,

$$\left\| \frac{\boldsymbol{B}_0}{a_{\max}^3} \right\|_{\mathrm{F}} \le \sqrt{k}\, z(1 + z^2), \tag{243}$$

and the time-one weights obey $w_{i,1} \ge |\bar{u}_{i,1}|^2 \ge 1$. Combining these facts with Equation (195) gives

$$e_1 \le C_k \left(z + \frac{\kappa\sqrt{dr}}{c_0} z^2 \right). \tag{244}$$

After the κ-free bootstrap threshold is fixed, choose $c_{\mathrm{init}} = c_{\mathrm{init}}(d, r, k, \kappa, \rho)$ small enough that $c_{\mathrm{init}} \le 1$, Equation (192) holds for every $z \le c_{\mathrm{init}}$ and

$$C_k \left(c_{\mathrm{init}} + \frac{\kappa\sqrt{dr}}{c_0} c_{\mathrm{init}}^2 \right) \le c_{\mathrm{stab}}. \tag{245}$$

Then Equation (244) and Equation (242) close the bootstrap. This exhibits explicitly why the rate and this sufficient polynomial exponent are condition-number free even though the admissible basin prefactor is not. At the canonical $\rho = \rho_0 = 3/4$, the dependence on ρ is absorbed into the constants in Appendix B.1.

Finally, reduce c_{init} once more so that the fixed-frame factor perturbation contributes at most $\varepsilon/4$ at time T_ε. Combining this with Equation (200) gives reconstruction error at most $\varepsilon/2$. In normalized units, for every $t \geq T = T_\varepsilon$, the stepsize tail and Equation (190) give

$$\|\boldsymbol{U}_t - \boldsymbol{U}_T\| \leq \sum_{s=T}^{\infty} \eta_s \leq 8\rho^{T-1} \leq \frac{x_\varepsilon}{40}. \tag{246}$$

The time-T error bound implies $\|\boldsymbol{U}_T\| \leq \sqrt{1 + x_\varepsilon/2} \leq \sqrt{3/2}$. Therefore,

$$\sup_{t \geq T} \|\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \leq \left(\frac{1}{2} + \frac{\sqrt{3/2}}{20} + \frac{1}{1600}\right) x_\varepsilon < x_\varepsilon. \tag{247}$$

No post-T buffer estimate is needed. Since $\mathcal{E}_{\mathrm{init}} \in \mathcal{F}_0$, the tower property gives

$$\mathbb{P}\big(\mathcal{E}_{\mathrm{init}} \cap \Omega^{(0)}_{\mathrm{amp}} \cap \Omega^{(0)}_{\mathrm{buf}}\big) \geq 1 - 0.005 - 0.001 - 0.001 = 0.993 > 0.99. \tag{248}$$

This proves the $\mu = 0$ case of part (b) in Appendix B.1, including its eventual-error guarantee, and establishes the zero-momentum part of Theorem 1. Since $\rho_0 = 3/4$ is numerical, Equation (190) is bounded by $C_0(1 + \log(\lambda^\star_{\max}/\varepsilon))$ after enlarging the same theorem-level constant C_0 if necessary. For arbitrary fixed $\rho \in [2/3, 1)$, the same horizon is $O((1 + \log(\lambda^\star_{\max}/\varepsilon))/(1-\rho))$.

B.9 Arbitrary admissible decay with independent amplitudes

Fix $0 < \mu < 1$ and any $\rho \in [\rho_\mu, 1)$, and use Equation (39). We prove the undercomplete extension stated in Appendix B.1. Normalize $a_{\max} = 1$, write $x = \varepsilon/\lambda^\star_{\max}$, and put $\ell_t = 4(1-\rho)\rho^{t-1}$ for $t \geq 1$. All constants in this subsection may depend on μ, ρ, d, k, but not on the target spectrum, except for the initialization prefactor identified below.

Aligned reference and horizon. The alignment argument in Appendix B.8 uses only the initial gradient and $\eta_0 \leq 2a_{\max}$, so it also applies at positive momentum. Under Equation (192), on $\mathcal{E}_{\mathrm{init}}$ of probability at least 0.995, use Equation (94) with $\widetilde{\boldsymbol{B}}_0 = \boldsymbol{0}$. The actual initial buffer remains $\boldsymbol{B}_0$. Every reference coordinate starts at $\bar{u}_{i,1} = -C_{\eta,0}$ and obeys

$$\bar{b}_{i,t} = \mu \bar{b}_{i,t-1} + g_{a_i}(\bar{u}_{i,t}), \qquad \bar{u}_{i,t+1} = \bar{u}_{i,t} - C_{\eta,t}\ell_t \,\mathsf{sign}(\bar{b}_{i,t}), \qquad \bar{b}_{i,0} = 0.$$

At time one the buffer is a nonzero cubic in $C_{\eta,0}$; at every later time it is a nonzero cubic in the fresh $C_{\eta,t-1}$, conditional on a nonzero previous buffer. Thus, almost surely, all reference buffers are nonzero at every finite time. This statement follows by induction and a countable union, without any restriction on polynomials in ρ.

The initial root distance is at most 2, and $\ell_1/(1-\rho) = 4$. By odd symmetry and the time-shifted versions of Lemmas 6 and 7,

$$D_{a_i}(\bar{u}_{i,t}) \leq K_{\mu,\rho}\ell_t, \qquad t \geq 1 + N_{\mu,\rho}.$$

The total path length gives $|\bar{u}_{i,t}| \leq 10$ at all times. Moreover, $K_{\mu,\rho} \leq 2P_\mu/\rho^2$ and $1 - \rho \leq (8P_\mu + 4)^{-1}$ imply $4K_{\mu,\rho}(1-\rho) < 1/\rho^2 < 2$. Consequently, after capture the reconstruction error of the reference is at most $8\rho^{t-1} \leq 80\rho^{t-1}$. Take

$$T = 1 + N_{\mu,\rho} + \left\lceil \frac{\log(320/x)}{-\log\rho} \right\rceil. \tag{249}$$

Then the reference error at T is at most $x/4$ and

$$T \leq 3 + \frac{\log(640/x)}{-\log\rho} \leq 3 + \frac{\log(640/x)}{1-\rho}.$$

Conditional inverse-buffer moment. Use the filtration in Equation (202) and the weights $w_{i,t} = \bar{u}_{i,t}^2 + a_i^2 + \ell_t^2$. For every $0 < p < 1/3$, we claim that, almost surely,

$$\mathbb{E}\left[\left(\frac{\ell_t w_{i,t}}{|\bar{b}_{i,t}|}\right)^p \middle| \mathcal{F}_{t-1}\right] \le K_p(d,k,\mu,\rho) < \infty, \qquad i \le k,\ t \ge 1. \tag{250}$$

At $t = 1$, the proof of Lemma 22 applies unchanged because $\bar{b}_{i,0} = 0$. For $t \ge 2$, condition on $\mathcal{F}_{t-1}$ and put $s = \ell_{t-1}$. The fresh prefactor enters as

$$\bar{u}_{i,t} = u \pm sC_{\eta,t-1}, \qquad \bar{b}_{i,t} = \mu\bar{b}_{i,t-1} + g_{a_i}(u \pm sC_{\eta,t-1}).$$

For $2 \le t \le 1 + N_{\mu,\rho}$, the factor and weight bounds are uniform, while $s \ge 4(1-\rho)\rho^{N_{\mu,\rho}-1} > 0$. Thus $\ell_t w_{i,t} \le C_{\mu,\rho} s^3$, and the cubic estimate in Lemma 21 proves the claim for this finite prefix.

For $t \ge 2 + N_{\mu,\rho}$, the previous state lies in the reference tube. For every candidate value of the fresh prefactor, the triangle inequality therefore gives $D_{a_i}(\bar{u}_{i,t}) \le K_0 s$, where $K_0 = K_{\mu,\rho} + 2$. If $a_i \le 4K_0 s$, then $|\bar{u}_{i,t}| \le 5K_0 s$ and $\ell_t w_{i,t} \le C_{\mu,\rho} s^3$, so the cubic estimate applies again. If $a_i > 4K_0 s$, the entire fresh affine interval satisfies

$$\frac{3}{4}a_i \le |\bar{u}_{i,t}| \le \frac{5}{4}a_i, \qquad |g'_{a_i}(\bar{u}_{i,t})| \ge \frac{11}{16}a_i^2, \qquad w_{i,t} \le 3a_i^2.$$

The buffer derivative with respect to the fresh prefactor has magnitude at least $(11/16)a_i^2 s$, while $\ell_t w_{i,t} \le 3a_i^2 s$. The simple-root estimate proves Equation (250). The term $\mu\bar{b}_{i,t-1}$ is fixed under this conditioning, so it changes neither the cubic leading coefficient nor the derivative estimate.

Stability and an explicit sufficient initialization. Use the fixed-frame states in Equation (204) for the present positive-momentum paths, and measure their error by $e_t = \|(\boldsymbol{U}_t^\circ - \overline{\boldsymbol{U}}_t, \boldsymbol{B}_{t-1}^\circ - \overline{\boldsymbol{H}}_{t-1})\|_{\star,t}$. The weighted differential and Taylor estimates in Lemmas 11 and 13 already apply at this momentum. Set $\theta = 1/12$ and $L_t^{\mathrm{lin}} = C_{\mathrm{lin}}(1 + 2\ell_t \max_i w_{i,t}/|\bar{b}_{i,t}|)$. By Equation (250), choose $\overline{K}_\theta \ge 1$ such that $\mathbb{E}[(L_t^{\mathrm{lin}})^\theta \mid \mathcal{F}_{t-1}] \le \overline{K}_\theta$. Define

$$q_{\mathrm{amp}}^{\mathrm{iid}} = \frac{\log \overline{K}_\theta}{\theta(-\log\rho)}, \qquad q_{\mathrm{iid}} = q_{\mathrm{amp}}^{\mathrm{iid}} + 12. \tag{251}$$

Repeated conditioning and Markov's inequality give an event $\Omega_{\mathrm{amp}}^{\mathrm{iid}}$ with $\mathbb{P}((\Omega_{\mathrm{amp}}^{\mathrm{iid}})^c \mid \mathcal{F}_0) \le 10^{-3}$, on which

$$P_T := \prod_{t=1}^{T-1} L_t^{\mathrm{lin}} \le C_{\mathrm{amp}}^{\mathrm{iid}} x^{-q_{\mathrm{amp}}^{\mathrm{iid}}}, \qquad C_{\mathrm{amp}}^{\mathrm{iid}} = (10^{-3})^{-1/\theta}\overline{K}_\theta^{(N_{\mu,\rho}+1)/\theta} 320^{q_{\mathrm{amp}}^{\mathrm{iid}}}.$$

Define $\Omega_{\mathrm{buf}}^{\mathrm{iid}}$ by Equation (227), using the present reference buffers, horizon, and the same head and future thresholds. Its conditional failure probability is at most 10^{-3}: the proof uses only the time-one cubic and the uniform cubic sublevel bound, both valid with the conditional momentum shift. Put $\ell_\star = \min_{1\le t\le T} \ell_t$ and $\beta_\star = \min_{1\le t<T} \beta_t$. Since $\rho^{T-1} \ge \rho^{N_{\mu,\rho}+1} x/320$, these thresholds give

$$\ell_\star \ge c_{\mu,\rho} x, \qquad \beta_\star \ge c_{k,\mu,\rho} T^{-3} x^3.$$

Thus the same Taylor bootstrap Equations (234) and (235) applies, with $M_T \le C_{d,k,\mu,\rho} T^9 x^{-9}$. Its three sufficient thresholds have powers $q_{\mathrm{amp}}^{\mathrm{iid}} + 1$, $q_{\mathrm{amp}}^{\mathrm{iid}} + 3$, and $q_{\mathrm{amp}}^{\mathrm{iid}} + 9$, respectively, and the last two carry factors T^{-3} and T^{-10}. By Equation (241), the choice Equation (251) ensures that $e_1 \le c_{\mathrm{stab}} x^{q_{\mathrm{iid}}}$, for a sufficiently small $c_{\mathrm{stab}} = c_{\mathrm{stab}}(d,k,\mu,\rho) > 0$, implies the bootstrap and makes the terminal reconstruction perturbation at most $x/4$.

To enforce this condition, take $c_{\mathrm{iid}} \le 1$ small enough that

$$c_{\mathrm{iid}}^2 \le \frac{c_0}{384\kappa\sqrt{dr}}, \qquad C_k\left(c_{\mathrm{iid}} + \frac{\kappa\sqrt{dr}}{c_0}c_{\mathrm{iid}}^2\right) \le c_{\mathrm{stab}},$$

and require $0 < \alpha_\varepsilon/a_{\max} \le c_{\mathrm{iid}} x^{q_{\mathrm{iid}}}$. The time-one error estimate Equation (244) holds also for positive momentum, because the actual previous buffer is $\boldsymbol{B}_0$ and the auxiliary previous buffer is

zero. It therefore gives the required bound on e_1. The exponent is independent of κ; the prefactor may depend on κ.

At time T, combining the reference and perturbation errors gives reconstruction error at most $x/2$. For every $t \geq T$, the remaining step length is at most $8\rho^{T-1} \leq x/40$. The calculation in Equation (247) therefore gives error at most x for all $t \geq T$, without any later buffer conditions. The alignment event and the two conditional failure budgets have joint success probability at least $1 - 0.005 - 0.001 - 0.001 = 0.993 > 0.99$. Rescaling proves the arbitrary-decay undercomplete claim and completes Theorem 1. This argument uses independent amplitudes; it does not remove the ratio-screening conditions of the separate single-amplitude refinement.

C Analysis of the infinite-prompt softmax ICL population objective (proof of Theorem 3)

C.1 Detailed conditions for Theorem 3

The scalar argument below allows any bounded amplitude sequence, and thus includes both the independent random amplitudes in the main theorem and the deterministic schedule as special cases.

Fix $b \geq 1$, $0 \leq \mu < 1$, and $(b+\mu)/(b+1) \leq \rho < 1$. Initialize $\boldsymbol{B}_{-1} = \boldsymbol{Q}_0 = \boldsymbol{0}$. Let $C_\eta \geq 1$ be a fixed numerical constant, independent of $\boldsymbol{S}$, choose any sequence $c_t \in [1, b]$, and set

$$\eta_t = \frac{C_\eta}{b}(\rho - \mu) c_t q^\star_{\max} \rho^t, \qquad t \geq 0. \tag{252}$$

Then, for every such amplitude sequence and every $T \geq 0$, `Muon` obeys

$$\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\| = \|\boldsymbol{Q}_T - \boldsymbol{S}^{-1}\| \leq C_\eta q^\star_{\max} \rho^T = C_\eta \rho^T \|\boldsymbol{S}^{-1}\|. \tag{253}$$

The main theorem takes $b = 2$, $C_\eta = 1$, and $c_t = C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1, 2]$. The deterministic guarantee follows by taking $b = 1$ and $c_t \equiv 1$: every $\rho \in [(1+\mu)/2, 1)$ admits $\eta_t = C_\eta(\rho - \mu) q^\star_{\max} \rho^t$ with the same envelope.

C.2 Scalar decomposition and tracking

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

$$\nabla f(\boldsymbol{Q}) = \boldsymbol{S}(\boldsymbol{S}\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{S} = \boldsymbol{S}^2\boldsymbol{Q}\boldsymbol{S} - \boldsymbol{S}^2. \tag{254}$$

To proceed, let us denote the eigen-decomposition of $\boldsymbol{S}$ as $\boldsymbol{S} = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}$, where $\boldsymbol{\Lambda}^\star = \mathsf{diag}\{\lambda_1^\star, \ldots, \lambda_d^\star\}$ is a diagonal matrix containing the eigenvalues $\{\lambda_i^\star\}$ of $\boldsymbol{S}$, and $\boldsymbol{V}^\star$ consists of orthonormal eigenvectors of $\boldsymbol{S}$.

Lemma 23 (Exact momentum decomposition for ICL)**.** *Initialize $\boldsymbol{Q}_0 = \boldsymbol{B}_{-1} = \boldsymbol{0}$ in Equation (22). For each $t \geq 0$, the resulting `Muon` iterates admit decompositions*

$$\boldsymbol{Q}_t = \boldsymbol{V}^\star \boldsymbol{\Theta}_t \boldsymbol{V}^{\star\top}, \tag{255}$$

$$\boldsymbol{B}_{t-1} = \boldsymbol{V}^\star \boldsymbol{P}_{t-1} \boldsymbol{V}^{\star\top}, \tag{256}$$

where $\boldsymbol{\Theta}_t = \mathsf{diag}\{\theta_{1,t}, \ldots, \theta_{d,t}\}$ and $\boldsymbol{P}_{t-1} = \mathsf{diag}\{p_{1,t-1}, \ldots, p_{d,t-1}\}$, with $\boldsymbol{P}_{-1} = \boldsymbol{0}$. These diagonal coordinates evolve as

$$\boldsymbol{P}_t = \mu \boldsymbol{P}_{t-1} + \boldsymbol{\Lambda}^{\star 3} \boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2}, \tag{257}$$

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \eta_t \, \mathsf{diag\text{-}sign}(\boldsymbol{P}_t). \tag{258}$$

Proof. The assertion holds at $t = 0$ because $\boldsymbol{Q}_0 = \boldsymbol{B}_{-1} = \boldsymbol{0}$. If it holds at time t, then

$$\nabla f(\boldsymbol{Q}_t) = \boldsymbol{S}^2 \boldsymbol{Q}_t \boldsymbol{S} - \boldsymbol{S}^2 \tag{259}$$

$$= \boldsymbol{V}^\star (\boldsymbol{\Lambda}^{\star 3} \boldsymbol{\Theta}_t - \boldsymbol{\Lambda}^{\star 2}) \boldsymbol{V}^{\star\top}. \tag{260}$$

Hence the buffer update gives $\boldsymbol{B}_t = \boldsymbol{V}^\star \boldsymbol{P}_t \boldsymbol{V}^{\star\top}$ with Equation (257). The partial-polar convention gives $\mathsf{msign}(\boldsymbol{B}_t) = \boldsymbol{V}^\star \, \mathsf{diag\text{-}sign}(\boldsymbol{P}_t) \boldsymbol{V}^{\star\top}$, including zero diagonal entries. Substitution into the iterate update yields Equation (258) and closes the induction. □

Consequently, every spectral coordinate follows

$$p_{i,t} = \mu p_{i,t-1} + \lambda_i^{\star 3}\left(\theta_{i,t} - \frac{1}{\lambda_i^\star}\right), \tag{261}$$

$$\theta_{i,t+1} = \theta_{i,t} - \eta_t \,\mathsf{sign}(p_{i,t}), \qquad \theta_{i,0} = 0, \quad p_{i,-1} = 0. \tag{262}$$

Lemma 24 (Global scalar tracking with momentum)**.** *Let $\lambda \geq \lambda_{\min} > 0$, $0 \leq \mu < 1$, and $b \geq 1$ with $(b+\mu)/(b+1) \leq \rho < 1$. For any sequence $c_t \in [1, b]$, consider*

$$p_t = \mu p_{t-1} + \lambda^3\left(\theta_t - \frac{1}{\lambda}\right), \tag{263}$$

$$\theta_{t+1} = \theta_t - c_t A \rho^t \,\mathsf{sign}(p_t), \qquad p_{-1} = 0, \tag{264}$$

with $\theta_0 = 0$ and $A = C_\eta(\rho - \mu)/(b\lambda_{\min})$ for $C_\eta \geq 1$. Then

$$\left|\theta_t - \frac{1}{\lambda}\right| \leq \frac{C_\eta}{\lambda_{\min}}\rho^t \qquad \textit{for every } t \geq 0. \tag{265}$$

Proof. Set $x_t = \theta_t - 1/\lambda$, $z_t = p_t/\lambda^3$, and

$$W_t \coloneqq \sum_{j=0}^{t} \mu^j, \qquad y_t \coloneqq \frac{z_t}{W_t}, \qquad D \coloneqq \frac{b\mu}{\rho - \mu}. \tag{266}$$

Unrolling the buffer gives $z_t = \sum_{j=0}^{t} \mu^j x_{t-j}$. Since $W_t > 0$, $\mathsf{sign}(p_t) = \mathsf{sign}(y_t)$. Moreover,

$$|x_{t-j} - x_t| \leq bA\rho^t \sum_{\ell=1}^{j} \rho^{-\ell}, \tag{267}$$

and therefore

$$|y_t - x_t| \leq bA\rho^t \frac{\sum_{j=0}^{t} \mu^j \sum_{\ell=1}^{j} \rho^{-\ell}}{\sum_{j=0}^{t} \mu^j} \leq DA\rho^t. \tag{268}$$

For the last inequality, if J_t has mass $\mathbb{P}(J_t = j) = \mu^j/W_t$ on $\{0, \ldots, t\}$, then for $1 \leq \ell \leq t$ and $0 < \mu < 1$,

$$\mathbb{P}(J_t \geq \ell) = \frac{\sum_{j=\ell}^{t} \mu^j}{\sum_{j=0}^{t} \mu^j} = \mu^\ell \frac{1 - \mu^{t-\ell+1}}{1 - \mu^{t+1}} \leq \mu^\ell. \tag{269}$$

The case $\mu = 0$ is immediate. Here $\rho > \mu$, so the geometric series converges. Thus the truncated law is stochastically dominated by the untruncated geometric law, and the tail-sum formula gives $\sum_{\ell \geq 1} (\mu/\rho)^\ell = \mu/(\rho - \mu)$.

Put $K = b + D = b\rho/(\rho - \mu)$. The condition $(b+1)\rho \geq b + \mu$ is exactly

$$K \leq \frac{\rho}{1 - \rho}. \tag{270}$$

We prove by induction that $|x_t| \leq (K/\rho)A\rho^t$. At $t = 0$,

$$|x_0| = \frac{1}{\lambda} \leq \frac{1}{\lambda_{\min}} \leq \frac{C_\eta}{\lambda_{\min}} = \frac{K}{\rho} A. \tag{271}$$

If $|x_t| \leq DA\rho^t$, the update and Equation (268) give

$$|x_{t+1}| \leq (D + b)A\rho^t = KA\rho^t. \tag{272}$$

If $|x_t| > DA\rho^t$, then $\mathsf{sign}(p_t) = \mathsf{sign}(x_t)$, and hence

$$|x_{t+1}| \leq \max\left\{b, \frac{K}{\rho} - 1\right\} A\rho^t \leq KA\rho^t, \tag{273}$$

where the final inequality uses $K \geq b$ and Equation (270). Since $KA\rho^t = (K/\rho)A\rho^{t+1}$, the induction closes. Finally, $(K/\rho)A = C_\eta/\lambda_{\min}$ by the definition of A, which is Equation (265). □

Applying Lemma 24 to Equations (261) and (262) gives

$$\|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\| = \max_{1\le i\le d}\left|\theta_{i,t} - \frac{1}{\lambda_i^\star}\right| \tag{274}$$

$$\le \frac{C_\eta}{\sigma_{\min}(\boldsymbol{S})}\rho^t, \qquad t \ge 0. \tag{275}$$

This proves the detailed guarantee Equation (253). More explicitly, absolute error $\varepsilon_{\mathrm{abs}}$ is reached once

$$T \ge \max\left\{0, \left\lceil \frac{\log\big(C_\eta/(\sigma_{\min}(\boldsymbol{S})\varepsilon_{\mathrm{abs}})\big)}{-\log\rho} \right\rceil\right\}, \tag{276}$$

whereas relative operator error $\varepsilon_{\mathrm{rel}}$ is reached once

$$T \ge \max\left\{0, \left\lceil \frac{\log\big(C_\eta/\varepsilon_{\mathrm{rel}}\big)}{-\log\rho} \right\rceil\right\}. \tag{277}$$

Setting $b = 2$, $C_\eta = 1$, and $c_t = C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1,2]$ proves Theorem 3 for every $\rho \in [(2+\mu)/3, 1)$, pathwise for all realizations. Its iteration bound follows from $-\log\rho \ge 1-\rho$. The matrix-factorization decay factors also satisfy this threshold, as verified next.

A common momentum-dependent decay factor. For $\mu = 0$, set $\rho_0 = 3/4$. For $0 < \mu < 1$, let ρ_μ be the canonical wide-regime matrix-factorization factor in Equation (51). Run Equation (22) with

$$\eta_t = \frac{1}{2}C_{\eta,t}(\rho_\mu - \mu)\|\boldsymbol{S}^{-1}\|\rho_\mu^t, \qquad C_{\eta,t} \overset{\text{iid}}{\sim} \mathsf{Unif}[1,2]. \tag{278}$$

Then, for every realization and simultaneously for every $T \ge 0$,

$$\frac{\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\|}{\|\boldsymbol{Q}^\star\|} \le \rho_\mu^T. \tag{279}$$

Consequently, every relative operator accuracy $\varepsilon \in (0,1)$ is reached in $O_\mu(1+\log(1/\varepsilon))$ iterations.

For $\mu = 0$, all matrix-factorization choices $\rho \ge 2/3$, including $\rho_0 = 3/4$, satisfy the condition for $b = 2$. For $0 < \mu < 1$, Equation (52) and $r_\mu \ge \mu$ give

$$\rho \ge \rho_\mu \ge \frac{3+r_\mu}{4} \ge \frac{3+\mu}{4} \ge \frac{2+\mu}{3}.$$

Thus Equation (279) follows from Equation (253) with $b = 2$, $\rho = \rho_\mu$, and $C_\eta = 1$, recovering the momentum-only iteration constant. The same envelope also holds for the deterministic schedule $\eta_t = (\rho_\mu - \mu)\|\boldsymbol{S}^{-1}\|\rho_\mu^t$, by the $b = 1$ case. With this shared choice of decay factor, combining the wide-regime ($k \ge d$) matrix-factorization guarantee in Theorem 1(a), under its initialization conditions and on its probability-one event, with the pathwise ICL guarantee gives, simultaneously for every $T \ge 0$, the normalized envelopes

$$\frac{\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\|}{\|\boldsymbol{M}^\star\|} \le C_\mu\rho_\mu^T, \qquad \frac{\|\boldsymbol{Q}_T - \boldsymbol{Q}^\star\|}{\|\boldsymbol{Q}^\star\|} \le \rho_\mu^T. \tag{280}$$

Objective-value conversion. Under the detailed conditions in Appendix C.1, for every $T \ge 0$,

$$f(\boldsymbol{Q}_T) \le \frac{C_\eta^2\,\mathrm{tr}(\boldsymbol{S}^3)}{2\sigma_{\min}(\boldsymbol{S})^2}\rho^{2T}, \tag{281}$$

$$\frac{f(\boldsymbol{Q}_T)}{f(\boldsymbol{Q}_0)} \le C_\eta^2\kappa(\boldsymbol{S})^2\rho^{2T} = C_\eta^2\kappa^{2/3}\rho^{2T}. \tag{282}$$

Thus the objective has a geometric envelope with rate ρ^2, but converting the operator guarantee to relative objective error introduces a spectrum-dependent prefactor and an additive $O(\log\kappa/(1-\rho))$ entrance term. For a decay factor chosen as a function of μ alone, this becomes $O_\mu(\log\kappa)$.

To verify the conversion, the diagonal decomposition in Lemma 23 gives $f(\boldsymbol{Q}_T) = \frac{1}{2}\sum_i \lambda_i(\boldsymbol{S})^3(\theta_{i,T} - 1/\lambda_i(\boldsymbol{S}))^2$ and $f(\boldsymbol{Q}_0) = \mathrm{tr}(\boldsymbol{S})/2$. Apply Equation (253) coordinatewise and use $\mathrm{tr}(\boldsymbol{S}^3)/(\sigma_{\min}(\boldsymbol{S})^2\,\mathrm{tr}(\boldsymbol{S})) \le \kappa(\boldsymbol{S})^2$.

D THE ZERO-MOMENTUM CONNECTION BETWEEN MUON AND SCALEDGD FOR MATRIX FACTORIZATION

A provably efficient preconditioned optimizer for matrix factorization is ScaledGD (Tong et al., 2021a), which also achieves convergence rates independent of the condition number. At $\mu = 0$, there are some inherent connections between Muon and ScaledGD. More concretely, the update rule of the zero-momentum specialization of Muon yields

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \nabla f(\boldsymbol{U}_t)\big(\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t)\big)^{\dagger/2} = \boldsymbol{U}_t - \eta_t \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{\Delta}_t^2 \boldsymbol{U}_t\big)^{\dagger/2}, \tag{283}$$

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

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \beta_t \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{U}_t\big)^{-1} \tag{284}$$

for some learning rate $\beta_t > 0$ when $\boldsymbol{U}_t$ has full column rank. Replacing the inverse by a Moore–Penrose inverse outside that domain is a natural heuristic extension, not part of the cited ScaledGD guarantee. In other words, zero-momentum Muon constructs its metric factor from the gradient, whereas ScaledGD constructs one from the iterate. In the idealized case where $\boldsymbol{\Delta}_t^2 \approx c_t \boldsymbol{U}_t\boldsymbol{U}_t^\top$ for some scalar $c_t > 0$, we have $\boldsymbol{U}_t^\top \boldsymbol{\Delta}_t^2 \boldsymbol{U}_t \approx c_t(\boldsymbol{U}_t^\top \boldsymbol{U}_t)^2$, so (283) becomes

$$\boldsymbol{U}_{t+1} \approx \boldsymbol{U}_t - \eta_t c_t^{-1/2} \nabla f(\boldsymbol{U}_t)\big(\boldsymbol{U}_t^\top \boldsymbol{U}_t\big)^{-1}, \tag{285}$$

which coincides with the ScaledGD update (284) after rescaling the learning rate.

In general, the condition $\boldsymbol{\Delta}_t^2 \approx c_t \boldsymbol{U}_t\boldsymbol{U}_t^\top$ need not hold, but it offers useful intuition in the local regime $\boldsymbol{U}_t\boldsymbol{U}_t^\top \approx \boldsymbol{M}^\star$. Adopting once again the simplifying assumption (11), we derive

$$\boldsymbol{\Delta}_t = \boldsymbol{V}^\star\left(\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star\right)\boldsymbol{V}^{\star\top} \quad \text{and} \quad \boldsymbol{U}_t\boldsymbol{U}_t^\top = \boldsymbol{V}^\star\boldsymbol{\Sigma}_t^2\boldsymbol{V}^{\star\top} \approx \boldsymbol{V}^\star\boldsymbol{\Lambda}^\star\boldsymbol{V}^{\star\top}. \tag{286}$$

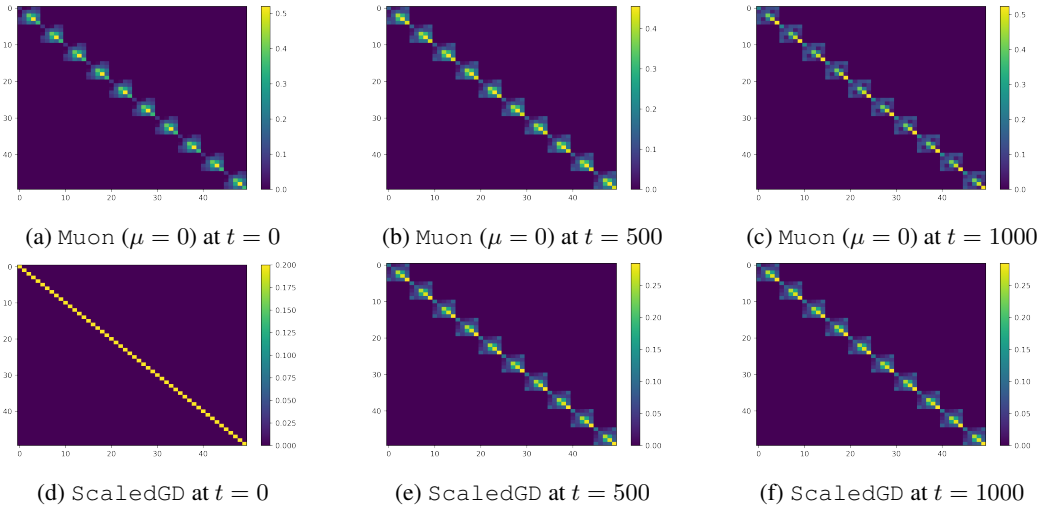
To support the heuristic $\boldsymbol{\Delta}_t^2 \approx c_t \boldsymbol{U}_t\boldsymbol{U}_t^\top$, one needs $(\boldsymbol{\Sigma}_t^2 - \boldsymbol{\Lambda}^\star)^2 \approx c_t\boldsymbol{\Lambda}^\star$, or equivalently,

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

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

$$4\big(|\sigma_{i,t}| - \sqrt{\lambda_i^\star}\big)^2 \approx c_t, \qquad 1 \le i \le r. \tag{287}$$

We next visualize the metric factors associated with zero-momentum Muon and ScaledGD along one synthetic trajectory; Figure 3 displays the matrices, and the setup and protocol caveats are described below.



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

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

Figure 3: Numerical comparison of metric factors associated with zero-momentum Muon and ScaledGD for matrix factorization at several steps along a zero-momentum Muon trajectory. The actual updates use their inverse or Moore–Penrose inverse.

To motivate (287), observe that when $\mu = 0$, at each nonstationary iteration, the scalar sequence $\{\sigma_{i,t}\}$ in (15) moves by signed length η_t irrespective of the gradient magnitude. In a local region where $|\sigma_{i,t}| \approx \sqrt{\lambda_i^\star}$, the scalar sequence is expected to follow a zigzag trajectory around its nearest signed root. Under random initialization, one may thus anticipate $\mathbb{E}\big[\big||\sigma_{i,t}| - \sqrt{\lambda_i^\star}\big|\big] \propto \eta_t$, a scale that is independent of the magnitude of $\lambda_i^\star$. This intuition suggests that in the local region, the `Muon` update may be approximated by `ScaledGD`. Note, however, that these arguments are heuristic in nature; a fully rigorous analysis of their connections is left for future work.

For this visualization, we consider a matrix factorization task with dimension $d = 10$ and target and search ranks $r = k = 5$, initialized at scale $\alpha = 10^{-10}$. The learning-rate schedule follows a plateau rule: multiply the learning rate by 0.3 after 50 iterations without loss improvement. Its initial learning rate and other protocol details remain unspecified, so we use it only as a heuristic illustration. At each step t, the displayed zero-momentum `Muon` metric factor is $\boldsymbol{H}_{\texttt{Muon},t} = \boldsymbol{I} \otimes \big(\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t)\big)^{1/2}$, while the `ScaledGD` factor is $\boldsymbol{H}_{\texttt{ScaledGD},t} = \boldsymbol{I} \otimes \big(\boldsymbol{U}_t^\top \boldsymbol{U}_t\big)$. Their inverse or Moore–Penrose inverse enters the corresponding update. On this synthetic trajectory, both factors display a block-diagonal pattern, illustrating a structural similarity. Their non-diagonal blocks also distinguish these updates from purely diagonal coordinatewise rescaling; this observation alone does not establish a performance claim about `Adam`.

E EXACT ALIGNED MOMENTUM REDUCTION FOR MATRIX FACTORIZATION

An aligned iterate and its momentum buffer remain simultaneously aligned, so every spectral coordinate follows an exact two-state scalar recursion.

Proposition 3 (Exact aligned momentum reduction)**.** *Let* $1 \le s \le \min\{d, k\}$, $\boldsymbol{V} \in \mathcal{O}_{d\times s}$, $\boldsymbol{R} \in \mathcal{O}_{k\times s}$, *and*

$$\boldsymbol{M}^\star = \boldsymbol{V}\mathsf{diag}\{\lambda_1, \dots, \lambda_s\}\boldsymbol{V}^\top, \qquad \lambda_i \ge 0. \tag{288}$$

Suppose

$$\boldsymbol{U}_0 = \boldsymbol{V}\mathsf{diag}\{u_{1,0}, \dots, u_{s,0}\}\boldsymbol{R}^\top, \qquad \boldsymbol{B}_{-1} = \boldsymbol{0}. \tag{289}$$

Then for every $t \ge 0$ *there are diagonal coordinates* $u_{i,t}$ *and* $b_{i,t-1}$ *such that*

$$\boldsymbol{U}_t = \boldsymbol{V}\mathsf{diag}\{u_{1,t}, \dots, u_{s,t}\}\boldsymbol{R}^\top, \tag{290}$$

$$\boldsymbol{B}_{t-1} = \boldsymbol{V}\mathsf{diag}\{b_{1,t-1}, \dots, b_{s,t-1}\}\boldsymbol{R}^\top. \tag{291}$$

Moreover, conditional on the common stepsize sequence, the coordinates decouple and satisfy

$$b_{i,t} = \mu b_{i,t-1} + (u_{i,t}^2 - \lambda_i)u_{i,t}, \tag{292a}$$

$$u_{i,t+1} = u_{i,t} - \eta_t\,\mathsf{sign}(b_{i,t}), \qquad b_{i,-1} = 0. \tag{292b}$$

Proof. Assume Equation (291) holds at time t. Since $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}_s$,

$$\nabla f(\boldsymbol{U}_t) = \boldsymbol{V}\mathsf{diag}\big\{(u_{i,t}^2 - \lambda_i)u_{i,t} : 1 \le i \le s\big\}\boldsymbol{R}^\top. \tag{293}$$

The buffer therefore has the same frames, with coordinates given by Equation (292a). The partial-polar convention in Equation (3) gives

$$\mathsf{msign}(\boldsymbol{B}_t) = \boldsymbol{V}\mathsf{diag}\{\mathsf{sign}(b_{1,t}), \dots, \mathsf{sign}(b_{s,t})\}\boldsymbol{R}^\top, \tag{294}$$

including when some $b_{i,t}$ vanishes. Substitution in Equation (5) yields Equation (292b) and closes the induction. □

F ADDITIONAL AUXILIARY LEMMAS FOR MATRIX FACTORIZATION

This section proves the initial-alignment event used in the undercomplete matrix-factorization analysis. We first compare the leading gradient component with its rank-r approximation, then establish the required Gaussian singular-value bound.

F.1 PROOF OF LEMMA 9

The rank-r truncation can be compared with $\boldsymbol{Q}$ directly, without a subspace-perturbation detour. Because $\boldsymbol{G}_0 = \boldsymbol{Q} + \alpha^3\boldsymbol{O}$ and $\mathrm{rank}(\boldsymbol{Q}) \le r$, the defining best-approximation property of $\boldsymbol{G}_{0,\le r}$ gives

$$\|\boldsymbol{G}_{0,\le r} - \boldsymbol{Q}\| \le \|\boldsymbol{G}_{0,\le r} - \boldsymbol{G}_0\| + \|\boldsymbol{G}_0 - \boldsymbol{Q}\| \tag{295}$$

$$\le 2\|\boldsymbol{G}_0 - \boldsymbol{Q}\| = 2\alpha^3, \tag{296}$$

and

$$\sigma_r(\boldsymbol{G}_{0,\le r}) = \sigma_r(\boldsymbol{G}_0) \ge \sigma_r(\boldsymbol{Q}) - \alpha^3. \tag{297}$$

On the event $\sigma_r(\boldsymbol{Q}) \ge 4\alpha^3$, both $\boldsymbol{Q}$ and $\boldsymbol{G}_{0,\le r}$ have rank r. Applying Lemma 29 therefore yields

$$\big\|\,\mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q})\big\| \le \frac{4\alpha^3}{\sigma_r(\boldsymbol{Q}) - \alpha^3} \le \frac{16\alpha^3}{\sigma_r(\boldsymbol{Q})}. \tag{298}$$

To continue upper bounding (298), we develop a lower bound on $\sigma_r(\boldsymbol{Q})$ in the lemma below, whose proof is provided in Section F.2.

Lemma 25 (Gaussian initial alignment)**.** *Let $1 \le r \le k \le d$, let $\boldsymbol{M}^\star$ be a fixed rank-r positive semidefinite matrix, and draw $\boldsymbol{O} \sim \mathsf{Unif}(\mathcal{O}_{d\times k})$. There exists a universal constant $c_0 \in (0,1]$ such that, with probability at least 0.995,*

$$\sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) \ge \frac{c_0 \lambda_r^\star}{\sqrt{dr}}. \tag{299}$$

Choose c_0 as in Lemma 25; then the event in Equation (90) has probability at least 0.995. On this same event, the definition $\boldsymbol{Q} = -\alpha \boldsymbol{M}^\star \boldsymbol{O}$ gives

$$\sigma_r(\boldsymbol{Q}) = \alpha \sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) \ge \frac{c_0 \alpha \lambda_r^\star}{\sqrt{dr}}. \tag{300}$$

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

$$\left\| \mathsf{msign}(\boldsymbol{G}_{0,\le r}) - \mathsf{msign}(\boldsymbol{Q}) \right\| \le \frac{16\alpha^3}{\sigma_r(\boldsymbol{Q})} \le \frac{16\alpha^2\sqrt{dr}}{c_0 \lambda_r^\star}. \tag{301}$$

F.2 Proof of Lemma 25

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

$$\sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) = \sigma_r(\boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top} \boldsymbol{O}) \ge \lambda_r^\star \cdot \sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{O}) \ge \lambda_r^\star \cdot \sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{O}_{:,1:r}), \tag{302}$$

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

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

$$\sigma_r\big(\boldsymbol{V}^{\star\top} \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1/2}\big) \ge \sigma_r\big(\boldsymbol{V}^{\star\top} \boldsymbol{G}\big) \sigma_r\big((\boldsymbol{G}^\top \boldsymbol{G})^{-1/2}\big) = \frac{\sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{G})}{\sigma_1(\boldsymbol{G})}. \tag{303}$$

The square matrix $\boldsymbol{V}^{\star\top} \boldsymbol{G} \in \mathbb{R}^{r\times r}$ also has i.i.d. standard Gaussian entries. Set

$$\varepsilon_0 := \frac{1}{400}, \qquad t_0 := \sqrt{2\log 400}, \qquad c_0 := \frac{\varepsilon_0}{2+t_0} \le 1.$$

By Lemmas 31 and 32 and a union bound, with probability at least $1 - \varepsilon_0 - e^{-t_0^2/2} = 0.995$,

$$\sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{G}) \ge \frac{\varepsilon_0}{\sqrt{r}}, \qquad \sigma_1(\boldsymbol{G}) \le \sqrt{d} + \sqrt{r} + t_0 \le (2+t_0)\sqrt{d}.$$

Hence, on this event,

$$\frac{\sigma_r(\boldsymbol{V}^{\star\top} \boldsymbol{G})}{\sigma_1(\boldsymbol{G})} \ge \frac{c_0}{\sqrt{dr}}, \qquad \sigma_r(\boldsymbol{M}^\star \boldsymbol{O}) \ge \frac{c_0 \lambda_r^\star}{\sqrt{dr}}. \tag{304}$$

G Coordinatewise history lower bound for matrix factorization

Detailed conditions for Theorem 2

The construction proves the following statement with an arbitrary local initialization radius. The method class is defined just below.

There are universal constants $c_{\mathrm{acc}}, c_{\mathrm{it}}, C > 0$ with the following property. For every $0 < \tau \le 10^{-2}$, $\kappa \ge 5$, and absolute objective tolerance

$$0 < \varepsilon_{\mathrm{abs}} < \frac{c_{\mathrm{acc}} \tau^2}{\kappa}, \tag{305}$$

there exist a dimension

$$d \le C\sqrt{\kappa}\left(1 + \log \frac{c_{\mathrm{acc}} \tau^2}{\kappa \varepsilon_{\mathrm{abs}}}\right), \tag{306}$$

a positive definite $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ of condition number κ, and a full-rank $\boldsymbol{U}_0 \in \mathbb{R}^{d\times d}$ such that $\lambda_{\min}(\boldsymbol{M}^\star) = 1-\tau$, $\left\|\boldsymbol{U}_0\boldsymbol{U}_0^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}} = \tau$, and $\mathsf{dist}_{\mathrm{F}}(\boldsymbol{U}_0, \{\boldsymbol{U} : \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{M}^\star\}) \le C\tau\kappa^{-1/4}$. Every possibly randomized method satisfying Equation (308) pathwise obeys

$$f(\boldsymbol{U}_T) \le \varepsilon_{\mathrm{abs}} \quad \Longrightarrow \quad T \ge c_{\mathrm{it}}\sqrt{\kappa}\log\frac{c_{\mathrm{acc}}\tau^2}{\kappa\varepsilon_{\mathrm{abs}}}. \tag{307}$$

Here $f(\boldsymbol{U}_0) = \tau^2/4$, so the high-accuracy range corresponds to $f(\boldsymbol{U}_T)/f(\boldsymbol{U}_0) \lesssim 1/\kappa$.

Normalization for the main theorem. Fix $\tau = 10^{-2}$ and put $a = \min\{c_{\mathrm{acc}}\tau^2, e^{-1}\}$ and $c = a^2$. For $0 < \varepsilon < c/\kappa$, take $\varepsilon_{\mathrm{abs}} = \varepsilon$. Then

$$\log\frac{c_{\mathrm{acc}}\tau^2}{\kappa\varepsilon} \ge \log\frac{a}{\kappa\varepsilon} \ge \frac{1}{2}\log\frac{1}{\kappa\varepsilon}.$$

Thus Equation (307) implies Theorem 2, with universal implicit constants.

METHOD CLASS

For a generic matrix objective, let $\boldsymbol{X}_t$ be the iterate, $\boldsymbol{H}_t = \nabla f(\boldsymbol{X}_t)$ its exact gradient, and let p index a scalar matrix entry. A method belongs to the *coordinatewise history class* if its realized trajectory satisfies, pathwise, the following condition: for every $t \ge 1$, there is a single scalar map Φ_t, possibly selected adaptively from the common past transcript and common randomness but used identically at every entry, such that

$$(\boldsymbol{X}_t - \boldsymbol{X}_0)_p = \Phi_t((\boldsymbol{H}_0)_p, \dots, (\boldsymbol{H}_{t-1})_p), \qquad \Phi_t(0,\dots,0) = 0. \tag{308}$$

The definition allows arbitrary nonlinear functions of the raw local gradient history and includes the affine sign-span condition $\boldsymbol{X}_t - \boldsymbol{X}_0 \in \mathrm{span}\{\mathsf{sign}(\boldsymbol{H}_0), \dots, \mathsf{sign}(\boldsymbol{H}_{t-1})\}$.

Proposition 4 (Full-batch `Adam` belongs to the coordinatewise history class)**.** *Fix shared scalars $0 \le \beta_1, \beta_2 < 1$, $\epsilon_{\mathrm{A}} > 0$, and a scalar learning-rate sequence $\{\gamma_t\}_{t\ge 0}$ that is common to all entries and may depend on the common past transcript but not on the coordinate. Starting from $\boldsymbol{M}_{-1} = \boldsymbol{V}_{-1} = \boldsymbol{0}$, define the deterministic full-batch recursion*

$$\boldsymbol{M}_t = \beta_1\boldsymbol{M}_{t-1} + (1-\beta_1)\boldsymbol{H}_t, \qquad \widehat{\boldsymbol{M}}_t = \frac{\boldsymbol{M}_t}{1-\beta_1^{t+1}}, \tag{309}$$

$$\boldsymbol{V}_t = \beta_2\boldsymbol{V}_{t-1} + (1-\beta_2)(\boldsymbol{H}_t \odot \boldsymbol{H}_t), \qquad \widehat{\boldsymbol{V}}_t = \frac{\boldsymbol{V}_t}{1-\beta_2^{t+1}}, \tag{310}$$

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \gamma_t \frac{\widehat{\boldsymbol{M}}_t}{\sqrt{\widehat{\boldsymbol{V}}_t} + \epsilon_{\mathrm{A}}}. \tag{311}$$

Then the trajectory satisfies Equation (308). The conclusion excludes nonzero initial moments, coordinate-dependent hyperparameters, stochastic gradients, and coupled or decoupled weight decay, including `AdamW`.

Proof. At every entry p, the bias-corrected moments at time t are the same scalar functions of $(\boldsymbol{H}_0)_p, \dots, (\boldsymbol{H}_t)_p$. Summing Equation (311) from 0 to $t-1$ therefore expresses $(\boldsymbol{X}_t - \boldsymbol{X}_0)_p$ through one common scalar map Φ_t of the local history. An all-zero history makes both moments and every displacement zero, so $\Phi_t(0,\dots,0) = 0$. □

The coordinatewise restriction is essential. At $\boldsymbol{Q}_0 = 0$ in ICL, the full matrix gradient $\nabla f(\boldsymbol{Q}_0) = -\boldsymbol{S}^2$ reveals $\boldsymbol{S} = (-\nabla f(\boldsymbol{Q}_0))^{1/2}$. Likewise, for square matrix factorization with invertible $\boldsymbol{U}_0$, one gradient reveals $\boldsymbol{M}^\star = \boldsymbol{U}_0\boldsymbol{U}_0^\top - \nabla f(\boldsymbol{U}_0)\boldsymbol{U}_0^{-1}$. Thus the lower bounds are structural results for Equation (308), not unrestricted matrix-aware first-order oracle lower bounds.

HARD INSTANCE

We prove the detailed statement above. Set

$$\varepsilon := \frac{\kappa \varepsilon_{\mathrm{abs}}}{c_{\mathrm{acc}} \tau^2} \in (0,1), \tag{312}$$

where the universal constant c_{acc} is chosen at the end of the proof. The construction combines a rank-one Cholesky downdate, which creates a hidden geometric tail, with two exactly initialized scalar anchors that set the condition number to precisely κ.

CONSTRUCTION AND LOCAL INITIALIZATION

Put $\chi = \sqrt{(1-\tau)\kappa}$, $q = (\chi - 1)/(\chi + 1)$, and

$$T_\varepsilon = \frac{\log(1/\varepsilon)}{4\log(1/q)}, \qquad \ell = 1 + \lceil T_\varepsilon \rceil, \qquad d = \ell + 2. \tag{313}$$

Since $\log(1/q) \geq 2/\chi$, this choice satisfies $d \leq C\sqrt{\kappa}(1 + \log(1/\varepsilon))$ for a universal C.

Let $\boldsymbol{J} \in \mathbb{R}^{\ell \times \ell}$ be the lower-shift matrix,

$$\boldsymbol{J}\boldsymbol{e}_i = \boldsymbol{e}_{i+1}, \quad 1 \leq i \leq \ell - 1, \qquad \boldsymbol{J}\boldsymbol{e}_\ell = \boldsymbol{0}, \tag{314}$$

and set

$$\boldsymbol{L} := \boldsymbol{I}_\ell - q\boldsymbol{J}, \qquad \gamma := \frac{\chi + 1}{2} = \frac{1}{1-q}, \qquad \widehat{\boldsymbol{U}}_0 := \gamma \boldsymbol{L}. \tag{315}$$

Define the hard target block by the rank-one downdate

$$\widehat{\boldsymbol{M}}^\star := \widehat{\boldsymbol{U}}_0 \widehat{\boldsymbol{U}}_0^\top - \tau \boldsymbol{e}_1 \boldsymbol{e}_1^\top, \tag{316}$$

and introduce two scalar anchor blocks:

$$\begin{aligned} \boldsymbol{M}^\star &:= \mathsf{diag}\Big\{ \widehat{\boldsymbol{M}}^\star,\ 1-\tau,\ \chi^2 \Big\}, \\ \boldsymbol{U}_0 &:= \mathsf{diag}\Big\{ \widehat{\boldsymbol{U}}_0,\ \sqrt{1-\tau},\ \chi \Big\}. \end{aligned} \tag{317}$$

Because $\|\boldsymbol{J}\| \leq 1$,

$$\sigma_{\min}(\widehat{\boldsymbol{U}}_0) \geq \gamma(1-q) = 1, \qquad \sigma_{\max}(\widehat{\boldsymbol{U}}_0) \leq \gamma(1+q) = \chi. \tag{318}$$

Consequently,

$$(1-\tau)\boldsymbol{I}_\ell \preceq \widehat{\boldsymbol{M}}^\star \preceq \chi^2 \boldsymbol{I}_\ell. \tag{319}$$

The two anchors attain the endpoints of this interval, so

$$\lambda_{\min}(\boldsymbol{M}^\star) = 1 - \tau, \qquad \lambda_{\max}(\boldsymbol{M}^\star) = \chi^2, \qquad \frac{\lambda_{\max}(\boldsymbol{M}^\star)}{\lambda_{\min}(\boldsymbol{M}^\star)} = \frac{\chi^2}{1-\tau} = \kappa. \tag{320}$$

Moreover,

$$\boldsymbol{U}_0 \boldsymbol{U}_0^\top - \boldsymbol{M}^\star = \mathsf{diag}\big\{ \tau \boldsymbol{e}_1 \boldsymbol{e}_1^\top, 0, 0 \big\}, \tag{321}$$

so $\big\| \boldsymbol{U}_0 \boldsymbol{U}_0^\top - \boldsymbol{M}^\star \big\|_{\mathrm{F}} = \tau$ and $f(\boldsymbol{U}_0) = \tau^2/4$.

We next verify the factor-space claim. Let

$$\boldsymbol{a} := \widehat{\boldsymbol{U}}_0^{-1} \boldsymbol{e}_1. \tag{322}$$

Then

$$\widehat{\boldsymbol{M}}^\star = \widehat{\boldsymbol{U}}_0 (\boldsymbol{I}_\ell - \tau \boldsymbol{a}\boldsymbol{a}^\top) \widehat{\boldsymbol{U}}_0^\top. \tag{323}$$

Since

$$\widehat{\boldsymbol{U}}_0^{-1} = \gamma^{-1}(\boldsymbol{I}_\ell + q\boldsymbol{J} + \cdots + q^{\ell-1}\boldsymbol{J}^{\ell-1}), \tag{324}$$

we have

$$\boldsymbol{a} = \gamma^{-1}(1, q, \dots, q^{\ell-1})^\top, \qquad \|\boldsymbol{a}\|_2^2 = \frac{1-q^{2\ell}}{\chi} \le \frac{1}{\chi}. \tag{325}$$

In particular, $\tau \|\boldsymbol{a}\|_2^2 < 1$, so the middle factor in Equation (323) is positive definite. One exact global factor is

$$\begin{aligned} \widehat{\boldsymbol{U}}^\star &:= \widehat{\boldsymbol{U}}_0(\boldsymbol{I}_\ell - \tau \boldsymbol{a}\boldsymbol{a}^\top)^{1/2}, \\ \boldsymbol{U}^\star &:= \mathsf{diag}\Big\{\widehat{\boldsymbol{U}}^\star, \sqrt{1-\tau}, \chi\Big\}. \end{aligned} \tag{326}$$

The exact square-root formula for a rank-one perturbation gives

$$(\boldsymbol{I}_\ell - \tau \boldsymbol{a}\boldsymbol{a}^\top)^{1/2} = \boldsymbol{I}_\ell - \frac{\tau}{1+\sqrt{1-\tau\|\boldsymbol{a}\|_2^2}} \boldsymbol{a}\boldsymbol{a}^\top. \tag{327}$$

Using $\widehat{\boldsymbol{U}}_0 \boldsymbol{a} = \boldsymbol{e}_1$ and Equation (325), we conclude that

$$\begin{aligned} \|\boldsymbol{U}_0 - \boldsymbol{U}^\star\|_{\mathrm{F}} &= \frac{\tau \|\boldsymbol{a}\|_2}{1+\sqrt{1-\tau\|\boldsymbol{a}\|_2^2}} \\ &\le \frac{\tau}{\sqrt{\chi}\big(1+\sqrt{1-\tau/\chi}\big)}, \end{aligned} \tag{328}$$

which is $O(\tau\kappa^{-1/4})$. The initialization is also nonstationary, since

$$\nabla f(\boldsymbol{U}_0) = \mathsf{diag}\big\{\tau\gamma \boldsymbol{e}_1\boldsymbol{e}_1^\top, 0, 0\big\} \neq \boldsymbol{0}. \tag{329}$$

Exact zero-chain propagation

For $0 \le s \le \ell$, define the hard-block support space

$$\mathcal{H}_s := \big\{\boldsymbol{A} \in \mathbb{R}^{\ell\times\ell} : \boldsymbol{A}_{ij} = 0 \text{ whenever } i > s \text{ or } j > s\big\}, \tag{330}$$

with $\mathcal{H}_0 = \{\boldsymbol{0}\}$, and embed it into the full space as

$$\mathcal{Z}_s := \{\mathsf{diag}\{\boldsymbol{A}, 0, 0\} : \boldsymbol{A} \in \mathcal{H}_s\}. \tag{331}$$

Lemma 26 (One-step support propagation)**.** *For every $0 \le s \le \ell - 1$,*

$$\boldsymbol{U} - \boldsymbol{U}_0 \in \mathcal{Z}_s \quad \Longrightarrow \quad \nabla f(\boldsymbol{U}) \in \mathcal{Z}_{s+1}. \tag{332}$$

Consequently, every trajectory satisfying Equation (308) obeys $\boldsymbol{U}_t - \boldsymbol{U}_0 \in \mathcal{Z}_t$ for $0 \le t \le \ell$.

Proof. Write the hard block of $\boldsymbol{U}$ as $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{U}}_0 + \boldsymbol{A}$ with $\boldsymbol{A} \in \mathcal{H}_s$. Its reconstruction residual is

$$\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \widehat{\boldsymbol{M}}^\star = \widehat{\boldsymbol{U}}_0\boldsymbol{A}^\top + \boldsymbol{A}\widehat{\boldsymbol{U}}_0^\top + \boldsymbol{A}\boldsymbol{A}^\top + \tau \boldsymbol{e}_1\boldsymbol{e}_1^\top. \tag{333}$$

Because $\widehat{\boldsymbol{U}}_0$ is lower bidiagonal, the right-hand side is supported on the leading $(s+1)\times(s+1)$ block. Multiplication on the right by $\widehat{\boldsymbol{U}}$ does not enlarge this support: only the first $s+1$ rows of $\widehat{\boldsymbol{U}}$ are sampled, and those rows have zero entries beyond column $s+1$ because $\widehat{\boldsymbol{U}}_0$ is lower bidiagonal and the perturbation lies in $\mathcal{H}_s$. The two scalar anchors remain exactly initialized, and all cross-block entries remain zero. Hence $\nabla f(\boldsymbol{U}) \in \mathcal{Z}_{s+1}$, proving the first claim.

For the second claim, proceed pathwise. The base case is $\boldsymbol{U}_0 - \boldsymbol{U}_0 \in \mathcal{Z}_0$. If the claim holds through time $t-1$, then Lemma 26 places each previous gradient $\boldsymbol{G}_j$ in $\mathcal{Z}_{j+1} \subseteq \mathcal{Z}_t$. Thus, at every coordinate outside $\mathcal{Z}_t$, the complete gradient history through time $t-1$ is identically zero. The condition $\Phi_t(0, \dots, 0) = 0$ in Equation (308) leaves every such coordinate at its initialized value, and hence $\boldsymbol{U}_t - \boldsymbol{U}_0 \in \mathcal{Z}_t$. □

The same support induction also applies to the exact zero-momentum polar-sign update $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\, \mathsf{msign}(\nabla f(\boldsymbol{U}_t))$: if $\boldsymbol{G} = \mathsf{diag}\{\boldsymbol{A}, 0, 0\} \in \mathcal{Z}_s$, then $\mathsf{msign}(\boldsymbol{G}) = \mathsf{diag}\{\mathsf{msign}(\boldsymbol{A}), 0, 0\} \in \mathcal{Z}_s$ under the partial-polar convention. Thus this particular hard construction also obstructs that update, although the theorem itself is stated for the coordinatewise history class.

THE LEADING-BLOCK DOWNDATE BARRIER

Lemma 27 (Downdate barrier)**.** *Fix $1 \le s \le \ell - 1$. If $\widehat{\boldsymbol{U}} - \widehat{\boldsymbol{U}}_0 \in \mathcal{H}_s$, then*

$$\left\| \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \widehat{\boldsymbol{M}}^\star \right\|_{\mathrm{F}} \ge \frac{\tau}{10}(1-q^2)q^{2s}. \tag{334}$$

Proof. Partition after coordinate s:

$$\widehat{\boldsymbol{U}} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}, \qquad \widehat{\boldsymbol{U}}_0 = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}, \tag{335}$$

where

$$\boldsymbol{B} = \gamma(\boldsymbol{I}_s - q\boldsymbol{J}_s), \qquad \boldsymbol{C} = -\gamma q \boldsymbol{e}_1 \boldsymbol{e}_s^\top. \tag{336}$$

Here $\boldsymbol{J}_s$ is the $s \times s$ lower-shift matrix, and the vectors in the expression for $\boldsymbol{C}$ have dimensions $\ell - s$ and s, respectively. Let

$$\boldsymbol{p} \coloneqq \boldsymbol{P}\boldsymbol{e}_s, \qquad \boldsymbol{b} \coloneqq \boldsymbol{B}\boldsymbol{e}_s = \gamma \boldsymbol{e}_s, \tag{337}$$

and abbreviate

$$\xi \coloneqq \left\| \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \widehat{\boldsymbol{M}}^\star \right\|_{\mathrm{F}}. \tag{338}$$

The upper-right residual block equals

$$(\boldsymbol{P} - \boldsymbol{B})\boldsymbol{C}^\top = -\gamma q(\boldsymbol{p} - \boldsymbol{b})\boldsymbol{e}_1^\top. \tag{339}$$

Both off-diagonal blocks contribute to the Frobenius norm, and hence

$$\|\boldsymbol{p} - \boldsymbol{b}\|_2 \le \frac{\xi}{\sqrt{2}\,\gamma q}. \tag{340}$$

Write $\boldsymbol{P} = [\boldsymbol{P}_- \ \boldsymbol{p}]$ and $\boldsymbol{B} = [\boldsymbol{B}_- \ \boldsymbol{b}]$, and define

$$\boldsymbol{K}_s \coloneqq \boldsymbol{B}_-\boldsymbol{B}_-^\top - \tau \boldsymbol{e}_1\boldsymbol{e}_1^\top. \tag{341}$$

The vector

$$\boldsymbol{z}_s \coloneqq (q^{s-1}, q^{s-2}, \dots, q, 1)^\top \tag{342}$$

satisfies $\boldsymbol{B}_-^\top \boldsymbol{z}_s = \boldsymbol{0}$. Therefore,

$$\frac{\boldsymbol{z}_s^\top \boldsymbol{K}_s \boldsymbol{z}_s}{\|\boldsymbol{z}_s\|_2^2} = -\tau \frac{(1-q^2)q^{2s-2}}{1-q^{2s}} \le -\beta_s, \qquad \beta_s \coloneqq \tau(1-q^2)q^{2s-2}. \tag{343}$$

Since $\boldsymbol{P}_-\boldsymbol{P}_-^\top \succeq \boldsymbol{0}$, evaluating along $\boldsymbol{z}_s$ yields

$$\left\| \boldsymbol{P}_-\boldsymbol{P}_-^\top - \boldsymbol{K}_s \right\| \ge \beta_s. \tag{344}$$

If $\boldsymbol{E}_{11}$ denotes the leading residual block, then

$$\boldsymbol{P}_-\boldsymbol{P}_-^\top - \boldsymbol{K}_s = \boldsymbol{E}_{11} - (\boldsymbol{p}\boldsymbol{p}^\top - \boldsymbol{b}\boldsymbol{b}^\top). \tag{345}$$

It follows that

$$\beta_s \le \xi + \left\| \boldsymbol{p}\boldsymbol{p}^\top - \boldsymbol{b}\boldsymbol{b}^\top \right\|_{\mathrm{F}}. \tag{346}$$

If $\xi > 1$, then the claimed result follows immediately because $\tau < 1$. Suppose instead that $\xi \le 1$. Since $\chi > 2$, we have $q > 1/3$ and $\gamma > 3/2$. Thus Equation (340) gives $\|\boldsymbol{p} - \boldsymbol{b}\|_2 \le \sqrt{2}$, and consequently

$$\|\boldsymbol{p}\|_2 + \|\boldsymbol{b}\|_2 \le 2\gamma + \sqrt{2} \le 3\gamma. \tag{347}$$

We therefore obtain

$$\begin{aligned} \left\| \boldsymbol{p}\boldsymbol{p}^\top - \boldsymbol{b}\boldsymbol{b}^\top \right\|_{\mathrm{F}} &\le (\|\boldsymbol{p}\|_2 + \|\boldsymbol{b}\|_2)\,\|\boldsymbol{p} - \boldsymbol{b}\|_2 \\ &\le \frac{3}{\sqrt{2}q}\,\xi < 7\xi. \end{aligned} \tag{348}$$

Combining this with Equation (346) yields $\xi \ge \beta_s/8$. Finally,

$$\frac{\beta_s}{8} = \frac{\tau}{8}(1-q^2)q^{2s-2} \ge \frac{\tau}{10}(1-q^2)q^{2s}, \tag{349}$$

which proves Equation (334). □

ITERATION LOWER BOUND

By Lemma 26, the hard-block perturbation at time t belongs to $\mathcal{H}_t$. For $1 \le t \le \ell - 1$, applying Lemma 27 with $s = t$, and using the initial residual for $t = 0$, gives

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \ge \frac{\tau}{10}(1-q^2)q^{2t}, \qquad f(\boldsymbol{U}_t) \ge \frac{\tau^2}{400}(1-q^2)^2 q^{4t} \tag{350}$$

for $0 \le t \le \ell - 1$. Because $1 - q^2 = 4\chi/(\chi+1)^2 \ge c_{\mathrm{tail}}/\sqrt{\kappa}$ for a universal $c_{\mathrm{tail}} > 0$, choosing $c_{\mathrm{acc}} \le c_{\mathrm{tail}}^2/400$ and using Equation (312) show that every integer $T < T_\varepsilon$ satisfies

$$f(\boldsymbol{U}_T) \ge \frac{c_{\mathrm{tail}}^2\tau^2}{400\kappa} q^{4T} > \varepsilon_{\mathrm{abs}}. \tag{351}$$

Thus $f(\boldsymbol{U}_T) \le \varepsilon_{\mathrm{abs}}$ requires $T \ge T_\varepsilon$. Finally, $\chi > 2$ and

$$\log\frac{1}{q} = \log\left(1 + \frac{2}{\chi - 1}\right) \le \frac{2}{\chi - 1} \le \frac{4}{\chi}. \tag{352}$$

This proves the detailed bound Equation (307); the normalization above then gives Theorem 2.

H INFINITE-PROMPT GAUSSIAN SOFTMAX FORMULATION OF THE ICL OBJECTIVE

This section defines the softmax-attention model used for the ICL case study and proves that its full-batch population risk is exactly the quadratic in Equation (19). We consider noiseless linear regression tasks with Gaussian random-design context tokens and a strict-causal query mask. In the $L \to \infty$ formulation, the empirical context-token measure is replaced by its limiting conditional law. This distribution-level model is the sole attention architecture analyzed in the paper; the finite-prompt setting is discussed separately below.

Proposition 5 (Exact infinite-prompt Gaussian softmax population objective)**.** *Let $\boldsymbol{S} \succ \boldsymbol{0}$ and $\gamma > 0$. Draw $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and, independently, $\boldsymbol{x}, \boldsymbol{x}_{\mathrm{q}} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{S})$. For each realized $\boldsymbol{w}$, let $\nu_{\boldsymbol{w}}$ be the conditional law of the context token*

$$\boldsymbol{z}_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{w}^\top\boldsymbol{x} \end{pmatrix}, \qquad \boldsymbol{z}_{\mathrm{q}}^0 = \begin{pmatrix} \boldsymbol{x}_{\mathrm{q}} \\ 0 \end{pmatrix},$$

and define the fixed value map and the temperature-scaled merged score by

$$\boldsymbol{W}_V = \begin{pmatrix} \boldsymbol{0}_{d\times d} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 1 \end{pmatrix}, \qquad \boldsymbol{U}_\gamma(\boldsymbol{Q}) = \gamma\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_d \\ \boldsymbol{0}_d^\top & 0 \end{pmatrix}.$$

The strict-causal infinite-prompt softmax output is

$$\boldsymbol{T}_\infty^{\mathsf{sm}}(\boldsymbol{Q}; \nu_{\boldsymbol{w}})(\boldsymbol{z}_{\mathrm{q}}^0) := \frac{\displaystyle\int \boldsymbol{W}_V\boldsymbol{z}\, \exp\left(\frac{\boldsymbol{z}^\top\boldsymbol{U}_\gamma(\boldsymbol{Q})\boldsymbol{z}_{\mathrm{q}}^0}{\gamma}\right) \mathrm{d}\nu_{\boldsymbol{w}}(\boldsymbol{z})}{\displaystyle\int \exp\left(\frac{\boldsymbol{z}^\top\boldsymbol{U}_\gamma(\boldsymbol{Q})\boldsymbol{z}_{\mathrm{q}}^0}{\gamma}\right) \mathrm{d}\nu_{\boldsymbol{w}}(\boldsymbol{z})}.$$

Then its label-coordinate prediction and squared population risk satisfy

$$\widehat{y}_{\mathrm{q}}^{\mathsf{sm},\infty} := \boldsymbol{e}_{d+1}^\top\boldsymbol{T}_\infty^{\mathsf{sm}}(\boldsymbol{Q}; \nu_{\boldsymbol{w}})(\boldsymbol{z}_{\mathrm{q}}^0) = \boldsymbol{w}^\top\boldsymbol{S}\boldsymbol{Q}\boldsymbol{x}_{\mathrm{q}},$$
$$f_\infty^{\mathsf{sm}}(\boldsymbol{Q}) := \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{\mathrm{q}}}\left[\left(\widehat{y}_{\mathrm{q}}^{\mathsf{sm},\infty} - \boldsymbol{w}^\top\boldsymbol{x}_{\mathrm{q}}\right)^2\right] = f(\boldsymbol{Q}).$$

Consequently, $\nabla f_\infty^{\mathsf{sm}}(\boldsymbol{Q}) = \boldsymbol{S}^2\boldsymbol{Q}\boldsymbol{S} - \boldsymbol{S}^2$.

Proof. Conditionally on $\boldsymbol{w}$, the context token is a possibly degenerate centered Gaussian with covariance

$$\boldsymbol{\Gamma}_{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{S} & \boldsymbol{S}\boldsymbol{w} \\ \boldsymbol{w}^\top\boldsymbol{S} & \boldsymbol{w}^\top\boldsymbol{S}\boldsymbol{w} \end{pmatrix}.$$

For any centered Gaussian $\boldsymbol{z}$ with covariance $\boldsymbol{\Gamma}$ and any vector $\boldsymbol{a}$, differentiate the moment-generating function $\mathbb{E}[\exp(\boldsymbol{z}^\top \boldsymbol{a})] = \exp(\boldsymbol{a}^\top \boldsymbol{\Gamma} \boldsymbol{a}/2)$ to obtain

$$\frac{\mathbb{E}[\boldsymbol{z} \exp(\boldsymbol{z}^\top \boldsymbol{a})]}{\mathbb{E}[\exp(\boldsymbol{z}^\top \boldsymbol{a})]} = \boldsymbol{\Gamma} \boldsymbol{a};$$

the identity remains valid for degenerate covariance matrices (Castin et al., 2025, Lemma 4.1); see also Boursier & Boyer (2026, Lemma 2.1). Here the effective score vector is

$$\boldsymbol{a} = \frac{\boldsymbol{U}_\gamma(\boldsymbol{Q}) \boldsymbol{z}_{\mathrm{q}}^0}{\gamma} = \begin{pmatrix} \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}} \\ 0 \end{pmatrix}.$$

Therefore,

$$\boldsymbol{W}_V \boldsymbol{\Gamma}_{\boldsymbol{w}} \boldsymbol{a} = \begin{pmatrix} \boldsymbol{0}_d \\ \boldsymbol{w}^\top \boldsymbol{S} \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}} \end{pmatrix},$$

which proves the prediction identity. With $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{Q} - \boldsymbol{I}$, independence and the two covariance identities give

$$f_\infty^{\mathsf{sm}}(\boldsymbol{Q}) = \frac{1}{2} \mathbb{E}\big[(\boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{x}_{\mathrm{q}})^2\big] = \frac{1}{2} \operatorname{tr}(\boldsymbol{A} \boldsymbol{S} \boldsymbol{A}^\top) = f(\boldsymbol{Q}).$$

Differentiating the resulting population risk yields the stated gradient. □

The distribution-level prediction is also the almost-sure limit of the empirical softmax prediction for each fixed $\boldsymbol{Q}$. Conditional on a task $\boldsymbol{w}$ and query $\boldsymbol{x}_{\mathrm{q}}$, both $\exp(\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}})$ and $(\boldsymbol{w}^\top \boldsymbol{x}) \exp(\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}})$ are integrable under the Gaussian context law. The conditional strong law therefore applies to the denominator and numerator separately, and the limiting denominator is strictly positive. Their ratio converges to $\boldsymbol{w}^\top \boldsymbol{S} \boldsymbol{Q} \boldsymbol{x}_{\mathrm{q}}$. The proposition computes the population risk of this limiting predictor; interchanging the prompt-length limit with the population expectation would require an additional argument.

The factor γ in $\boldsymbol{U}_\gamma(\boldsymbol{Q})$ is essential for the convention $\boldsymbol{Q}^\star = \boldsymbol{S}^{-1}$: equivalently, the temperature-normalized merged score is $\boldsymbol{U}_\gamma(\boldsymbol{Q})/\gamma = \operatorname{blkdiag}(\boldsymbol{Q}, 0)$. Without this scaling the prediction is divided by γ, and the minimizer is $\gamma \boldsymbol{S}^{-1}$ rather than $\boldsymbol{S}^{-1}$. One may realize the merged score with, for example, $\boldsymbol{W}_K = \boldsymbol{I}$ and $\boldsymbol{W}_Q = \boldsymbol{U}_\gamma(\boldsymbol{Q})$, but Proposition 5 concerns optimization of the structured merged parameter $\boldsymbol{Q}$ with fixed $\boldsymbol{W}_V$, not joint training of separately factorized key, query, and value matrices. Including the zero-label query as one additional attention atom gives the same infinite-prompt limit because its mass is $1/(L+1)$, although it changes every finite-L objective.

A finite-prompt obstruction. Let f_L^{sm} denote the population risk of the masked version of Proposition 5 with L i.i.d. context tokens, where the expectation is over $\boldsymbol{w}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$, and $\boldsymbol{x}_{\mathrm{q}}$. Write $\bar{\boldsymbol{x}} = L^{-1} \sum_i \boldsymbol{x}_i$ and $\boldsymbol{C}_L = L^{-1} \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top - \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^\top$. At $\boldsymbol{Q} = \boldsymbol{0}$, the prediction is $\boldsymbol{w}^\top \bar{\boldsymbol{x}}$. Differentiating the uniform attention weights and averaging over $\boldsymbol{w}$ therefore gives

$$\nabla f_L^{\mathsf{sm}}(\boldsymbol{0}) = \mathbb{E}\big[\boldsymbol{C}_L (\bar{\boldsymbol{x}} - \boldsymbol{x}_{\mathrm{q}}) \boldsymbol{x}_{\mathrm{q}}^\top\big] = -\mathbb{E}[\boldsymbol{C}_L] \boldsymbol{S} = -\left(1 - \frac{1}{L}\right) \boldsymbol{S}^2.$$

Here the query is independent of the context, has mean zero and covariance $\boldsymbol{S}$, and $\mathbb{E}[\boldsymbol{C}_L] = (1 - 1/L)\boldsymbol{S}$. Thus the difference from the infinite-prompt gradient at initialization is $\boldsymbol{S}^2/L$. In the eigenbasis of $\boldsymbol{S}$, a proposed relative bound $|e_{i,L}| \le \lambda_i(\boldsymbol{S})^3 \delta_L$ must already satisfy $\delta_L \ge 1/(L \lambda_{\min}(\boldsymbol{S}))$. The available finite-prompt concentration and gradient-flow results in Boursier & Boyer (2026) do not provide the trajectory-uniform, covariance-relative control needed to pass this discontinuous partial-polar update through finite prompt length.

I Coordinatewise history lower bound for softmax ICL

Detailed conditions for Theorem 4

There are universal constants $c, C > 0$ with the following property. For every $\kappa \ge 4$, $0 < \varepsilon < 1$, and $r_0 > 0$, there exist a dimension $d \le C\sqrt{\kappa}(1 + \log(1/\varepsilon))$, a positive-definite population

covariance $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ with $\kappa(\boldsymbol{S})^3 = \kappa$, and a positive definite $\boldsymbol{Q}_0$ satisfying $\left\|\boldsymbol{Q}_0 - \boldsymbol{S}^{-1}\right\|_{\mathrm{F}} = r_0$ and $\nabla f(\boldsymbol{Q}_0) \neq \boldsymbol{0}$. Every possibly randomized method satisfying Equation (308) pathwise obeys

$$f(\boldsymbol{Q}_T) \leq \frac{\varepsilon}{8} f(\boldsymbol{Q}_0) \quad \Longrightarrow \quad T \geq c\sqrt{\kappa}\log\frac{1}{\varepsilon}. \tag{353}$$

In particular, this applies to ordinary affine sign-span methods and to the deterministic full-batch `Adam` recursion in Proposition 4.

Since $r_0 > 0$ is arbitrary, this detailed statement immediately implies Theorem 4.

We now prove the detailed statement. The main point is to embed a classical nearest-neighbor zero chain while respecting the structured Hessian $\boldsymbol{H} \mapsto \boldsymbol{S}^2\boldsymbol{H}\boldsymbol{S}$ of the ICL objective.

A CIRCULANT EMBEDDING

Set $q = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ and

$$\begin{aligned} T_\varepsilon &= \frac{\log(1/\varepsilon)}{2\log(1/q)}, & R_\varepsilon &= \left\lceil \frac{\log 2}{2\log(1/q)} \right\rceil, \\ m_\varepsilon &= \lceil T_\varepsilon \rceil, & d &= 4(m_\varepsilon + R_\varepsilon + 1). \end{aligned} \tag{354}$$

Thus $d = O(\sqrt{\kappa}(1+\log(1/\varepsilon)))$. Index coordinates by $\{0,\dots,d-1\}$ modulo d, and let $\boldsymbol{P}$ be the cyclic shift $\boldsymbol{P}\boldsymbol{e}_j = \boldsymbol{e}_{j+1 \bmod d}$. Define

$$\boldsymbol{A} \coloneqq \frac{\kappa+1}{2}\boldsymbol{I} - \frac{\kappa-1}{4}(\boldsymbol{P} + \boldsymbol{P}^\top). \tag{355}$$

Because d is even, the eigenvalues of $\boldsymbol{A}$ are

$$\lambda_j(\boldsymbol{A}) = \frac{\kappa+1}{2} - \frac{\kappa-1}{2}\cos\frac{2\pi j}{d}, \qquad 0 \leq j \leq d-1. \tag{356}$$

Thus

$$\lambda_{\min}(\boldsymbol{A}) = 1, \qquad \lambda_{\max}(\boldsymbol{A}) = \kappa. \tag{357}$$

Set

$$\boldsymbol{S} \coloneqq \boldsymbol{A}^{1/3}, \tag{358}$$

where the principal positive definite cube root is used. Then $\boldsymbol{S}$ is real symmetric and circulant, and $\kappa(\boldsymbol{S})^3 = \kappa(\boldsymbol{A}) = \kappa$. This positive-definite matrix is a valid population covariance for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S})$, so the hard objective is exactly the infinite-prompt softmax population risk in Proposition 5.

For $\boldsymbol{x} \in \mathbb{R}^d$, write

$$\mathrm{Circ}(\boldsymbol{x}) \coloneqq \sum_{j=0}^{d-1} x_j \boldsymbol{P}^j, \tag{359}$$

the circulant matrix with first column $\boldsymbol{x}$. We use the identities

$$\begin{aligned} \|\mathrm{Circ}(\boldsymbol{x})\|_{\mathrm{F}}^2 = d\,\|\boldsymbol{x}\|_2^2, \qquad \langle \mathrm{Circ}(\boldsymbol{x}), \mathrm{Circ}(\boldsymbol{y})\rangle = d\boldsymbol{x}^\top\boldsymbol{y}, \\ \mathrm{Circ}(\boldsymbol{x})\,\mathrm{Circ}(\boldsymbol{y}) = \mathrm{Circ}(\mathrm{Circ}(\boldsymbol{x})\boldsymbol{y}). \end{aligned} \tag{360}$$

Let $\boldsymbol{v} = \boldsymbol{A}^{-1}\boldsymbol{e}_0$ and choose

$$\alpha \coloneqq \frac{r_0}{\|\boldsymbol{A}^{-1}\|_{\mathrm{F}}}, \qquad \boldsymbol{Q}^\star = \boldsymbol{S}^{-1}, \qquad \boldsymbol{Q}_0 = \boldsymbol{S}^{-1} + \alpha\boldsymbol{A}^{-1}. \tag{361}$$

Both summands in $\boldsymbol{Q}_0$ are positive definite, and

$$\|\boldsymbol{Q}_0 - \boldsymbol{Q}^\star\|_{\mathrm{F}} = r_0, \qquad \nabla f(\boldsymbol{Q}_0) = \boldsymbol{S}^2(\boldsymbol{Q}_0 - \boldsymbol{S}^{-1})\boldsymbol{S} = \alpha\boldsymbol{I} \neq \boldsymbol{0}. \tag{362}$$

Thus the initialization has the claimed radius and can be arbitrarily close to the optimum.

EXACT ZERO-CHAIN PROPAGATION

We first verify that every trajectory in Equation (308) stays in the circulant subspace. Suppose inductively that the previous iterates and gradients are circulant. Entries on the same cyclic diagonal then have identical gradient histories. Because the same map Φ_t is used at every entry, their displacements from $\boldsymbol{Q}_0$ are identical, and hence $\boldsymbol{Q}_t$ is circulant. Since circulant matrices form a commuting algebra, $\nabla f(\boldsymbol{Q}_t) = \boldsymbol{S}^2(\boldsymbol{Q}_t - \boldsymbol{S}^{-1})\boldsymbol{S}$ is circulant as well. This closes the induction. It applies pathwise under common randomness and, in particular, to both subclasses stated in the theorem.

Write

$$\boldsymbol{X}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star = \mathrm{Circ}(\boldsymbol{x}_t). \tag{363}$$

Using $\boldsymbol{S}^3 = \boldsymbol{A}$ and commutativity,

$$\nabla f(\boldsymbol{Q}_t) = \boldsymbol{S}^2\boldsymbol{X}_t\boldsymbol{S} = \boldsymbol{A}\boldsymbol{X}_t = \mathrm{Circ}(\boldsymbol{A}\boldsymbol{x}_t), \tag{364}$$

$$f(\boldsymbol{Q}_t) = \frac{1}{2}\langle \boldsymbol{X}_t, \boldsymbol{A}\boldsymbol{X}_t\rangle = \frac{d}{2}\,\boldsymbol{x}_t^\top \boldsymbol{A}\boldsymbol{x}_t. \tag{365}$$

At initialization,

$$\boldsymbol{x}_0 = \alpha\boldsymbol{A}^{-1}\boldsymbol{e}_0, \qquad \boldsymbol{A}\boldsymbol{x}_0 = \alpha\boldsymbol{e}_0. \tag{366}$$

For an integer $s \geq 0$, let

$$B_s \coloneqq \{j : \mathsf{dist}_{\mathrm{cyc}}(j,0) \leq s\}, \qquad \mathsf{dist}_{\mathrm{cyc}}(j,0) = \min\{j, d-j\}, \tag{367}$$

and set $B_{-1} = \varnothing$. If $\boldsymbol{z}_t$ is the first column of $\boldsymbol{Q}_t - \boldsymbol{Q}_0$, then

$$\mathrm{supp}(\boldsymbol{z}_t) \subseteq B_{t-1}, \qquad \mathrm{supp}(\boldsymbol{A}\boldsymbol{x}_t) \subseteq B_t. \tag{368}$$

Indeed, the claim is immediate at $t = 0$ from Equation (366). If $\mathrm{supp}(\boldsymbol{z}_t) \subseteq B_{t-1}$, the nearest-neighbor structure of $\boldsymbol{A}$ gives

$$\mathrm{supp}(\boldsymbol{A}\boldsymbol{x}_t) = \mathrm{supp}(\alpha\boldsymbol{e}_0 + \boldsymbol{A}\boldsymbol{z}_t) \subseteq B_t.$$

Conversely, outside B_{t-1} every gradient entry through time $t-1$ has been exactly zero. The condition $\Phi_t(0,\dots,0) = 0$ therefore gives $\mathrm{supp}(\boldsymbol{z}_t) \subseteq B_{t-1}$. This proves Equation (368) by induction.

THE CYCLIC GREEN-FUNCTION BARRIER

Put $s = \sqrt{\kappa}$. The value $q = (s-1)/(s+1)$ satisfies

$$q + q^{-1} = \frac{2(\kappa+1)}{\kappa - 1}. \tag{369}$$

A direct substitution into the cyclic three-term recurrence for $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{e}_0$ gives

$$v_j = \frac{q^j + q^{d-j}}{s(1-q^d)}, \qquad 0 \leq j \leq d-1. \tag{370}$$

Indeed, the homogeneous relation for $1 \leq j \leq d-1$ follows from the displayed identity for $q + q^{-1}$, including the neighbor pair across the cyclic cut. At $j = 0$, substituting v_0, v_1, and v_{d-1} gives exactly one; the factor $s(1-q^d)$ is the resulting normalization. Thus the wrap-around equation is included rather than replaced by a noncyclic boundary condition. By the definition of R_ε, $q^{2R_\varepsilon} \leq 1/2$. Moreover, $d \geq 4R_\varepsilon$, so $q^d \leq 1/4$.

Fix $1 \leq t \leq m_\varepsilon$. By Equation (368), every cyclic distance $j = t, \dots, d/4 - 1$ is untouched at time t. There are two coordinates at each such distance, and Equation (370) yields

$$\begin{aligned}\|\boldsymbol{x}_t\|_2^2 &\geq \frac{2\alpha^2}{\kappa}\sum_{j=t}^{d/4-1} q^{2j} \\ &\geq \frac{\alpha^2}{\kappa(1-q^2)}q^{2t}.\end{aligned} \tag{371}$$

For the last inequality, note that $d/4 - 1 = m_\varepsilon + R_\varepsilon$; hence the sum contains at least $R_\varepsilon + 1$ terms and $2(1 - q^{2(R_\varepsilon+1)}) \geq 1$.

Since $\boldsymbol{A} \succeq \boldsymbol{I}$, Equation (365) gives

$$f(\boldsymbol{Q}_t) \geq \frac{d\alpha^2}{2\kappa(1-q^2)} q^{2t}. \tag{372}$$

At time zero, Equation (370) gives

$$f(\boldsymbol{Q}_0) = \frac{d\alpha^2}{2} \boldsymbol{e}_0^\top \boldsymbol{A}^{-1} \boldsymbol{e}_0 = \frac{d\alpha^2}{2s} \frac{1+q^d}{1-q^d} \leq \frac{d\alpha^2}{s}. \tag{373}$$

Finally,

$$s(1-q^2) = \frac{4s^2}{(s+1)^2} \leq 4. \tag{374}$$

Combining the last three displays proves

$$f(\boldsymbol{Q}_t) \geq \frac{1}{8} f(\boldsymbol{Q}_0) q^{2t} \tag{375}$$

for $1 \leq t \leq m_\varepsilon$; the case $t = 0$ is immediate.

ITERATION CONSEQUENCE

If an integer $T \leq m_\varepsilon$ satisfies the target in Equation (353), then the barrier forces $q^{2T} \leq \varepsilon$, or $T \geq T_\varepsilon$. If $T > m_\varepsilon$, the same conclusion is automatic. Finally, because $s = \sqrt{\kappa} \geq 2$,

$$\log \frac{1}{q} = \log\left(1 + \frac{2}{s-1}\right) \leq \frac{2}{s-1}, \qquad \frac{1}{2\log(1/q)} \geq \frac{s-1}{4} \geq \frac{s}{8}. \tag{376}$$

This proves Equation (353) and completes the proof of Theorem 4.

J A MATCHED LOCAL SEPARATION BETWEEN STABILIZED ADAM AND MUON

We prove Proposition 6. Unlike the leading-block construction in Appendix G, the construction here permits a fast polar trajectory. Its lower bound is specific to stabilized Adam with bounded scalar learning rates.

We use the full-batch Adam recursion in Equation (311), with zero initial moments and no weight decay, and add a finite cap on its scalar learning rates.

Proposition 6 (Same-instance local separation)**.** *Fix $\beta_1, \beta_2 \in [0,1)$, $\epsilon_{\mathrm{A}} > 0$, and $\Gamma > 0$. For every $\kappa \geq 4$, $\varepsilon \in (0,1)$, and $r_0 > 0$, there exist $d = O(\sqrt{\kappa}(1 + \log(1/\varepsilon)))$, a positive-definite target $\boldsymbol{M}^\star$ with condition number κ, and a common square initialization $\boldsymbol{U}_0$ with $0 < \|\boldsymbol{U}_0 - (\boldsymbol{M}^\star)^{1/2}\|_{\mathrm{F}} \leq r_0$ such that:*

- *Every* Adam *trajectory with $0 \leq \gamma_t \leq \Gamma$, including adaptively or randomly chosen scalar rates, satisfies*

$$f(\boldsymbol{U}_T^{\mathrm{A}}) \leq \frac{\varepsilon}{64} f(\boldsymbol{U}_0) \quad \Longrightarrow \quad T \geq \frac{\sqrt{\kappa}}{8} \log \frac{1}{\varepsilon}. \tag{377}$$

- *For every fixed $\mu \in [0,1)$, exact-polar* Muon *with zero initial buffer and a deterministic geometric schedule satisfies*

$$f(\boldsymbol{U}_T^{\mathrm{M}}) \leq \frac{\varepsilon}{64} f(\boldsymbol{U}_0) \quad \textit{for all} \quad T \geq \left\lceil \frac{\log(144\kappa^2/\varepsilon)}{-2\log\rho} \right\rceil, \qquad \rho = \frac{4+\mu}{5}. \tag{378}$$

The construction takes $\boldsymbol{M}^\star = \boldsymbol{A}/2$ and $\boldsymbol{U}_0 = (\boldsymbol{A}/2)^{1/2} + \delta \boldsymbol{A}^{-1}$, where $\boldsymbol{A}$ is the circulant quadratic Hessian in Equation (355) and $\delta > 0$ is sufficiently small. The Muon schedule is $\eta_t = (1-\rho)\delta\rho^t$. The proof uses a uniform finite-horizon limit: near this factor, stabilized Adam approaches a scalar-stepsize momentum method on the hard quadratic, whereas Muon contracts the aligned spectral errors geometrically. The scale δ may depend on the Adam parameters, rate cap, condition number, and accuracy. This is a stabilizer-dominated local separation; it does not cover $\epsilon_{\mathrm{A}} = 0$, the unrestricted coordinatewise history class, or the isotropic initializations of Theorem 1.

J.1 A COMMON LOCAL INITIALIZATION

Fix the parameters of Proposition 6. Set

$$q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad T_\varepsilon = \frac{\log(1/\varepsilon)}{2\log(1/q)}, \qquad N = \lceil T_\varepsilon \rceil,$$
$$R = \left\lceil \frac{\log 2}{2\log(1/q)} \right\rceil, \qquad d = 4(N+R+1). \tag{379}$$

Thus $d = O(\sqrt{\kappa}(1+\log(1/\varepsilon)))$. Let $\boldsymbol{P}$ be the cyclic shift in dimension d, and define

$$\boldsymbol{A} = \frac{\kappa+1}{2}\boldsymbol{I} - \frac{\kappa-1}{4}(\boldsymbol{P}+\boldsymbol{P}^\top), \qquad \boldsymbol{M}^\star = \frac{\boldsymbol{A}}{2}, \qquad \boldsymbol{S} = (\boldsymbol{A}/2)^{1/2},$$
$$\boldsymbol{U}_0 = \boldsymbol{S} + \delta \boldsymbol{A}^{-1}. \tag{380}$$

The spectrum in Equation (357) gives $\lambda_{\min}(\boldsymbol{A}) = 1$ and $\lambda_{\max}(\boldsymbol{A}) = \kappa$. Both $\boldsymbol{M}^\star$ and $\boldsymbol{U}_0$ are positive definite for $\delta > 0$, and $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{M}^\star$. We will choose δ sufficiently small that

$$0 < \delta \le \min\left\{ \frac{1}{4\sqrt{2}}, \frac{r_0}{\|\boldsymbol{A}^{-1}\|_{\mathrm{F}}} \right\}, \tag{381}$$

which ensures $0 < \|\boldsymbol{U}_0 - \boldsymbol{S}\|_{\mathrm{F}} \le r_0$.

The real symmetric circulant matrices form a commuting algebra. They are also closed under applying a common scalar function entrywise. Starting from Equation (380), the exact gradient, the two Adam moment states, and the Adam iterates therefore remain symmetric circulant. Indeed, matrix products preserve this algebra, while entrywise squaring, square roots, and division preserve symmetry and cyclic-diagonal equality; the denominator is strictly positive. For any symmetric circulant $\boldsymbol{X}$, commutativity and $2\boldsymbol{S}^2 = \boldsymbol{A}$ give the exact identities

$$\frac{\nabla f(\boldsymbol{S}+\delta\boldsymbol{X})}{\delta} = \boldsymbol{A}\boldsymbol{X} + 3\delta\boldsymbol{S}\boldsymbol{X}^2 + \delta^2\boldsymbol{X}^3, \tag{382}$$
$$F_\delta(\boldsymbol{X}) \coloneqq \frac{f(\boldsymbol{S}+\delta\boldsymbol{X})}{\delta^2} = \frac{1}{4}\|2\boldsymbol{S}\boldsymbol{X} + \delta\boldsymbol{X}^2\|_{\mathrm{F}}^2. \tag{383}$$

The latter extends continuously to $\delta = 0$, with

$$F_0(\boldsymbol{X}) = \mathcal{Q}(\boldsymbol{X}) \coloneqq \frac{1}{2}\langle \boldsymbol{X}, \boldsymbol{A}\boldsymbol{X}\rangle. \tag{384}$$

J.2 DETERMINISTIC LOCAL CONTRACTION FOR MUON

The following scalar lemma holds for arbitrary small aligned errors, not only the particular spectral values in Equation (380).

Lemma 28 (Local scalar contraction with momentum)**.** *Let $a > 0$, $\mu \in [0,1)$, and $0 < \delta \le a/4$. Set $\rho = (4+\mu)/5$ and $\eta_t = (1-\rho)\delta\rho^t$. The recursion*

$$b_t = \mu b_{t-1} + h_a(e_t), \qquad h_a(e) = 2a^2 e + 3ae^2 + e^3,$$
$$e_{t+1} = e_t - \eta_t \,\mathsf{sign}(b_t), \qquad b_{-1} = 0 \tag{385}$$

with $\mathsf{sign}(0) = 0$ and $|e_0| \le \delta$ satisfies $|e_t| \le \delta\rho^t$ for every $t \ge 0$.

Proof. For $|e| \le a/4$, direct differentiation and factorization give

$$\mathsf{sign}(h_a(e)) = \mathsf{sign}(e), \qquad |h_a(e)| \ge a^2|e|, \qquad |h_a'(e)| \le 4a^2. \tag{386}$$

Suppose inductively that $|e_j| \le \delta\rho^j$ for $0 \le j \le t$. Put $W_t = \sum_{j=0}^t \mu^j$ and $y_t = b_t/W_t$. Every past error lies in $[-a/4, a/4]$, and $|e_{t-j} - e_t| \le \eta_t \sum_{\ell=1}^j \rho^{-\ell}$. Hence

$$|y_t - h_a(e_t)| \le \frac{4a^2\eta_t}{W_t}\sum_{j=0}^t \mu^j \sum_{\ell=1}^j \rho^{-\ell} \le 4a^2\eta_t \frac{\mu}{\rho-\mu}. \tag{387}$$

For the second inequality, exchange the finite sums and use $W_t^{-1}\sum_{j=\ell}^{t}\mu^j \le \mu^\ell$, followed by $\sum_{\ell=1}^{t}(\mu/\rho)^\ell \le \mu/(\rho-\mu)$; here $\rho > \mu$.

If $e_t b_t > 0$, the signs agree and

$$|e_{t+1}| = \big||e_t| - \eta_t\big| \le \max\{\eta_t, \delta\rho^t - \eta_t\} \le \delta\rho^{t+1},$$

since $\rho \ge 1/2$. If $e_t b_t \le 0$, including either zero case, Equations (386) and (387) imply $a^2|e_t| \le |h_a(e_t)| \le |y_t - h_a(e_t)|$. Therefore

$$|e_{t+1}| \le \left(1 + \frac{4\mu}{\rho-\mu}\right)\eta_t = \frac{1+4\mu}{5}\delta\rho^t \le \delta\rho^{t+1}.$$

This closes the induction. □

Let λ_i be the eigenvalues of $\boldsymbol{A}$ and set $a_i = \sqrt{\lambda_i/2}$. The exact aligned reduction in Proposition 3 applies with identical left and right eigenframes. Writing the spectral coordinates of $\boldsymbol{U}_t^{\mathrm{M}}$ as $a_i + e_{i,t}$ gives Equation (385) and $e_{i,0} = \delta/\lambda_i$. Condition (381) ensures $|e_{i,0}| \le \delta \le a_i/4$, so Lemma 28 yields

$$|e_{i,t}| \le \delta\rho^t \qquad \text{for all } i, t. \tag{388}$$

Consequently, using $|2a_i + e_{i,t}| \le 3a_i$,

$$f(\boldsymbol{U}_t^{\mathrm{M}}) = \frac{1}{4}\sum_i e_{i,t}^2(2a_i + e_{i,t})^2 \le \frac{9}{8}\delta^2\rho^{2t}\,\mathrm{tr}(\boldsymbol{A}),$$

$$f(\boldsymbol{U}_0) \ge \sum_i a_i^2 e_{i,0}^2 = \frac{\delta^2}{2}\,\mathrm{tr}(\boldsymbol{A}^{-1}). \tag{389}$$

The lower bound uses $e_{i,0} > 0$. Since the eigenvalues of $\boldsymbol{A}$ lie in $[1, \kappa]$, we obtain

$$\frac{f(\boldsymbol{U}_t^{\mathrm{M}})}{f(\boldsymbol{U}_0)} \le \frac{9}{4}\frac{\mathrm{tr}(\boldsymbol{A})}{\mathrm{tr}(\boldsymbol{A}^{-1})}\rho^{2t} \le \frac{9}{4}\kappa^2\rho^{2t}. \tag{390}$$

This proves Equation (378). Since $-\log\rho \ge 1 - \rho = (1-\mu)/5$, its iteration count is $O((\log\kappa + \log(1/\varepsilon))/(1-\mu))$.

J.3 A UNIFORM QUADRATIC LIMIT FOR STABILIZED ADAM

For a prescribed rate vector $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_{N-1}) \in [0, \Gamma]^N$, apply Equation (311) to the same instance and initialization. Normalize its states by

$$\boldsymbol{X}_t^{(\delta)} = \frac{\boldsymbol{U}_t^{\mathrm{A}} - \boldsymbol{S}}{\delta}, \qquad \overline{\boldsymbol{M}}_t = \frac{\boldsymbol{M}_t}{\delta}, \qquad \overline{\boldsymbol{V}}_t = \frac{\boldsymbol{V}_t}{\delta^2}. \tag{391}$$

With $\overline{\boldsymbol{M}}_{-1} = \overline{\boldsymbol{V}}_{-1} = \boldsymbol{0}$ and $\boldsymbol{X}_0^{(\delta)} = \boldsymbol{A}^{-1}$, their exact recursions are

$$\begin{aligned}
\overline{\boldsymbol{G}}_t &= \boldsymbol{A}\boldsymbol{X}_t^{(\delta)} + 3\delta\boldsymbol{S}(\boldsymbol{X}_t^{(\delta)})^2 + \delta^2(\boldsymbol{X}_t^{(\delta)})^3,\\
\overline{\boldsymbol{M}}_t &= \beta_1\overline{\boldsymbol{M}}_{t-1} + (1-\beta_1)\overline{\boldsymbol{G}}_t,\\
\overline{\boldsymbol{V}}_t &= \beta_2\overline{\boldsymbol{V}}_{t-1} + (1-\beta_2)(\overline{\boldsymbol{G}}_t \odot \overline{\boldsymbol{G}}_t),\\
\boldsymbol{X}_{t+1}^{(\delta)} &= \boldsymbol{X}_t^{(\delta)} - \gamma_t\frac{\widehat{\overline{\boldsymbol{M}}}_t}{\epsilon_{\mathrm{A}} + \delta\sqrt{\widehat{\overline{\boldsymbol{V}}}_t}}.
\end{aligned} \tag{392}$$

Hats denote the same bias corrections as in Equation (311), and the division and square root are entrywise. Setting $\delta = 0$ defines the reference trajectory

$$\begin{aligned}
\boldsymbol{Y}_0 &= \boldsymbol{A}^{-1}, & \boldsymbol{L}_{-1} &= \boldsymbol{0},\\
\boldsymbol{L}_t &= \beta_1\boldsymbol{L}_{t-1} + (1-\beta_1)\boldsymbol{A}\boldsymbol{Y}_t,\\
\boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \frac{\gamma_t}{\epsilon_{\mathrm{A}}(1-\beta_1^{t+1})}\boldsymbol{L}_t.
\end{aligned} \tag{393}$$

This is a scalar-stepsize momentum method for $\mathcal{Q}$; the second moment drops out of the update.

For each fixed finite N, all states in Equation (392) are jointly continuous in $(\delta, \boldsymbol{\gamma}) \in [0,1] \times [0,\Gamma]^N$. This follows by induction: the moments are continuous, the second moment is entrywise nonnegative, and every update denominator is at least $\epsilon_{\mathrm{A}} > 0$. The continuous finite-step maps are uniformly continuous on this compact set. Together with Equations (383) and (384), this proves

$$\sup_{\boldsymbol{\gamma}\in[0,\Gamma]^N} \max_{0\le t\le N} \left| F_\delta(\boldsymbol{X}_t^{(\delta)}) - \mathcal{Q}(\boldsymbol{Y}_t) \right| \longrightarrow 0 \qquad \text{as } \delta \downarrow 0. \tag{394}$$

No continuity of a rate-selection policy is needed: for adaptive or randomized rates, feed the reference method the realized rate vector. The bound is uniform over all such vectors and therefore holds pathwise.

J.4 The reference barrier and transfer to actual Adam

At time zero, $\boldsymbol{A}\boldsymbol{Y}_0 = \boldsymbol{I}$. Multiplication by $\boldsymbol{A}$ enlarges cyclic-diagonal support by at most one. Induction in Equation (393) therefore shows that $\boldsymbol{Y}_t - \boldsymbol{A}^{-1}$ is supported on cyclic diagonals of distance at most $t-1$, with empty support at $t = 0$. This statement holds for every scalar rate vector, independently of how its entries are chosen.

For completeness, write $\boldsymbol{Y}_t = \mathrm{Circ}(\boldsymbol{y}_t)$ and $\boldsymbol{v} = \boldsymbol{A}^{-1}\boldsymbol{e}_0$. The cyclic Green function from Equation (370) is

$$v_j = \frac{q^j + q^{d-j}}{\sqrt{\kappa}(1-q^d)}, \qquad 0 \le j \le d-1. \tag{395}$$

For $1 \le t \le N$, the entries at cyclic distances $j = t, \dots, d/4-1$ are still equal to those of $\boldsymbol{v}$. There are two entries at each such distance. Since $q^{2R} \le 1/2$ and $d/4 - 1 = N + R$, the same finite-tail calculation as in Equation (371) gives

$$\|\boldsymbol{y}_t\|_2^2 \ge \frac{2}{\kappa} \sum_{j=t}^{d/4-1} q^{2j} \ge \frac{q^{2t}}{\kappa(1-q^2)}. \tag{396}$$

Moreover, $q^d \le 1/4$ and $\boldsymbol{A} \succeq \boldsymbol{I}$, so

$$\begin{aligned} \mathcal{Q}(\boldsymbol{Y}_t) &= \frac{d}{2}\boldsymbol{y}_t^\top \boldsymbol{A}\boldsymbol{y}_t \ge \frac{d}{2\kappa(1-q^2)} q^{2t}, \\ \mathcal{Q}(\boldsymbol{Y}_0) &= \frac{d}{2\sqrt{\kappa}} \frac{1+q^d}{1-q^d} \le \frac{d}{\sqrt{\kappa}}. \end{aligned} \tag{397}$$

Using $\sqrt{\kappa}(1-q^2) \le 4$, and including the immediate case $t = 0$, yields the uniform reference barrier

$$\mathcal{Q}(\boldsymbol{Y}_t) \ge \frac{1}{8}\mathcal{Q}(\boldsymbol{Y}_0) q^{2t}, \qquad 0 \le t \le N, \tag{398}$$

Now choose $\delta > 0$ small enough to satisfy Equation (381) and, by Equation (394),

$$\begin{aligned} \sup_{\boldsymbol{\gamma}\in[0,\Gamma]^N} \max_{0\le t\le N} \left| F_\delta(\boldsymbol{X}_t^{(\delta)}) - \mathcal{Q}(\boldsymbol{Y}_t) \right| &\le \frac{\varepsilon}{16}\mathcal{Q}(\boldsymbol{Y}_0), \\ F_\delta(\boldsymbol{A}^{-1}) &\le 2\mathcal{Q}(\boldsymbol{Y}_0). \end{aligned} \tag{399}$$

This choice is common to all rate policies and all $\mu \in [0,1)$. For every integer $t < T_\varepsilon$, we have $q^{2t} > \varepsilon$; Equations (398) and (399) imply

$$\frac{f(\boldsymbol{U}_t^{\mathrm{A}})}{\delta^2} > \frac{\varepsilon}{16}\mathcal{Q}(\boldsymbol{Y}_0), \qquad \frac{f(\boldsymbol{U}_t^{\mathrm{A}})}{f(\boldsymbol{U}_0)} > \frac{\varepsilon}{32}. \tag{400}$$

Thus no such iterate reaches the common target $\varepsilon f(\boldsymbol{U}_0)/64$. Finally, for $\sqrt{\kappa} \ge 2$,

$$\log(1/q) = \log\left(1 + \frac{2}{\sqrt{\kappa}-1}\right) \le \frac{4}{\sqrt{\kappa}}, \qquad T_\varepsilon \ge \frac{\sqrt{\kappa}}{8}\log\frac{1}{\varepsilon}. \tag{401}$$

This proves Equation (377) and completes the proposition.

The instance and initialization are chosen for each requested accuracy; the proposition does not assert one fixed hard initialization for all accuracies. The finite-horizon continuity argument relies on both $\epsilon_{\mathrm{A}} > 0$ and the fixed cap Γ. It gives a local, stabilizer-dominated result, not a lower bound for arbitrary nonlinear coordinatewise history maps, which may amplify the small nonlinear gradient terms in Equation (382).

K TECHNICAL LEMMAS

In this section, we gather several technical lemmas used in our analysis. We begin with a polar-perturbation bound and a partial-isometry completion lemma.

Lemma 29 (Common-rank partial-polar perturbation)**.** *Let $1 \le r \le \min\{m, n\}$, and let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$ both have rank r. Then*

$$\|\mathsf{msign}(\boldsymbol{X}) - \mathsf{msign}(\boldsymbol{Y})\| \le \frac{2}{\min\{\sigma_r(\boldsymbol{X}), \sigma_r(\boldsymbol{Y})\}} \|\boldsymbol{X} - \boldsymbol{Y}\|. \tag{402}$$

Proof. For $\boldsymbol{Z} = \boldsymbol{X}, \boldsymbol{Y}$, form the symmetric dilation

$$\boldsymbol{H}_{\boldsymbol{Z}} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{Z} \\ \boldsymbol{Z}^\top & \boldsymbol{0} \end{pmatrix}.$$

Its positive and negative spectral subspaces both have dimension r, and its nonzero eigenvalues have absolute value at least $s := \min\{\sigma_r(\boldsymbol{X}), \sigma_r(\boldsymbol{Y})\}$. Write $\boldsymbol{P}^+_{\boldsymbol{Z}}$ and $\boldsymbol{P}^-_{\boldsymbol{Z}}$ for the corresponding spectral projections. The positive subspace of $\boldsymbol{H}_{\boldsymbol{X}}$ is separated by at least s from the nonpositive subspace of $\boldsymbol{H}_{\boldsymbol{Y}}$. To see the projection bound, put $\boldsymbol{P} = \boldsymbol{P}^+_{\boldsymbol{X}}$, $\boldsymbol{Q} = \boldsymbol{P}^+_{\boldsymbol{Y}}$, and $\boldsymbol{E} = \boldsymbol{H}_{\boldsymbol{X}} - \boldsymbol{H}_{\boldsymbol{Y}}$. The cross projection $\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{P}$ satisfies a Sylvester equation $\boldsymbol{B}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{A} = -(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{E}\boldsymbol{P}$, where $\boldsymbol{A} = \boldsymbol{H}_{\boldsymbol{X}}|_{\mathrm{range}\,\boldsymbol{P}} \succeq s\boldsymbol{I}$ and $\boldsymbol{B} = \boldsymbol{H}_{\boldsymbol{Y}}|_{\mathrm{range}(\boldsymbol{I}-\boldsymbol{Q})} \preceq 0$. The solution $\boldsymbol{Z} = \int_0^\infty e^{t\boldsymbol{B}}(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{E}\boldsymbol{P}e^{-t\boldsymbol{A}}\,\mathrm{d}t$ gives $\|\boldsymbol{Z}\| \le \|\boldsymbol{E}\|/s$. Interchanging $\boldsymbol{X}$ and $\boldsymbol{Y}$ bounds the other cross projection; the standard identity for orthogonal projections then gives

$$\|\boldsymbol{P}^+_{\boldsymbol{X}} - \boldsymbol{P}^+_{\boldsymbol{Y}}\| \le \frac{\|\boldsymbol{H}_{\boldsymbol{X}} - \boldsymbol{H}_{\boldsymbol{Y}}\|}{s} = \frac{\|\boldsymbol{X} - \boldsymbol{Y}\|}{s}.$$

The same argument applies to the negative projections. Since $\mathsf{sign}(\boldsymbol{H}_{\boldsymbol{Z}}) = \boldsymbol{P}^+_{\boldsymbol{Z}} - \boldsymbol{P}^-_{\boldsymbol{Z}}$ and its off-diagonal block is $\mathsf{msign}(\boldsymbol{Z})$, the triangle inequality gives the result. Related polar-factor perturbation bounds appear in Li & Sun (2006). □

Lemma 30 (Rectangular partial-isometry completion)**.** *Let $d \ge k$ and $1 \le r \le k$. Suppose that $\boldsymbol{A}, \boldsymbol{C} \in \mathbb{R}^{d\times k}$ are partial isometries of ranks r and $k - r$, respectively, with*

$$\boldsymbol{A}^\top\boldsymbol{C} = \boldsymbol{0}, \qquad \boldsymbol{A}\boldsymbol{C}^\top = \boldsymbol{0}, \qquad \boldsymbol{F} := \boldsymbol{A} + \boldsymbol{C}, \qquad \boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_k. \tag{403}$$

For every rank-r partial isometry $\boldsymbol{B} \in \mathbb{R}^{d\times k}$, there exists a rank-$(k - r)$ partial isometry $\boldsymbol{D}$ such that

$$\boldsymbol{B}^\top\boldsymbol{D} = \boldsymbol{0}, \qquad \boldsymbol{B}\boldsymbol{D}^\top = \boldsymbol{0}, \qquad (\boldsymbol{B} + \boldsymbol{D})^\top(\boldsymbol{B} + \boldsymbol{D}) = \boldsymbol{I}_k, \tag{404}$$

and

$$\big\|\boldsymbol{F} - (\boldsymbol{B} + \boldsymbol{D})\big\| \le 25\|\boldsymbol{A} - \boldsymbol{B}\|. \tag{405}$$

When $r = k$, the conclusion holds with $\boldsymbol{C} = \boldsymbol{D} = \boldsymbol{0}$.

Proof. The case $r = k$ is immediate, so assume $r < k$ and put

$$\delta := \|\boldsymbol{A} - \boldsymbol{B}\|, \quad \boldsymbol{P}_A := \boldsymbol{A}\boldsymbol{A}^\top, \quad \boldsymbol{Q}_A := \boldsymbol{A}^\top\boldsymbol{A}, \quad \boldsymbol{P} := \boldsymbol{B}\boldsymbol{B}^\top, \quad \boldsymbol{Q} := \boldsymbol{B}^\top\boldsymbol{B}.$$

If $\delta \ge 1/12$, choose any rank-$(k - r)$ partial isometry $\boldsymbol{D}$ whose initial space is $\mathrm{range}(\boldsymbol{I}_k - \boldsymbol{Q})$ and whose final space is a $(k - r)$-dimensional subspace of $\mathrm{range}(\boldsymbol{I}_d - \boldsymbol{P})$. Such a choice exists because $d - r \ge k - r$. Then $\boldsymbol{B} + \boldsymbol{D}$ has orthonormal columns, and

$$\|\boldsymbol{F} - (\boldsymbol{B} + \boldsymbol{D})\| \le 2 \le 25\delta.$$

It remains to consider $\delta < 1/12$. Define

$$\boldsymbol{Z} \coloneqq (\boldsymbol{I}_d - \boldsymbol{P})\boldsymbol{C}(\boldsymbol{I}_k - \boldsymbol{Q}). \tag{406}$$

The standard product bounds give

$$\|\boldsymbol{P} - \boldsymbol{P}_A\| \le 2\delta, \qquad \|\boldsymbol{Q} - \boldsymbol{Q}_A\| \le 2\delta.$$

Because $\boldsymbol{P}_A\boldsymbol{C} = \boldsymbol{0}$ and $\boldsymbol{C}\boldsymbol{Q}_A = \boldsymbol{0}$, it follows that

$$\begin{aligned} e \coloneqq \|\boldsymbol{Z} - \boldsymbol{C}\| &\le \|\boldsymbol{P}\boldsymbol{C}\| + \|\boldsymbol{C}\boldsymbol{Q}\| + \|\boldsymbol{P}\boldsymbol{C}\boldsymbol{Q}\| \\ &\le 6\delta < \frac{1}{2}. \end{aligned} \tag{407}$$

The $k - r$ positive singular values of $\boldsymbol{C}$ equal one. Weyl's inequality therefore gives $\sigma_{k-r}(\boldsymbol{Z}) \ge 1 - e > 0$. Since the right range of $\boldsymbol{Z}$ lies in the $(k-r)$-dimensional space $\operatorname{range}(\boldsymbol{I}_k - \boldsymbol{Q})$, the matrix $\boldsymbol{Z}$ has rank exactly $k - r$. Set $\boldsymbol{D} \coloneqq \mathsf{msign}(\boldsymbol{Z})$. Its initial space is $\operatorname{range}(\boldsymbol{I}_k - \boldsymbol{Q})$ and its final space is orthogonal to $\operatorname{range}(\boldsymbol{P})$, so the required orthogonality relations hold and $\boldsymbol{B} + \boldsymbol{D}$ has orthonormal columns.

Finally, Lemma 29 and eq. (407) yield

$$\|\boldsymbol{D} - \boldsymbol{C}\| \le \frac{2e}{1 - e} \le 4e \le 24\delta.$$

Thus

$$\|\boldsymbol{F} - (\boldsymbol{B} + \boldsymbol{D})\| \le \|\boldsymbol{A} - \boldsymbol{B}\| + \|\boldsymbol{C} - \boldsymbol{D}\| \le 25\delta,$$

which proves the claim. □

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

Lemma 31 (Adapted from Theorem 6.1 in Wainwright (2019))**.** *Suppose that $\boldsymbol{G} \in \mathbb{R}^{d_1 \times d_2}$ is a standard Gaussian matrix, where $d_1 \ge d_2$. Then, for every $t \ge 0$,*

$$\mathbb{P}\left(\|\boldsymbol{G}\| \ge \sqrt{d_1} + \sqrt{d_2} + t\right) \le \exp\left(-\frac{t^2}{2}\right). \tag{408}$$

Lemma 32 (Adapted from Equation (3.2) in Rudelson & Vershynin (2010))**.** *Suppose that the entries of $\boldsymbol{G} \in \mathbb{R}^{d \times d}$ are i.i.d. standard Gaussian random variables. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{G}) \le \varepsilon d^{-1/2}\right) \le \varepsilon. \tag{409}$$

L ADDITIONAL NUMERICAL EXPERIMENTS

This section gives the protocols and tuning comparisons behind Figures 1 and 2, then examines the search-rank regimes of Theorem 1 and the zero-initialized operator envelope underlying Theorem 3. It also tests finite orthogonalization and the matched local Adam–Muon mechanism of Proposition 6. All experiments use full-batch gradients; Muon uses exact partial polar updates except in the finite-orthogonalization comparison in Appendix L.5. Step amplitudes are deterministic; variability comes from the stated initialization or eigenbasis draws.

L.1 FACTORIZATION OF THE HARD QUADRATIC HESSIAN

Target and initialization. For Figure 1, we use the same circulant $\boldsymbol{A}$ used in the ICL study, including its dimension $d = 4096$ and grid $\kappa \in \{5, 25, 125, 625, 3125, 15625\}$, and set

$$\boldsymbol{M}^\star = \frac{\boldsymbol{A}}{\kappa} = \frac{\kappa+1}{2\kappa}\boldsymbol{I} - \frac{\kappa-1}{4\kappa}(\boldsymbol{P} + \boldsymbol{P}^\top). \tag{410}$$

Its eigenvalues range from $1/\kappa$ to 1. The target and factor are full rank, $r = k = d$, so the factorization is exactly parameterized and lies in the wide regime of Theorem 1. For each seed, draw $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and set $\boldsymbol{U}_0 = 0.1\boldsymbol{O}$, where $\boldsymbol{O} = \mathsf{msign}(\mathrm{Circ}(\boldsymbol{g}))$. The Gaussian circulant is nonsingular almost surely, so $\boldsymbol{O}$ is real orthogonal and $\boldsymbol{U}_0\boldsymbol{U}_0^\top = 0.01\boldsymbol{I}$. The initial reconstruction error is 0.99 at every κ. The same draw is paired across condition numbers and algorithms, independently of the target. This samples the circulant orthogonal class, not the full Haar measure.

Exact algorithms and tuning. All three methods use the zero-initialized buffer recursion stated in Section 2.1.4, with $\mu \in \{0, 0.9\}$. The circulant algebra is invariant: in Fourier coordinates, if u_j and m_j are the factor and target eigenvalues, the gradient has eigenvalues $(|u_j|^2 - m_j)u_j$. We use this exact representation for Muon and GD; the polar direction is the SVD partial polar, including zero singular values. SignGD takes the entrywise sign in the original matrix coordinates, represented by its first column. Dense checks verify all six optimizer recurrences, the gradient, and the norm calculations. Muon and GD reconstruction trajectories are invariant to the initial orthogonal phase; their seed bands therefore coincide. SignGD depends on this phase.

For multiplier ξ, Muon uses $\eta_t = 2(1-\rho)\xi\rho^t$ with $\rho = \max\{0.98, (1+\mu^{1/3})/2\}$; SignGD starts at $\eta_0 = \xi/\sqrt{d}$ and multiplies the rate by 0.3 after 50 consecutive nonimproving pre-update objectives; GD uses constant $\eta = (1+\mu)\xi/4$. The last scale is a local normalized reference, not a global smoothness bound for this quartic objective. On seeds 1000, 1001, we search $\xi = 2^{j/2}$ for $j = -16, \ldots, 8$, extending selected boundaries by four half-octaves before evaluation. No final selection is on a search boundary. We select rates separately for first hits of reconstruction error 10^{-6} and relative objective 10^{-6}, ranking numerical failures, unreached seeds, median and mean hitting times, final metric, then the smaller multiplier. The fixed protocol transfers the $\kappa = 5$ choice unchanged; the tuned protocol selects separately at each κ. Both evaluate held-out seeds $0, \ldots, 9$ and save all 12,000 updates without convergence-based stopping. The archive retains 1,832 tuning runs and 750 distinct evaluation runs, including 127 divergent tuning trials and no evaluation numerical failures.

Condition-number dependence. Figure 4 compares the protocols. At $\mu = 0.9$, Muon reaches reconstruction error 10^{-6} in 665–670 iterations throughout the grid under both protocols. Momentum SignGD remains nonmonotone for $\kappa \le 625$, then slows at $\kappa = 3125, 15625$. The separately tuned rate at $\kappa = 15625$ reaches 10^{-6} on $9/10$ evaluation seeds; its identifiable median is 6,203.5, while the fixed rate reaches all ten with median 6,035.5. Thus selection on separate seeds can transfer imperfectly. Without momentum, neither SignGD nor GD reaches the tolerance at $\kappa = 3125, 15625$ on any evaluation seed under either protocol. These finite-grid observations support weak condition-number dependence for Muon. Reusing the quadratic Hessian does not transfer its lower-bound argument to the quartic factorization objective.

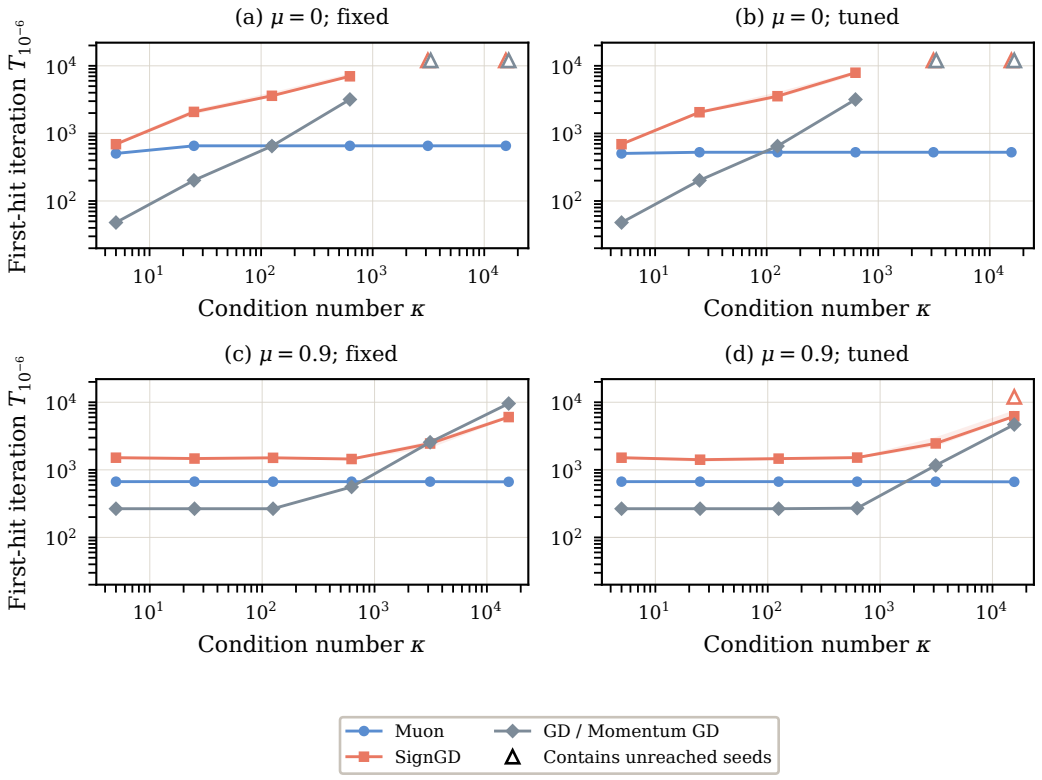


Figure 4: Full-rank factorization of $\boldsymbol{A}/\kappa$: first hits of $\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star\| \le 10^{-6}$. Top: $\mu = 0$; bottom: $\mu = 0.9$. Left: fixed normalized rates; right: per-κ rates selected on two separate seeds. Curves show medians and identifiable 20th–80th percentiles over ten evaluation seeds. Open triangles flag groups containing unreached seeds at 12,000 updates; missing medians are not connected.

L.2 Circulant ICL with nonlocal random initialization

Covariance and initialization. For Figure 2, we use $d = 4096$ at every $\kappa \in \{5, 25, 125, 625, 3125, 15625\}$. With $\boldsymbol{P}$ the cyclic shift,

$$\boldsymbol{A} = \frac{\kappa+1}{2}\boldsymbol{I} - \frac{\kappa-1}{4}(\boldsymbol{P} + \boldsymbol{P}^\top), \qquad \boldsymbol{S} = \boldsymbol{A}^{1/3}, \qquad \boldsymbol{Q}^\star = \boldsymbol{S}^{-1}. \tag{411}$$

This is the covariance family in Equation (355), with $\lambda_{\min}(\boldsymbol{S}) = 1$ and $\lambda_{\max}(\boldsymbol{S}) = \kappa^{1/3}$. For each seed, draw independent $g_0, \dots, g_{d/2} \sim \mathcal{N}(0,1)$, form $\boldsymbol{c} = (g_0, \dots, g_{d/2}, g_{d/2-1}, \dots, g_1)^\top$, and set

$$\boldsymbol{Q}_0 = \frac{2\,\mathrm{Circ}(\boldsymbol{c})}{\|\,\mathrm{Circ}(\boldsymbol{c})\|}. \tag{412}$$

The draw does not use $\boldsymbol{Q}^\star$ and is paired across condition numbers and algorithms. Since $\|\boldsymbol{Q}_0\| = 2$ and $\|\boldsymbol{Q}^\star\| = 1$, $1 \le e_0^{\mathrm{ICL}} \le 3$; observed initial errors range from 1.9316 to 2.8626 over all tuning and evaluation instances. The initialization is random within the symmetric circulant class, which commutes with $\boldsymbol{S}$. It is not the local start in Equation (361): its initial gradient has every cyclic diagonal nonzero, so the coordinate-support lower-bound argument does not apply. It also differs from $\boldsymbol{Q}_0 = \boldsymbol{0}$ in the stated `Muon` upper theorem. Appendix L.4 checks the zero-initialized operator envelope with an untuned schedule.

Algorithms and rate selection. We use the three buffer updates in Section 2.1.4 with $\mu \in \{0, 0.9\}$. The exact circulant spectral representation of `Muon` agrees with the dense SVD partial polar; `SignGD` applies the entrywise sign in the original coordinates. The uniform bound $R = 3$ gives the untuned `Muon` schedule $\eta_t = 3(0.98 - \mu)0.98^t$. For `SignGD`, $\eta_0 = (3/16)\xi$, with either plateau decay by

0.3 after 50 consecutive nonimproving pre-update objectives or geometric decay $\eta_t = \eta_0(0.999)^t$. GD uses constant $\eta = ((1+\mu)/\kappa)\xi$, accounting for the exact smoothness κ.

We search $\xi = 2^{j/2}$, $j = -28, \dots, 8$, on seeds $1000, 1001$. Rates are selected separately for first hits of operator error 10^{-6} and relative objective 10^{-6}. Candidates are ranked by numerical failures, unreached seeds, median and mean hitting times, final metric, then the smaller multiplier. The fixed protocol transfers the $\kappa = 5$ choice across κ; the tuned protocol selects separately at each κ. Both use held-out seeds $0, \dots, 9$ for evaluation. Muon is unchanged between protocols. No selected rate lies on a search boundary. Fixed geometric SignGD applies the same entire step sequence across κ and seeds; plateau events can depend on the trajectory.

Horizon, metrics, and outcomes. Every nondivergent run records 12,000 updates without early stopping. The complete archive retains divergent tuning trials; there are no numerical failures in the evaluations reported here. First hits at $10^{-4}, 10^{-6}, 10^{-8}$, subsequent exits, and final-100-state confirmations are saved for operator error and relative objective, together with objectives, Frobenius errors, rates, and decay events. Unreached times are censored. Medians and percentile bands are reported only when their required order statistics are observed; means and standard deviations require all ten times. The numerical checks verify gradients, norms, the dense and reduced optimizer recurrences, and every threshold crossing. Figures read the saved arrays without rerunning optimization.

Condition number, tuning, and the error metric. Figure 5 compares the two rate-selection protocols at $\mu = 0.9$. For operator error 10^{-6}, Muon's median is 680 throughout the grid. At $\kappa = 15625$, tuned plateau SignGD reaches on none of ten seeds, while tuned Momentum GD takes a median 8,223 iterations. Thus both baselines deteriorate at large κ even after tuning. On the smaller grid $\kappa \le 625$, SignGD's counts remain comparable and nonmonotone, while tuned GD is faster than Muon. In relative objective, tuned GD remains competitive throughout: its median count at $\kappa = 15625$ is 310, versus 309 for Muon. Relative objective weights spectral directions differently and divides by a κ-dependent initial objective, consistent with the spectrum-dependent conversion in Equation (282). We therefore keep the two metrics separate and infer no asymptotic exponent from this grid.

The geometric SignGD control applies an identical step sequence across κ in the fixed protocol. At $\mu = 0.9$, no seed reaches operator error 10^{-6} at any tested κ under that protocol. Some separately tuned geometric runs have transient hits, and their relative-objective counts can decrease with κ (Figure 5(d)). Thus the observed separation depends on the schedule and error metric; a first hit need not imply sustained accuracy.

L.3 Search rank and overparameterization

This study examines the undercomplete and wide regimes in Theorem 1. We fix $d = 100$, $r = 2$, and $\alpha = 0.1$, and vary $k \in \{2, 3, 5, 10, 100\}$ at $\kappa \in \{1, 125, 625\}$ for $\mu \in \{0, 0.9, 0.95\}$.

Instances and protocol. The positive target eigenvalues are 1 and $1/\kappa$. We obtain the target eigenbasis and an independent initialization frame by QR-orthonormalizing Gaussian matrices, and set $\boldsymbol{U}_0 = 0.1\boldsymbol{O}$. The columns of $\boldsymbol{O}$ are orthonormal for $k < d$; $\boldsymbol{O}$ is orthogonal for $k = d$. We use the exact Muon recursion with zero initial buffer and $\eta_t = 2(1-\rho)\xi\rho^t$, $\rho = \max\{0.98, (1+\mu^{1/3})/2\}$. For fixed rates, seeds $2000, 2001$ on a $d = 20$, $\kappa = 5$, $k = 2$ pilot select $\xi = 1, 2^{3/2}, 2^{-1/2}$ for $\mu = 0, 0.9, 0.95$, respectively; these multipliers are fixed across all evaluation ranks and condition numbers. The tuned protocol selects separately for each (k, κ, μ) at $d = 100$ using seeds $1000, 1001$. Both search $\xi = 2^{j/2}$ for $j = -12, \dots, 8$, extending any selected boundary once by four half-octave points before evaluation. No final selection is on a search boundary. Candidates are ranked by numerical failures, unreached tuning seeds, median and mean first-hit times at 10^{-6}, median final error, then the smaller multiplier. Evaluation uses paired seeds $0, \dots, 9$.

The budget is 12,000 updates. Early stopping requires at least 1,000 updates and 100 consecutive states with reconstruction error at most 10^{-8}. Nonfinite arithmetic or operator error above 10^6 is a numerical failure; finite unreached times are censored. We report first hits of reconstruction error

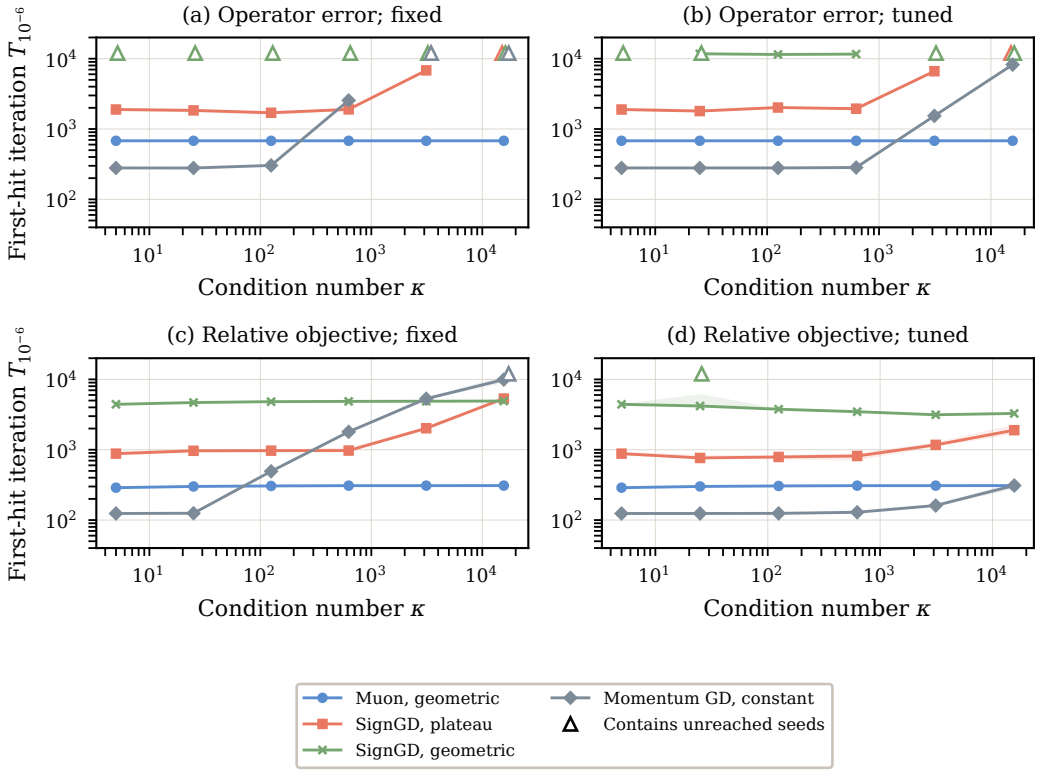


Figure 5: Nonlocal random-start ICL, $d = 4096$, $\mu = 0.9$. Top: first hit of $\|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\| \le 10^{-6}$; bottom: first hit of $f(\boldsymbol{Q}_t)/f(\boldsymbol{Q}_0) \le 10^{-6}$. Left: fixed normalized rates; right: per-κ rates selected on two separate seeds, independently for each metric. `Muon` uses the same untuned schedule throughout. Curves show medians and identifiable 20th–80th percentiles over ten evaluation seeds. Open triangles flag groups containing unreached seeds; missing medians are not connected. The budget is 12,000 updates.

10^{-6}, with medians and 20th–80th percentiles only when their order statistics are observed. First hits are distinct from the eventual-accuracy guarantee in the theorem.

Results and scope. With fixed geometric rates at $\kappa = 625$, the median counts across these five ranks are $505.5, 527, 654.5, 561.5, 588$ for $\mu = 0.9$ and $1{,}021, 1{,}014.5, 979.5, 949, 997$ for $\mu = 0.95$, with $10/10$ successes in each group. Thus the qualitative positive-momentum behavior persists across exact parameterization, undercomplete overparameterization, and the wide regime.

The tuned protocol again exposes transfer failures. At $\kappa = 625$, geometric $\mu = 0$ reaches on $4/10$ seeds for $k = 3$ and $7/10$ for $k = 5$. At $\kappa = 125$, geometric $\mu = 0.9$ reaches on $8/10$ for $k = 10$. These outcomes are reported in Figure 6. The common $\alpha = 0.1$ is not selected for the accuracy-dependent initialization requirement of the undercomplete theorem.

L.4 An untuned check of the ICL operator envelope

We check the operator envelope underlying Theorem 3 using the deterministic $b = 1$, $C_\eta = 1$ specialization in Appendix C.1. Here $d = 100$, $\boldsymbol{B}_{-1} = \boldsymbol{Q}_0 = \boldsymbol{0}$, and the covariance has a Gaussian-QR orthonormal eigenbasis with eigenvalues $\lambda_i(\boldsymbol{S}) = \kappa^{(i-1)/(3(d-1))}$, $1 \le i \le d$. Thus $\|\boldsymbol{S}^{-1}\| = 1$; at $\kappa = 1$, we set $\boldsymbol{S} = \boldsymbol{I}$ exactly. The untuned schedule is

$$\eta_t = (\rho - \mu)\rho^t, \qquad \rho = \max\{0.98, (1+\mu)/2\}, \tag{413}$$

The control was specified before held-out evaluation. It covers $\kappa \in \{1, 2, 5, 10, 25, 50, 125, 250, 625, 1250\}$ at $\mu = 0, 0.9, 0.95$; $\kappa \in \{1, 5, 25, 125, 625\}$ at

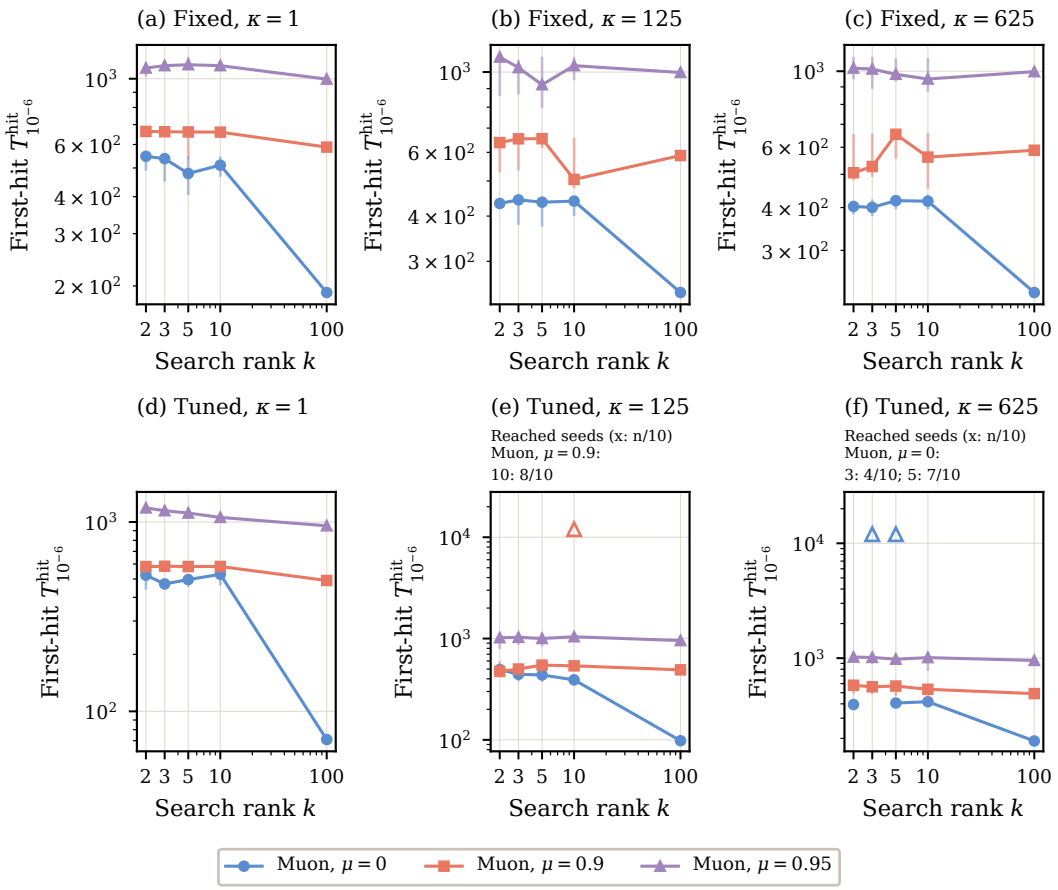


Figure 6: Matrix-factorization search-rank study under geometric decay: fixed normalized rates (top) and rates selected on separate tuning seeds (bottom). Columns correspond to $\kappa = 1, 125, 625$; $d = 100$, $r = 2$, and $\alpha = 0.1$ throughout. Each point summarizes ten evaluation seeds at operator accuracy 10^{-6}. Whiskers give the 20th–80th percentiles when identifiable. Open triangles and success-count annotations retain the censored outcomes in the tuned protocol.

$\mu = 0.5, 0.99$; and $\kappa \in \{1, 125, 625\}$ at $\mu = 0.8, 0.98$. Each of these 46 settings uses seeds $0, \ldots, 9$, with the iteration budget, stopping rule, and first-hit statistic of Appendix L.3, applied to ICL operator error. Trajectory panels use separate 1,500-update replays without early stopping. The exact spectral implementation is checked against dense SVD updates.

The observed operator errors satisfy the envelope $e_t^{\text{ICL}} \le \rho^t$ up to an absolute roundoff discrepancy of 1.72×10^{-15}. For $\mu = 0.9$, median first-hit counts are 610–611 at $\kappa = 5, 25, 125, 625$; for $\mu = 0.99$, the count is 2,757 at every condition number on the five-point grid.

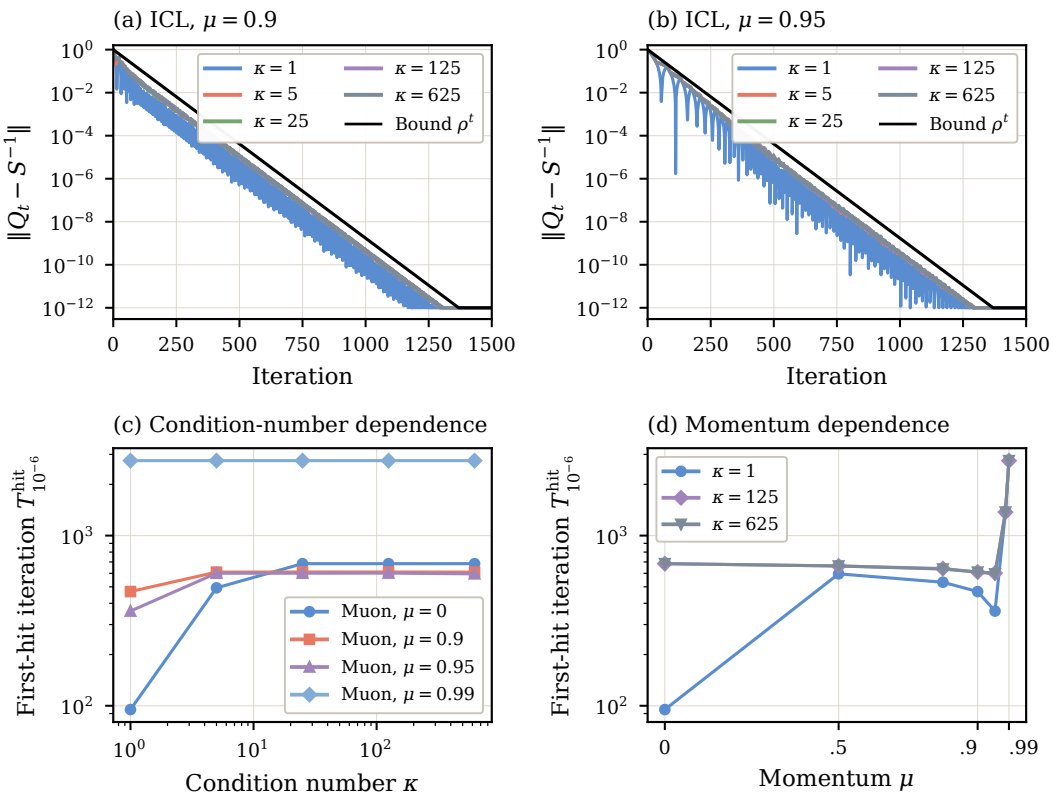


Figure 7: Untuned ICL control with $d = 100$, $\boldsymbol{Q}_0 = \boldsymbol{0}$, and the exact schedule in Equation (413). Top: $\|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\|$ trajectories and the theorem envelope ρ^t at $\mu = 0.9$ and 0.95. Bottom: first-hit iteration counts versus condition number and momentum. Each setting uses ten seeds; rotation invariance makes the spectral trajectories coincide across seeds. All shown groups reach 10^{-6}. This control uses the deterministic $b = 1$, $C_\eta = 1$ specialization, independently of the pilot and tuned protocols.

L.5 Finite Newton–Schulz orthogonalization

We examine whether the weak condition-number dependence of exact-polar `Muon` survives finite practical orthogonalization. We use the quintic Newton–Schulz (NS) map of Jordan et al. (2024): starting from $\boldsymbol{X} = \boldsymbol{G}/(\|\boldsymbol{G}\|_{\mathrm{F}} + 10^{-7})$, repeat

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{X}^\top, \qquad \boldsymbol{X} \leftarrow 3.4445\boldsymbol{X} - 4.7750\boldsymbol{A}\boldsymbol{X} + 2.0315\boldsymbol{A}^2\boldsymbol{X}.$$

The zero buffer returns zero. We use dense float64 arithmetic to isolate truncation from low-precision arithmetic, and retain the full-batch buffer $\boldsymbol{B}_t = \mu\boldsymbol{B}_{t-1} + \nabla f(\boldsymbol{Q}_t)$. Only the direction changes to $\mathrm{NS}_s(\boldsymbol{B}_t)$; the exact comparator uses the same SVD partial-polar convention as Appendix L.4. The finite map is approximate orthogonalization, not an asymptotically exact polar iteration.

We reuse that appendix's covariance construction, $d = 100$, $\boldsymbol{B}_{-1} = \boldsymbol{Q}_0 = \boldsymbol{0}$, and $\|\boldsymbol{S}^{-1}\| = 1$, with $\kappa \in \{1, 5, 25, 125, 625, 3125, 15625\}$ and $\mu = 0.9$. We compare $s \in \{1, 2, 3, 5, 10\}$ against exact partial polar. Protocol A keeps Equation (413) unchanged. Protocol B multiplies it by one ξ_s, selected only at $\kappa = 5$ on seeds 1000, 1001 and then frozen across κ. Searching $\xi = 2^{j/2}$, $j = -12, \ldots, 8$, with the failure, censoring, first-hit, final-error, and tie-breaking rules of Appendix L.3, selects $(\xi_1, \xi_2, \xi_3, \xi_5, \xi_{10}) = (2^{-1/2}, 2^{-3/2}, 1/4, 2^{-3/2}, 2^{-3/2})$; none is on a search boundary. Exact polar remains untuned. Each evaluation uses seeds $0, \ldots, 9$ and saves all 12,000 updates without early stopping. The primary statistic T_{hit} is the first hit of $e_t = \|\boldsymbol{Q}_t - \boldsymbol{S}^{-1}\| \le 10^{-6}$; unreached times remain censored. Table 2 complements these first hits with sustained accuracy and final error from the same saved trajectories.

In Figure 8, five and ten NS steps reach every tested condition number under both protocols. For $\kappa \ge 5$, the untuned five-step medians are 603–621, compared with 608–631 for dense exact polar.

One and two untuned steps miss the tolerance at $\kappa = 3125, 15625$; three steps miss at 15625. Ten steps are slightly slower than five throughout the grid, so increasing the inner count is not uniformly beneficial. A repetition at $\mu = 0.95$ on seeds $0, 1$ also reaches the whole grid with five and ten steps (both select $\xi_s = 1$), although the ten-step seed-1 runs at $\kappa = 3125, 15625$ finish at 4.91×10^{-7} and 1.34×10^{-7}, respectively, and never reach the stricter tolerance 10^{-8}. Both remain below 10^{-6} throughout the last 1,000 saved states. No pilot or evaluation run has a numerical failure.

Calibration lowers the five-step medians to 552–569 for $\kappa \geq 5$, but does not remove the small-s failures. It also introduces failures to reach on easier instances: calibrated NS-1 and NS-3 miss at $\kappa = 1$, NS-2 misses already at 625, and NS-3 reaches only the reference $\kappa = 5$. The saved direction spectra show distortion beyond a common scale factor: at $\kappa = 15625$, the initial minimum-to-maximum singular-value ratio is 0.0019 for NS-1 and 0.2233 for NS-5. Nevertheless, the same-buffer discrepancy $\| \mathrm{NS}_5(\boldsymbol{B}_t) - \mathsf{msign}(\boldsymbol{B}_t) \|$ can be near one even on successful runs. Similar outer counts therefore do not imply uniformly accurate polar directions.

Sustained accuracy and final error. For $H = 12{,}000$ and $\tau = 10^{-6}$, define

$$T_{\mathrm{stay}}(\tau; H) = \min\{t \in \{0, \ldots, H\} : e_j \leq \tau \text{ for every } j = t, \ldots, H\}.$$

An empty set is censored. This measures persistence through the recorded horizon, not an asymptotic guarantee. At $\mu = 0.9$, every run that first reaches τ also attains this sustained criterion and stays below τ for the final 1,000 states ($t = 11{,}001, \ldots, 12{,}000$). However, first hits can precede sustained accuracy by up to 194 updates: for calibrated NS-5 at $\kappa = 1$, the first hit is 297 and $T_{\mathrm{stay}} = 491$ on every rotation. Table 2 retains these delays and the largest final error over every tested condition number and rotation, including unreached runs. Five and ten steps finish below 1.30×10^{-15} in all primary runs under both protocols; the unreached small-s runs retain large final errors.

Checks verify the gradients, dense NS recurrence, singular-value polynomial, and saved threshold crossings. Single-map orthogonal equivariance holds to numerical precision, but long dense trajectories are sensitive to roundoff across rotations. The plotted medians summarize this numerical sensitivity, not independent statistical variation. The exact baseline reproduces the existing dense code; its first hits can differ from the reduced implementation in Appendix L.4 (at $\kappa = 625$, 620–624 versus 610). These experiments test robustness of the exact-polar phenomenon to truncated practical orthogonalization on a controlled quadratic family; they do not provide a finite-Newton–Schulz counterpart of Theorem 3.

Table 2: First-hit and sustained accuracy at $\tau = 10^{-6}$, and final operator error at $H = 12{,}000$, for the finite-NS study with $\mu = 0.9$. Each pair $(n_{\mathrm{hit}}, n_{\mathrm{stay}})$ gives two counts, each out of 70 runs (seven condition numbers and ten rotations). $\Delta T_{\max}$ is the largest $T_{\mathrm{stay}} - T_{\mathrm{hit}}$ among successful runs; $e_H^{\max}$ is the largest final error over all 70 runs, including those that never reach the tolerance. The exact-polar entries reuse the same trajectories in both protocols.

	Protocol A: untuned			Protocol B: calibrated once		
Method	$(n_{\mathrm{hit}}, n_{\mathrm{stay}})$	$\Delta T_{\max}$	$e_H^{\max}$	$(n_{\mathrm{hit}}, n_{\mathrm{stay}})$	$\Delta T_{\max}$	$e_H^{\max}$
Exact polar	$(70, 70)$	126	7.81×10^{-17}	$(70, 70)$	126	7.81×10^{-17}
NS-1	$(50, 50)$	145	7.16×10^{-1}	$(40, 40)$	98	7.43×10^{-1}
NS-2	$(50, 50)$	64	6.27×10^{-1}	$(40, 40)$	84	6.80×10^{-1}
NS-3	$(60, 60)$	92	2.35×10^{-1}	$(10, 10)$	35	4.84×10^{-1}
NS-5	$(70, 70)$	87	1.24×10^{-15}	$(70, 70)$	194	1.29×10^{-15}
NS-10	$(70, 70)$	95	2.04×10^{-16}	$(70, 70)$	124	2.18×10^{-16}

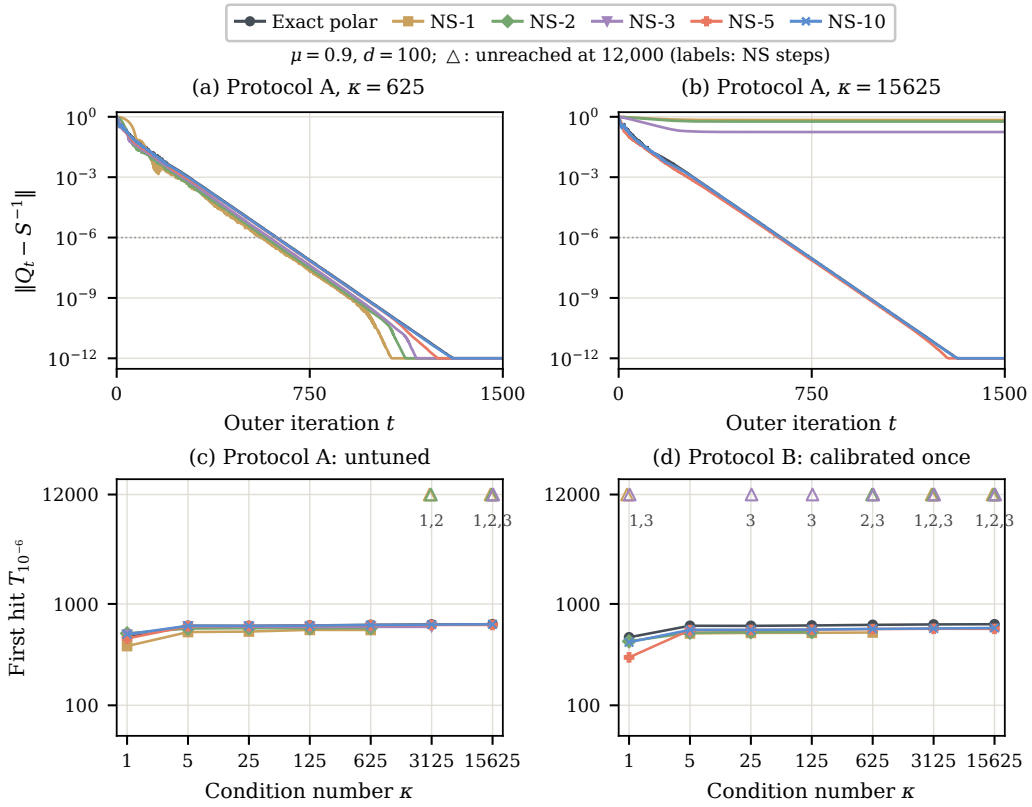


Figure 8: Finite quintic NS orthogonalization in the zero-initialized ICL control, $d = 100$, $\mu = 0.9$. (a,b) Operator-error trajectories under the untuned Protocol A at $\kappa = 625, 15625$, shown through 1,500 updates. (c,d) First hits of $e_t \le 10^{-6}$ under Protocol A and the once-calibrated Protocol B. Curves show medians across ten rotations, used as numerical sensitivity checks. Open triangles flag settings where all ten runs are unreached at 12,000 updates; nearby numbers give their NS step counts. Censored medians are not connected. Table 2 reports sustained accuracy and final errors. All full-horizon histories and direction diagnostics are saved.

L.6 Actual Adam on the matched local factorization instance

This experiment tests the local mechanism of Proposition 6 on its common instance and initialization, while varying the scale that controls the quadratic approximation. We use Equation (380): $\boldsymbol{M}^\star = \boldsymbol{A}/2$, $\boldsymbol{S} = (\boldsymbol{A}/2)^{1/2}$, and $\boldsymbol{U}_0 = \boldsymbol{S} + \delta \boldsymbol{A}^{-1}$, for $\kappa \in \{25, 625, 15625\}$. With the proposition's accuracy parameter $\varepsilon = 10^{-2}$, Equation (379) gives $d = 32, 140, 668$, respectively. We sweep $\delta \in \{10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}, 10^{-16}\}$; each run is deterministic and saves all 2,000 updates.

Algorithms and stable evaluation. Actual full-batch `Adam` follows Equation (311), with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_{\mathrm{A}} = 10^{-8}$, zero initial moments, both bias corrections, and no weight decay. Its constant scalar rate is $\gamma_t = \epsilon_{\mathrm{A}}/\kappa$, fixed across δ; the resulting quadratic reference has effective rate $1/\kappa$. This rule is chosen to keep the reference stable, without rate tuning. Exact-polar `Muon` uses $\mu = 0.9$, zero initial buffer, and the proposition's schedule $\eta_t = 0.02\delta(0.98)^t$. We also run the quadratic momentum reference in Equation (393) with the same Adam rates and first-moment bias correction.

To resolve tiny initial errors, we evolve $\boldsymbol{E} = \boldsymbol{U} - \boldsymbol{S}$, evaluating the exact commuting identities

$$\nabla f(\boldsymbol{S} + \boldsymbol{E}) = \boldsymbol{A}\boldsymbol{E} + 3\boldsymbol{S}\boldsymbol{E}^2 + \boldsymbol{E}^3, \qquad f(\boldsymbol{S} + \boldsymbol{E}) = \frac{1}{4}\|2\boldsymbol{S}\boldsymbol{E} + \boldsymbol{E}^2\|_{\mathrm{F}}^2. \tag{414}$$

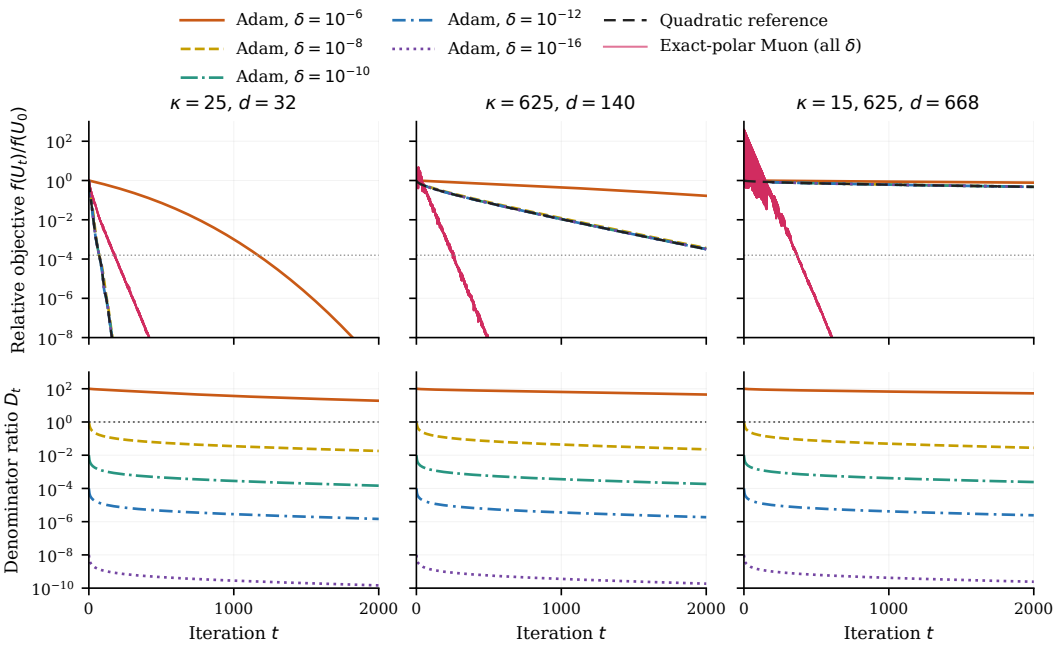


Figure 9: Actual full-batch `Adam` and exact-polar `Muon` on the common local factorization instance of Proposition 6. Columns vary κ; colors and line styles vary δ. Top: relative objective, with the quadratic momentum reference dashed black and all five `Muon` runs in pink (visually coincident). The dotted line marks $\varepsilon/64 = 1.5625 \times 10^{-4}$; values below 10^{-8} lie outside the displayed range. Bottom: Adam's D_t, with a dotted line at one. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_{\mathrm{A}} = 10^{-8}$, and fixed $\gamma = \epsilon_{\mathrm{A}}/\kappa$; `Muon` uses $\mu = 0.9$ and $\eta_t = 0.02\delta(0.98)^t$. These are deterministic, untuned runs on the same initialization for each (κ, δ).

In implementation, the normalized states $\boldsymbol{X} = \boldsymbol{E}/\delta$, $\boldsymbol{M}/\delta$, and $\boldsymbol{V}/\delta^2$ obey the exact recursions in Equation (392). This avoids subtracting $\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star$ or adding a tiny error to $\boldsymbol{S}$. Fourier coordinates evaluate matrix products and the exact polar factor; `Muon` additionally rescales its error and buffer by ρ^t to resolve the tail of the geometric schedule. Adam's moments and division are entrywise in the original matrix coordinates, represented by one circulant column.

Diagnostics and scope. Besides relative objective $f(\boldsymbol{U}_t)/f(\boldsymbol{U}_0)$, we record

$$D_t = \max_p \frac{\sqrt{\widehat{V}_{t,p}}}{\epsilon_{\mathrm{A}}}, \qquad \Delta_\delta = \max_{0 \le t \le 2000} \frac{|F_\delta(\boldsymbol{X}_t) - \mathcal{Q}(\boldsymbol{Y}_t)|}{\mathcal{Q}(\boldsymbol{Y}_0)}. \tag{415}$$

Here p ranges over matrix entries, and D_t uses the bias-corrected second moment immediately before update t. Thus $D_t \ll 1$ certifies stabilizer dominance in every entry, whereas $D_t > 1$ identifies at least one entry whose root second moment exceeds the stabilizer. The error Δ_δ measures the finite-horizon approximation in Equation (394) along the chosen rate vector. This is a fixed-rate mechanism diagnostic, not a comparison of optimally tuned optimizers or an empirical estimate of the proposition's $\sqrt{\kappa}$ lower-bound exponent. The support-barrier proof applies on its stated horizon $N = 6, 29, 144$; the longer trajectories diagnose the observed dynamics.

The local limit and its boundary. In Figure 9, D_0 is approximately $\delta/\epsilon_{\mathrm{A}}$, ranging from 100 to 10^{-8}. For $\delta = 10^{-6}$, $D_t > 18$ throughout the displayed horizon at every κ, and the quadratic reference does not closely describe Adam's objective trajectory. For $\delta \le 10^{-10}$, $D_t \le 0.0101$ throughout, and the discrepancy decreases with δ. Table 3 quantifies this limit: the largest discrepancy across the three instances falls from 0.881 at $\delta = 10^{-6}$ to 1.92×10^{-9} at $\delta = 10^{-16}$. Thus even small-error initializations can lie outside the stabilizer-dominated regime; small distance to $\boldsymbol{S}$ alone does not justify replacing Adam by the quadratic reference.

Table 3: Maximum normalized objective discrepancy Δ_δ from Equation (415), over all 2,001 saved states. Adam parameters and rates are those of Figure 9.

δ	$\kappa = 25$	$\kappa = 625$	$\kappa = 15625$
10^{-6}	8.81×10^{-1}	6.35×10^{-1}	3.02×10^{-1}
10^{-8}	1.22×10^{-1}	3.45×10^{-2}	7.52×10^{-3}
10^{-10}	1.91×10^{-3}	4.43×10^{-4}	9.01×10^{-5}
10^{-12}	1.92×10^{-5}	4.45×10^{-6}	9.04×10^{-7}
10^{-16}	1.92×10^{-9}	4.45×10^{-10}	9.04×10^{-11}

The comparison is not a uniform win for `Muon`. At $\kappa = 25$, Adam reaches $f(\boldsymbol{U}_t)/f(\boldsymbol{U}_0) \le \varepsilon/64$ in 72 steps for $\delta \le 10^{-10}$, 73 steps for $\delta = 10^{-8}$, and 1,168 steps for $\delta = 10^{-6}$; `Muon` takes 178 steps at every scale. At $\kappa = 625, 15625$, `Muon` takes $251, 367$ steps, respectively, whereas neither Adam nor its quadratic reference reaches the target within 2,000 steps under this rate rule. Muon is also nonmonotone in relative objective: its transient maximum is about 4.70 at $\kappa = 625$ and 387 at $\kappa = 15625$ before contraction. The experiment therefore supports the stabilizer-dominated mechanism and exhibits its limits, without extending the proposition to second-moment-dominated Adam or to tuned practical performance. Dense checks verify the stable objective and gradient, both Adam bias corrections, and agreement of the spectral polar update with an SVD.