Towards Understanding the Power and Limits of the Muon Optimizer: A River-Valley Perspective

Tianqi Shen[1] **Runze Shi**[1] **Jinji Yang**[1] **Jianhao Ma**[2] **Jiaye Teng**[3] **Ziye Ma**[1]
[1]Department of Computer Science, City University of Hong Kong
[2]Department of Industrial Engineering, Tsinghua University
[3]School of Statistics and Data Science, Shanghai University of Finance and Economics
tianqshen5-c@my.cityu.edu.hk runzeshi3-c@my.cityu.edu.hk
jinji.yang@cityu.edu.hk jianhao@tsinghua.edu.cn
tengjiaye@sufe.edu.cn ziyema@cityu.edu.hk

Abstract

Momentum Orthogonalized by Newton–Schulz (Muon) has emerged as a promising optimizer for large language model (LLM) training, yet when and why it outperforms established alternatives such as AdamW remain insufficiently understood. This gap matters in practice, as a sub-optimal choice of optimizer or even hyper-parameters can slow progress throughout expensive pre-training and waste substantial computational resources. Existing theories largely adopt a local-improvement perspective, emphasizing singular-value flattening or implicit curvature adaptation, but they neither determine which optimizer is preferable nor explain why Muon is typically fast only during early training. To better explain Muon's complicated behavior over the entire training trajectory, we develop a trajectory-level characterization of Muon dynamics based on a new mixed-spectrum matrix sensing (MSMS) model motivated by the river–valley view of LLM loss landscapes. Under simplifying assumptions, our continuous-time analysis reveals a qualitatively different convergence mechanism: Muon advances piecewise linearly toward the solution, whereas gradient descent and Adam-like flows approach it asymptotically along curved paths. With finite step sizes, however, the same normalized update magnitude makes Muon prone to overshooting near the solution, so its admissible step size shrinks with the residual scale. This links Muon's early advantage and late-stage limitations to a common mechanism and explains its sensitivity to learning-rate scheduling. Language model pre-training experiments support our discovery and suggest that switching from Muon to AdamW can preserve rapid initial progress while improving late-stage refinement. Code: `https://github.com/MOFA-LAB/River-Valley-Muon`.

1 Introduction

Non-Euclidean gradient-based optimizers [Crawshaw et al., 2026], e.g. Momentum Orthogonalized by Newton–Schulz (Muon) [Jordan et al., 2024], are increasingly used for training large language models (LLMs), as documented in recent technical reports. Some notable examples include Kimi K2 [Team et al., 2025], GLM-5 [Zeng et al., 2026], DeepSeek-V4 [Xu et al., 2026], and Qwen3.8-Flash [Qiu et al., 2026]. In parallel, substantial system-level engineering efforts have been devoted to optimizing Muon across modern software and hardware stacks [Khona et al., 2026]. Despite this growing adoption, it remains unclear under what conditions these methods could outperform the traditional Euclidean optimizers [Bernstein and Newhouse, 2024] that have long dominated large-scale training, such as AdamW [Loshchilov and Hutter, 2017]. Given the enormous cost of LLM pretraining [Cottier et al., 2024], even modest differences in optimizer efficiency can lead to substantial

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

financial and environmental consequences [Schlotthauer et al., 2025, Strubell et al., 2019]. Without guidance on when Muon or AdamW is more efficient, practitioners risk committing substantial resources to a suboptimal optimizer [van der Schelling et al., 2026].

Recent studies [Qi et al., 2026, Dragutinović et al., 2026, Shumaylov et al., 2026, Wang et al., 2026a] have begun to address this question, but a definitive picture has yet to emerge. A prevailing explanation attributes Muon's advantage to its ability to flatten the singular-value spectrum of the momentum or gradient matrix [Boreiko et al., 2025, Nguyen et al., 2026, Wang et al., 2026a, Beneventano et al., 2026], while another line of work interprets Muon as a pseudo-second-order method that implicitly exploits Hessian information [Shah et al., 2025, Si et al., 2025, Du and Su, 2026, Che, 2026]. Together, they suggest that Muon can mitigate the effects of heterogeneous curvature and rapidly reach a reasonably good solution. If these explanations were sufficient, however, Muon's advantages should persist throughout optimization, leading to consistently faster convergence. Regrettably, existing empirical and theoretical results do not yet support such a conclusion. Most theoretical studies only establish a faster convergence or progress in the early stage, and typically for toy models that have a drastically different landscape to LLMs, such as Gaussian-mixture model [Vasudeva et al., 2026], isotropic-curvature model [Su, 2025] and associative-memory models [Wang et al., 2026b,c, Kim et al., 2026a]. Moreover, both empirical evidence [Liu et al., 2025] and recent theoretical studies indicate that Muon may even suffer from poor late-stage convergence: it can exhibit suboptimal convergence rates or even fail to converge under certain optimization regimes [Chang et al., 2025, Parshakova et al., 2026, Do et al., 2026].

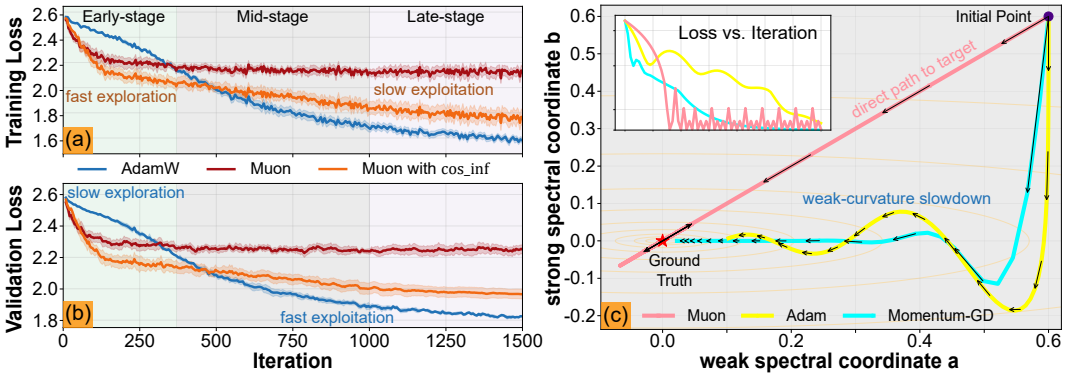

Figure 1: Key observations comparing Non-Euclidean & Euclidean gradient-based optimizers. **(a)** and **(b)** report the losses for pre-training NanoChat on the Shakespeare dataset. The shaded bands denote the spread of losses obtained under different learning rates, reflecting each optimizer's sensitivity to the scheduler. **(c)** illustrates our proposed mixed-spectrum matrix sensing testbed in a simplified two-dimensional setting, designed to capture salient features of LLM training dynamics.

These existing results indicate a non-uniform behavior of Muon convergence throughout the pre-training phase, which we summarize as a "fast-then-slow" dynamic. This means that Muon usually makes rapid early progress but slows down during the later exploitation stage. To showcase this dynamic more clearly, we have pre-trained NanoChat [Karpathy, 2025] on the Tiny Shakespeare dataset [Karpathy, 2015] with both Muon and AdamW, plotted in Figure 1 (a) and (b). As our plot indicates, with a constant learning rate (LR), Muon rapidly reduces the loss in the first 500 iterations, but subsequently enters a slow exploitation phase and plateaus at a relatively high loss. An LR-decay scheduler alleviates this late-stage stagnation, but compromises Muon's fast exploration in the early stage. This fast-then-slow dynamic cannot be explained by existing theoretical frameworks discussed above which focus solely on spectral flattening or local second-order improvements.

To address this critical gap, we instead provide a trajectory-level characterization of Muon convergence that captures its time-varying optimization behavior. As illustrated in Figure 1(c), our analysis shows that, in the mixed-spectrum matrix sensing (MSMS) setting, Muon makes approximately linear progress toward the ground truth. In contrast, the convergence rates of Adam [Kingma and Ba, 2014] and gradient descent (GD) with momentum [Ma et al., 2026] gradually decrease as their iterates approach the ground truth, with their early-stage trajectories typically following gradient-dominated directions. As another key contribution, we introduce MSMS as a new mathematical model designed to better capture the heterogeneous spectral structure observed in the optimization

landscapes of real-world LLM training. To make the resulting high-dimensional dynamics analytically tractable, we further construct an invariant manifold that reduces the original training dynamics to a three-dimensional system of ordinary differential equations (ODEs).

This leads to the following insights. When step-sizes are allowed to be infinitely small (i.e., gradient flow setting), the speed of Muon's progress need not diminish as its trajectory approaches a critical point, whereas the speed of GD/Adam-like flows typically decreases as their distance to the critical point shrinks. This distinction explains Muon's advantage in rapid early-stage exploration. Once finite step sizes are reintroduced through discretization, however, Figure 1's (c) shows that Muon can overshoot the target solution (i.e., GT), while GD/Adam-like refinement can remain stable. This provides an explanation for Muon's difficulty in late-stage exploitation: avoiding overshooting requires meticulously tuning the LR scheduler, which becomes particularly difficult as the trajectory approaches the target; an overly small late-stage LR, in turn, can slow its convergence. These conclusions are not limited to the MSMS setting: we further establish the same insights in a more general setting in Section 5 and corroborate them with experiments on LLMs in Section 6.

2 Related Work

Current understanding of Muon. Much of the existing understanding of Muon is derived from empirical observations [Liu et al., 2025] or stylized case studies that do not fully capture the characteristics of LLM training. Su [2025] studies Muon through an isotropic-curvature model, showing that spectral updates improve conditioning by homogenizing singular values of gradient matrix. Vasudeva et al. [2026] demonstrate that, in a Gaussian-mixture data model, Muon implicitly learns all principal components at comparable rates, thereby promoting a more balanced recovery of the underlying data components. Wang et al. [2026b] analyze associative-memory learning under heavy-tailed data and prove that Muon yields more balanced class-wise learning than Adam through a more isotropic update spectrum. Following them, Kim et al. [2026a] further show, in linear associative-memory models, that Muon achieves a substantially faster initial recovery rate than stochastic gradient descent (SGD) and admits a larger storage capacity. Other studies [Tveit et al., 2025, Gronich and Vardi, 2026, Gonon et al., 2026] rely on toy Transformers, smooth homogeneous neural networks, or even simple quadratic objectives. Overall, these works primarily explain Muon's advantage in fast early-stage exploration [Wen et al., 2026]; by contrast, Muon's poor late-stage exploitation remains less well understood. For instance, Chang et al. [2025] show that Muon variants achieve only a suboptimal ergodic convergence rate in stochastic non-convex optimization and Do et al. [2026] prove non-convergence in asymptotic stochastic optimization. Parshakova et al. [2026] find that Muon can even fail to converge on convex Lipschitz functions regardless of the LR schedule. To narrow this gap, we provide a trajectory-level analysis that tracks Muon's behavior across both the early-stage exploratory and late-stage exploitative phases to understand its strengths and weaknesses.

LLM landscape geometry. A pervasive observation in LLM training is strong anisotropy. Godey et al. [2024] show that anisotropic directions in Transformers can characterize inductive preferences toward spiked features. More directly, Zhu et al. [2026], Huang et al. [2026], Bernas et al. [2026] suggest that LLM loss landscapes and gradient spectra are highly anisotropic: loss reduction is often dominated by progress along a small number of spectral directions, while useful information is distributed across a long tail of weaker directions. Similar anisotropic phenomena [Weickert et al., 1998, Boscaini et al., 2016, Yu et al., 2026, Choi et al., 2026] also arise in vision models. This anisotropy motivates the river–valley picture, which decomposes the landscape [Davis et al., 2025] into slowly varying directions and sharply curved transverse directions. Wen et al. [2025] introduce this geometry to explain warmup–stable–decay (WSD) schedules: optimization progresses along a flat "river" while oscillating across steep valleys, and LR decay suppresses these oscillations to reveal the accumulated progress. Empirical landscape visualizations further support this interpretation, both during cooldown [Dremov et al., 2025] and with multi-power-law LR schedules [Luo et al., 2025]. Related analyses [Lu et al., 2026] also explain how momentum can stabilize optimization in such anisotropic geometries. Concurrent work by Kim et al. [2026b] likewise adopts this geometric picture to interpret Muon's rapid progress and oscillatory behavior in LLM pretraining. Although the river–valley assumption is only an approximate mechanistic model, it has become a useful explanatory framework for several phenomena in modern LLM training. We thus develop an MSMS model that captures the anisotropy of LLM landscapes and whose dynamics reduce to ODEs with natural river–valley decomposition on invariant manifolds, enabling a detailed characterization.

3 Preliminaries

Muon Optimizer. Unlike GD/Adam-like methods, which update parameters using the raw gradient or its coordinate-wise rescaling, Muon replaces each matrix-valued gradient or momentum with its orthogonalized counterpart. This operation preserves the singular directions of the gradient or momentum while discarding its singular values. Consequently, Muon not only flattens the singular-value spectrum of the update but also normalizes its overall magnitude. Concretely, given a matrix parameter X and an objective function $h(X)$, Muon performs the update[1]

$$\begin{aligned} M_t^{\mathrm{Muon}} &= \nabla h\big(X_t^{\mathrm{Muon}}\big) + \xi_t M_{t-1}^{\mathrm{Muon}}, \\ X_{t+1}^{\mathrm{Muon}} &= X_t^{\mathrm{Muon}} - \eta_t \,\mathtt{msign}\big(M_t^{\mathrm{Muon}}\big)\,, \end{aligned} \qquad \text{(Muon update)}$$

where ξ_t is the momentum coefficient and η_t is the learning rate. The operator $\mathtt{msign}(\cdot)$ returns the polar factor of its input matrix and is defined as

$$\mathtt{msign}(Y) := \arg\min_{Q} \left\{ \|Y - Q\|_F \;\middle|\; \text{either } QQ^\top = I \text{ or } Q^\top Q = I \right\}. \qquad (\mathtt{msign} \text{ operator})$$

In Section 4, we begin with a simplified version of Muon that removes momentum and implementation details, e.g. Newton–Schulz iterations. This simplified setting allows us to isolate the effect of gradient orthogonalization itself. In Section 5, we further study Muon with momentum in more general settings, with the additional preliminaries required for this extension collected in Appendix B.1.

4 Mixed-spectrum Matrix Sensing Model

Classical matrix sensing. Matrix sensing (MS) problem [Davenport et al., 2012] aims to recover a low-rank positive semidefinite (PSD) matrix $M^\star = X^\star {X^\star}^\top \in \mathbb{R}^{n\times n}$ with rank $r_\star < n$ from a set of linear measurements $\mathcal{A}$. Burer-Monteiro (BM) factorized MS replaces the rank constraint [Tillmann and Pfetsch, 2013] with a factor $X \in \mathbb{R}^{n\times r_{\mathrm{search}}}$ such that $XX^\top \approx M^\star$. Here, r_{search} is a tunable search-rank hyperparameter. The BM-factorized MS problem is then formulated as follows:

$$\min_{X\in\mathbb{R}^{n\times r_{\mathrm{search}}}} \frac{1}{2}\|\mathcal{A}(XX^\top) - b\|_2^2, \qquad b = \mathcal{A}(X^\star {X^\star}^\top). \qquad \text{(BM-factorized MS)}$$

Here, $b \in \mathbb{R}^m$ is the known measurement vector, and $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ (whose adjoint operator is denoted as $\mathcal{A}^*$) is a known operator defined as $\mathcal{A}(M) = [\langle A_1, M\rangle, \langle A_2, M\rangle, \cdots \langle A_m, M\rangle]^\top$, with symmetric sensing matrices A_i. $\langle A_i, M\rangle = \mathrm{Tr}(A_i^\top M)$ is the Frobenius inner product. This problem arises in numerous applications [Li et al., 2011, Candes et al., 2015, Rajani et al., 2014, Jin et al., 2019], and is equivalent to the training of two-layer quadratic neural networks [Li et al., 2018].

4.1 Problem Setting of Mixed-spectrum MS

Empirical observations. The optimization geometry of BM-factorized MS is determined by the sensing distribution: each sample corresponds to an A_i, and the squared gradient is proportional to $\mathcal{A}^*\mathcal{A}(XX^\top - M^\star)$. We therefore design the sensing matrices in MSMS to be anisotropic, so that they can induce anisotropic gradient and Hessian geometries. This choice is motivated by empirical evidence from both vision and language models. Figure 2 illustrates the covariance spectra of pre-classifier feature maps from fine-tuned vision models. In panels (a) and (b), the Marchenko–Pastur upper edge [Marčenko and Pastur, 1967] is used as a reference threshold under the isotropic assumption: $\lambda_+ = (1 + \sqrt{d/n})^2$, where d and n denote the number of channels and pixels, respectively. Most eigenvalues (gray) lie below λ_+, forming an isotropic bulk or background noise, while only a small fraction lie above it. Among these larger eigenvalues, panels (c) and (d) show that many (purple) have little association with label information, due to a low ANOVA coefficient R^2 [Neter et al., 1996], revealing nuisance spikes; only a small remainder (orange) corresponds to task-relevant signal. An analogous observation has been reported in LLM gradients, where energy concentrates in a compact, low-rank spike while weaker, context-specific semantic information spreads across a long spectral tail [Huang et al., 2026]. Together, these observations suggest that mixed-spectrum structure is common in deep learning and motivate its use as a tractable abstraction in MSMS.

[1] M with superscript and subscript denotes momentum instead of plain matrix used in the Section 4.

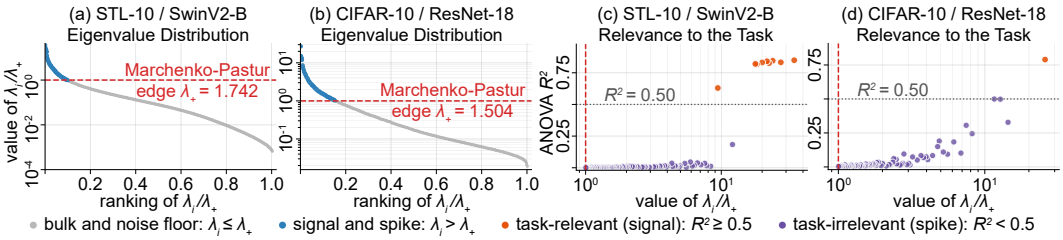

Figure 2: Empirical evidence for the mixed-spectrum covariance structure in image representations. Across datasets and architectures, the covariance spectrum of pre-classifier feature maps exhibits a pronounced bulk–signal–spike structure. Larger-scale settings with SwinV2-B [Liu et al., 2022] or STL-10 [Coates et al., 2011], produce more task-relevant eigendirections; smaller-scale settings with ResNet-18 [He et al., 2016] or CIFAR-10 [Krizhevsky and Hinton, 2009] exhibit fewer.

Mixed-spectrum MS model. To model such anisotropic structure in a tractable way, we introduce a mixed-spectrum MS model with choice $r_{\text{search}} = n$. We assume that the GT matrix is an orthogonal projector $P_\star = UU^\top$ with $U^\top U = I_r$. Let $\mathbb{R}^n$ admit the orthogonal decomposition

$$P_\star + P_\diamond + P_\circ = I_n, \qquad \dim(P_\star) = \dim(P_\diamond) = \dim(P_\circ) = r, \tag{1}$$

where $P_\diamond = VV^\top$ with $V^\top V = I_r$, and $P_\circ = WW^\top$ with $W^\top W = I_r$. The subspaces spanned by U, V, and W are mutually orthogonal with $3r = n = r_{\text{search}}$. We construct sensing matrix as

$$A = G + \pi_\star P_\star + \pi_\diamond P_\diamond + \pi_\circ P_\circ, \tag{2}$$

where $G \sim \mathrm{GOE}(n)$ is an isotropic Gaussian noise component, and $\pi := [\pi_\star, \pi_\diamond, \pi_\circ]^\top \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succeq 0$ being a covariance matrix. Thus, this sensing model consists of an isotropic Gaussian background together with three structured spectral components: $P_\star$, which is directly aligned with the target signal; $P_\diamond$, which represents dominant but task-irrelevant spiked directions; and $P_\circ$, which captures the remaining bulk or residual directions. This signal–spike–bulk decomposition is inspired by [Braun et al., 2026] but more closely aligned with the LLM observations [Huang et al., 2026].

Under this model, we study the corresponding population objective (see Lemma A.2)

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} \frac{1}{2}\mathbb{E}\left[\left\langle A, XX^\top - P_\star \right\rangle^2\right]. \tag{Mixed-spectrum MS}$$

Equivalently, the population loss can be calculated as follows if denoting $M := XX^\top \in \mathbb{R}^{n \times n}$:

$$\mathcal{F}(M) = \frac{1}{n}\left\|M - P_\star\right\|_F^2 + \frac{1}{2}\begin{bmatrix}\langle P_\star, M - P_\star\rangle \\ \langle P_\diamond, M\rangle \\ \langle P_\circ, M\rangle\end{bmatrix}^\top \Sigma \begin{bmatrix}\langle P_\star, M - P_\star\rangle \\ \langle P_\diamond, M\rangle \\ \langle P_\circ, M\rangle\end{bmatrix}. \tag{3}$$

Invariant manifold and reduced loss. Here, an *invariant manifold* means that if the optimization is initialized on the manifold, then the subsequent iterates remain on it. Specifically, for the variable $M = XX^\top$ and the factor form X, we define $\mathcal{M}$ for vanilla GD and $\mathcal{X}$ for simplified Muon:

$$\mathcal{M} := \{M = \alpha P_\star + \beta P_\diamond + \gamma P_\circ : \ \alpha, \beta, \gamma \in \mathbb{R}\}, \tag{Invariant manifold $\mathcal{M}$}$$

$$\mathcal{X} := \left\{X = \begin{bmatrix} aU & bV & cW \end{bmatrix} : \ a, b, c \in \mathbb{R}\right\}. \tag{Invariant manifold $\mathcal{X}$}$$

A key property of the population loss (3) is that these manifolds are invariant under the corresponding optimization dynamics. If vanilla GD is initialized in $\mathcal{M}$, then the entire trajectory remains in $\mathcal{M}$ (see Lemma A.3). Similarly, if Muon is initialized in $\mathcal{X}$, then its trajectory stays in $\mathcal{X}$ (see Lemma A.4). Moreover, $\mathcal{X}$ maps exactly into $\mathcal{M}$, since for any $X \in \mathcal{X}$, $XX^\top = a^2 P_\star + b^2 P_\diamond + c^2 P_\circ \in \mathcal{M}$. Studying the two invariant manifolds has three advantages. First, they reduce matrix-valued dynamics to a low-dimensional scalar system, substantially simplifying the analysis. Second, the invariant manifold is precisely $\mathrm{span}\{P_\star, P_\diamond, P_\circ\}$, which constitutes the entire subspace of interest. Third, common small initializations of the form ϵI [Stöger and Soltanolkotabi, 2021] lie on the manifold because $I = P_\star + P_\diamond + P_\circ$. The reduced scalar loss on $\mathcal{M}$ is

$$f(\alpha, \beta, \gamma) = \frac{1}{2}\begin{bmatrix}\alpha - 1 \\ \beta \\ \gamma\end{bmatrix}^\top \hat{K} \begin{bmatrix}\alpha - 1 \\ \beta \\ \gamma\end{bmatrix}, \qquad \hat{K} := \frac{2r}{n} I_3 + r^2 \Sigma \succeq 0. \tag{4}$$

Its minimizer is $(\alpha, \beta, \gamma) = (1, 0, 0)$, yielding a natural river–hill splitting in which α (signal) measures progress along the river direction, while β (spike) and γ (bulk) measure progress along the hill directions. On $\mathcal{X}$, the induced coordinates satisfy $\alpha = a^2, \beta = b^2, \gamma = c^2$; hence the reduced objective $f(a, b, c)$ is obtained by substituting them into $f(\alpha, \beta, \gamma)$. These scalar reductions later yield a convenient low-dimensional ODE system for comparing simplified Muon with vanilla GD.

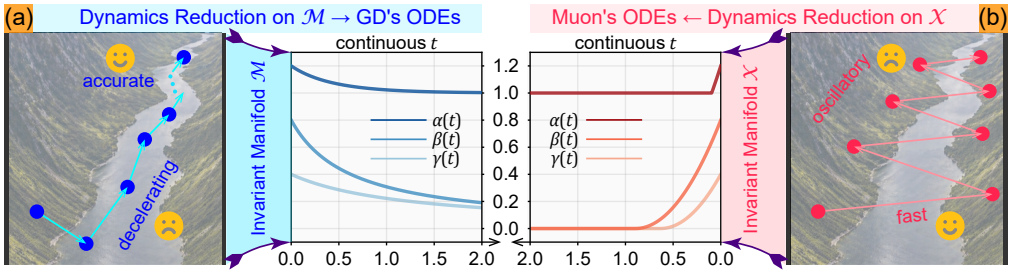

Figure 3: Reduction of the optimization dynamics in a river–valley to scalar ODEs on the invariant manifolds: (a) vanilla GD on $\mathcal{M}$, using (5); (b) simplified Muon on $\mathcal{X}$, using (6).

4.2 Dynamics Analysis of Mixed-spectrum MS from a River-valley View

We now analyze the continuous/discrete dynamics of vanilla GD and simplified Muon on the invariant manifolds introduced above under zero-momentum setting. We use the river–hill splitting

$$\hat{K} =: \begin{bmatrix} \hat{a} & \hat{b}^\top \\ \hat{c} & \hat{D} \end{bmatrix} \in \mathbb{S}^{3\times 3}, \quad \hat{a} \in \mathbb{R}, \quad \hat{b} = \hat{c} \in \mathbb{R}^2, \quad 0 \prec \hat{D} \in \mathbb{S}^{2\times 2}.$$

A diagonal special case. To build intuition, we first consider a particularly simple case in which $\hat{K}$ is diagonal. In this case, the dynamics of α, β, and γ are completely decoupled, and the river is given by $\beta = \gamma = 0$. For notational simplicity, we define $\kappa := \hat{K}/r$.

Proposition 4.1 (Reduced ODE system of vanilla GD and simplified Muon, diagonal case)**.** *Assume that $\kappa = \texttt{Diag}(\kappa_{11}, \kappa_{22}, \kappa_{33})$, the initialization satisfies $\beta_0, \gamma_0 > 0$ and $\alpha_0 > 1$. The vanilla GD flow restricted to $\mathcal{M}$ is governed by the following ODE system with its closed-form solution:*

$$\begin{cases} \dot{\alpha} = -4\kappa_{11}\alpha(\alpha - 1) & \Rightarrow \quad \alpha(t) = \frac{1}{1+((1-\alpha_0)/\alpha_0)e^{-4\kappa_{11}t}}, \\ \dot{\beta} = -4\kappa_{22}\beta^2 & \Rightarrow \quad \beta(t) = \frac{\beta_0}{1+4\kappa_{22}\beta_0 t}, \\ \dot{\gamma} = -4\kappa_{33}\gamma^2 & \Rightarrow \quad \gamma(t) = \frac{\gamma_0}{1+4\kappa_{33}\gamma_0 t}. \end{cases} \tag{5}$$

The simplified Muon flow restricted to $\mathcal{X}$ also induces the ODE system with its closed-form solution:

$$\begin{cases} \dot{\alpha} \in -2\sqrt{\alpha}\,\texttt{sign}[\kappa_{11}(\alpha - 1)] & \Rightarrow \quad \alpha(t) = (\max\{\sqrt{\alpha_0} - t, 1\})^2, \\ \dot{\beta} \in -2\sqrt{\beta}\,\texttt{sign}(\kappa_{22}\beta) & \Rightarrow \quad \beta(t) = (\max\{\sqrt{\beta_0} - t, 0\})^2, \\ \dot{\gamma} \in -2\sqrt{\gamma}\,\texttt{sign}(\kappa_{33}\gamma) & \Rightarrow \quad \gamma(t) = (\max\{\sqrt{\gamma_0} - t, 0\})^2. \end{cases} \tag{6}$$

Therefore, in the diagonal case, Figure 3 plots the closed-form solutions, enabling a direct comparison of the optimization behaviors. Without finite-step-size effects, simplified Muon progresses substantially faster than vanilla GD along both the river and hill components. However, Figure 7 shows that with a finite step size $\eta > 0$, simplified Muon may overshoot switching surfaces such as $\alpha = 1$, inducing oscillations, whereas vanilla GD converges more slowly but more accurately.

General PSD case. While the diagonal special case is a useful sanity check, it precludes coupling between the α (signal), β (spikes), and γ (bulk). In our MSMS model, $\pi \sim \mathcal{N}(0, \Sigma)$ is intended to capture correlated spectral anisotropy in learned representations, e.g. empirical covariances in vision/language models exhibit heterogeneous directions with nontrivial cross-correlations. We thus study the general PSD case $\hat{K} \succeq 0$, where the off-diagonal blocks of $\hat{K}$ encode these couplings.

The reduced discrete iterations are summarized in Table 3, where $\hat{a}_{\text{eff}} := \hat{a} - \bar{q}$ with $\bar{q} := \hat{c}^\top \hat{D}^{-1}\hat{c}$. It gives a natural river-valley decomposition (the interpretation is different from the diagonal case):

Table 1: River-side quantities in the general PSD case.

River item and notation	Geometric interpretation
variable $\theta_k = \alpha_k - 1$	Coordinate along the coupled river. It measures progress of the signal component toward the target value $\alpha = 1$.
force $\phi_k = \hat{a}\theta_k + \hat{b}^\top \omega_k$	Reduced gradient component in the river direction before eliminating the hill variables.
energy $E_k = \hat{a}_{\text{eff}}\theta_k^2$	Effective loss remaining after the hill variables are optimally adjusted. Equivalently, it is the Schur-complement energy along the valley floor.
GD river iteration: $\theta_{k+1} = \theta_k - 4\eta(1+\theta_k)\phi_k$	
Muon river iteration: $\theta_{k+1} = \theta_k - 2\eta\sqrt{1+\theta_k}\,\texttt{sign}(\phi_k)$	

* This table highlights only the river-side quantities. The full river-hill notation and discrete dynamics are given in Table 3, (38), and (45). The hill energy is measured by the hill force ψ_k rather than the raw hill coordinate ω_k (see Remark A.14).

when $\hat{K}$ is not diagonal, the raw coordinates (θ, ω) are coupled, and the river-valley structure is effectively rotated. Consequently, the magnitude of the raw hill variable ω_k alone does not correctly measure the transverse deviation from the river. The relevant transverse quantity is instead the hill force ψ_k. We then define the reduced river as the no-hill-force (as well as no-hill-energy) manifold

$$\mathcal{R} := \{(\theta, \omega) : \psi = 0\} = \{(\theta, \omega) : \omega = -\hat{D}^{-1}\hat{c}\,\theta\}. \qquad \text{(River in PSD case)}$$

Thus the river is generally not the set $\omega = 0$; it is a tilted one-dimensional manifold parameterized by θ. On this manifold, the river force becomes $\phi = \hat{a}\theta - \hat{b}^\top \hat{D}^{-1}\hat{c}\,\theta = \hat{a}_{\text{eff}}\theta$. Therefore, if $\hat{a}_{\text{eff}} > 0$, the only point on the river with zero river force is $\theta = 0$, which further gives $\omega = 0$. In other words, $\psi = 0$ alone means that the iterate lies on the river, but exact recovery also requires the river force to vanish. This rotated decomposition also separates the reduced quadratic energy. Indeed, by completing the square, (4) becomes $\hat{a}_{\text{eff}}\theta^2 + \psi^\top \hat{D}^{-1}\psi$ (see Definition A.12), which is exactly the river and hill energies defined in Table 3. Hence $H_k = O(\eta^2)$ means that the iterate lies inside an $\psi_k = O(\eta)$ vicinity (we call a hill-force tube) around the river. In the following Theorem 4.2, we formally study this regime when the hill energy has reached the $O(\eta^2)$ scale and $E_k \gg H_k$. The proof follows from Propositions A.19, A.25, and A.27; see Section A.3 for details.

Theorem 4.2 (Dynamics along river for the PSD case)**.** *Assume that the river curvature $\hat{a}_{\text{eff}} > 0$. Suppose that, on the time interval under consideration, $0 < \underline{\alpha} \le 1 + \theta_k \le \overline{\alpha}$. For an algorithm $\square \in \{\mathrm{GD}, \mathrm{Muon}\}$, after a hill-reduction time $N_\square$, the trajectory remains in an $O(\eta)$ hill-force tube:*

$$H_k \le R_\square^2 \eta^2, \qquad \forall k \ge N_\square, \tag{7}$$

for some constant $R_\square > 0$ independent of η. Hitting times $T_\square$ are defined as

$$T_\square := \inf\{k \ge N_\square \,:\, |\theta_k| \le C_\square \eta\}. \tag{8}$$

Then the following statements hold.

(i) GD makes geometric progress along the river. *Assume that the GD iterate satisfies (7) with constants $N_{\mathrm{GD}}, R_{\mathrm{GD}}$ and η is sufficiently small. Define $C_{\mathrm{GD}} := \max\{2\sqrt{\bar{q}}/\hat{a}_{\text{eff}}, \sqrt{\Gamma_{\mathrm{GD}}/\hat{a}_{\text{eff}}}\}R_{\mathrm{GD}}$ with a fixed constant $\Gamma_{\mathrm{GD}} \gg 1$. Then, for every $k \ge N_{\mathrm{GD}}$ such that $|\theta_k| \ge C_{\mathrm{GD}}\eta$,*

$$|\theta_{k+1}| \le (1 - 2\underline{\alpha}\,\hat{a}_{\text{eff}}\eta)\,|\theta_k|, \tag{9}$$

and the iterate is in a river-dominant phase with $E_k \gg H_k$. Hence the number of post-transition GD iterations required to reduce a constant-scale river error to the $O(\eta)$ scale is bounded by

$$T_{\mathrm{GD}} - N_{\mathrm{GD}} \le \left\lceil \frac{1}{2\underline{\alpha}\,\hat{a}_{\text{eff}}\eta} \log \frac{|\theta_{N_{\mathrm{GD}}}|}{C_{\mathrm{GD}}\eta} \right\rceil = O\left(\frac{1}{\eta}\log\frac{1}{\eta}\right). \tag{10}$$

(ii) Muon makes linear progress along the river. *Assume that the Muon iterate satisfies (7) with constants $N_{\mathrm{Muon}}, R_{\mathrm{Muon}}$. Define $C_{\mathrm{Muon}} := \max\{2\sqrt{\bar{q}}R_{\mathrm{Muon}}/\hat{a}_{\text{eff}}, 2\sqrt{\overline{\alpha}}, R_{\mathrm{Muon}}\sqrt{\Gamma_{\mathrm{Muon}}/\hat{a}_{\text{eff}}}\}$ with a fixed constant $\Gamma_{\mathrm{Muon}} \gg 1$. Then, for every $k \ge N_{\mathrm{Muon}}$ such that $|\theta_k| \ge C_{\mathrm{Muon}}\eta$,*

$$|\theta_{k+1}| \le |\theta_k| - 2\sqrt{\underline{\alpha}}\,\eta, \tag{11}$$

and the iterate is in a river-dominant phase with $E_k \gg H_k$. Therefore the number of post-transition Muon iterations required to reduce a constant-scale river error to the $O(\eta)$ scale is bounded by

$$T_{\mathrm{Muon}} - N_{\mathrm{Muon}} \le \left\lceil \frac{(|\theta_{N_{\mathrm{Muon}}}| - C_{\mathrm{Muon}}\eta)_+}{2\sqrt{\underline{\alpha}}\,\eta} \right\rceil = O\left(\frac{1}{\eta}\right). \tag{12}$$

Finally, assume that once θ_k enters the $O(\eta)$-neighborhood of the river equilibrium, both vanilla GD and simplified Muon remain in a relative river tube (see Remark A.28):

$$\sqrt{\bar{q}\, H_k} \le \bar{\rho}\, \hat{a}_{\text{eff}} |\theta_k|, \qquad \bar{\rho} \in (0,1). \tag{13}$$

Under this condition, GD can further refine the river variable from $O(\eta)$ to any target accuracy $\varepsilon \ll \eta$, whereas fixed-step simplified Muon increases the river energy after one step, i.e., $E_{k+1} > E_k$.

At the hitting times $T_\square$, both algorithms satisfy $|\theta_{T_\square}| = O(\eta)$ and $H_{T_\square} = O(\eta^2)$, and therefore $\|\omega_{T_\square}\| \le \|\hat{D}^{-1}\|(\|\psi_{T_\square}\| + \|\hat{c}\|\,|\theta_{T_\square}|) = O(\eta)$. Both methods thus reach an $O(\eta)$-neighborhood of the river equilibrium $(\theta, \omega) = (0, 0)$, but their behavior there is qualitatively different. Vanilla GD is force-proportional and can continues to refine as the residual forces vanish (Remark A.20, Proposition A.27). By contrast, Muon remains sign-based and non-smooth; with a fixed step size, it may overshoot the equilibrium and generate persistent oscillations (Remark A.26, Proposition A.27). Consequently, simplified Muon typically reaches the neighborhood faster, whereas GD attains better final accuracy, consistent with Figure 3 (also see the Lyapunov analyses in Remarks A.18 and A.24).

5 River-Valley as Tool in Generalized Settings

In the previous sections, we have mostly focused on the simplified Muon without momentum and our specific Mixed-spectrum MS model to provide fine-grained analysis and explicit trajectory dynamics. However, we argue that the river–valley framework can extend beyond these simplified assumptions and help validate our insights in more generalized settings. Let's generalize the problem[2] to optimize over an objective function $\mathcal{F}(X) : \mathbb{R}^{n\times n} \to \mathbb{R}$ to search for a critical point of interest $X^\star \in \mathbb{R}^{n\times n}$. This point could be the global optimum or any point that generalizes well. Suppose our target $X^\star$ admits a singular value decomposition (SVD) as $X^\star = U_\star \Sigma_\star V_\star^\top$, then following suit of our river-valley analysis, we define the spectral river as follows:

$$\mathcal{R}^\star := \{X^\star + \mathcal{B}(x) \;\mid\; x \in \mathbb{R}^n\}, \qquad \mathcal{B}(x) := U_\star \texttt{Diag}(x) V_\star^\top,$$

such that along the river, we are only away from the target $X^\star$ up to a spectral scale difference, so only the singular values are different.

In this section, we first show, once the iterate enters the spectral river, the initial convergence speed of Muon is faster than that of GD, whereas in the final phase close to $X^\star$, Muon could easily overshoot and oscillate, hindering convergence. These phenomena reflect the same set of observations made in the mixed-spectrum MS case, which indicates a broader applicability of our theory. Our generalized proofs hold contingent on two basic requirement: 1) the stored momentum $M_0^{\text{Muon/GD}}$ at X_0 has to point mostly in the river direction and 2) the river is approximate gradient-invariant near $X^\star$. Here, gradient-invariance near $X^\star$ holds if, for all sufficiently small x, $\nabla\mathcal{F}(X^\star + \mathcal{B}(x)) \in \mathcal{R}^\star$. When this condition is met, it holds that $\nabla^2 f(X^\star)[\mathcal{B}(x)] = \mathcal{B}(Kx)$ for a fixed K (please find details in Appendix B.1). With these assumptions in mind, we first present the informal early-stage result:

Theorem 5.1 (Early-stage Descent along the River, informal)**.** *Suppose the iterate $X = X^\star + \mathcal{B}(x)$ lies on the spectral river, and define the river-restricted objective $f(x) := \mathcal{F}(X^\star + \mathcal{B}(x)) - \mathcal{F}(X^\star)$. Let L_{full} denote a smoothness constant of $\mathcal{F}$ in the ambient matrix space, and let L_{riv} denote a smoothness constant of f restricted to the river. Then, under the assumption that the orthogonalized momentum after update is mostly aligned with $\nabla f(x)$, we show that the one-step decreases of Muon and Momentum GD respectively satisfy*

$$\textit{Muon decrease} \;\gtrsim\; \frac{\rho^2 \|\nabla f(x)\|_1^2}{2L_{\text{riv}} n}, \qquad \textit{Momentum GD decrease} \;\lesssim\; \frac{\|\nabla f(x)\|_2^2}{2L_{\text{full}}}.$$

Here, $\rho \in (0, 1]$ measures the alignment between the effective momentum direction and the river gradient $\nabla f(x)$, and the ratio $\|\nabla f(x)\|_1/\|\nabla f(x)\|_2$ quantifies how uniformly the gradient mass is distributed across coordinates (i.e., the degree of non-sparsity).

The formal statements and its associated proof appear in Theorem B.10. Since L_{riv} is the smoothness constant of the objective restricted to the spectral river, whereas L_{full} is the ambient smoothness constant over all directions, one always has $L_{\text{riv}} \le L_{\text{full}}$. Moreover, when the local Hessian is much

[2] X can be easily generalized to a non-square matrix, but we keep the shape to be squared for simplicity.

flatter on the river than on its transverse hill complement, we have $L_{\mathrm{riv}} \ll L_{\mathrm{full}}$, thereby making Muon's decrease much bigger. We refer to Remark B.11 for further discussion.

Then, we turn to the late-stage analysis of starting from a point X_0 along the river that is close to $X^\star$ (i.e., already in the final phase of convergence). We show that momentum GD could provide steady decrease in loss whereas Muon with momentum diverges in two steps. Before presenting the results, we discuss why this would happen. For a proposed update $X \leftarrow X - \eta D$, Taylor expansion gives

$$\mathcal{F}(X - \eta D) = \mathcal{F}(X) - \eta \langle \nabla \mathcal{F}(X), D \rangle + \frac{\eta^2}{2} \left\langle D, \nabla^2 \mathcal{F}(X)[D] \right\rangle + \mathcal{O}(\eta^3 \|D\|_F^3).$$

If we assume that η is sufficiently small such that $\mathcal{O}(\eta^3) \approx 0$, then the quadratic model decreases when $0 < \eta < \eta_D^{\max}(X)$ with $\eta_D^{\max}(X) := 2\langle \nabla \mathcal{F}(X), D \rangle / \langle D, \nabla^2 \mathcal{F}(X)[D] \rangle$, provided $\langle \nabla \mathcal{F}(X), D \rangle > 0$ and $\langle D, \nabla^2 \mathcal{F}(X)[D] \rangle > 0$. At a river point $X = X^\star + a\mathcal{B}(x)$, suppose gradient invariance holds such that $\nabla \mathcal{F}(X) \approx a\mathcal{B}(Kx)$. Regardless of the momentum, for the spectrally normalized muon update $D = \mathcal{B}(\mathbf{1})$, the one-step safe range is

$$0 < \eta < \eta_D^{\max}(X) = 2\frac{\langle a\mathcal{B}(Kx), \mathcal{B}(\mathbf{1}) \rangle}{\langle \mathcal{B}(\mathbf{1}), \mathcal{B}(K\mathbf{1}) \rangle} = 2ay, \qquad y := \frac{\mathbf{1}^\top K x}{\mathbf{1}^\top K \mathbf{1}}. \tag{14}$$

If momentum keeps the same polar direction for two consecutive Muon steps, then two steps give the displacement $2\eta\mathcal{B}(\mathbf{1})$, and the two-step safe range is approximately $0 < \eta < ay$, thus Muon's admissible step-size scales with $\mathcal{O}(a)$, which necessitates a slowing progress. For GD, however, we can easily show that the admissible step-size is agnostic of residual scale a (see Proposition B.12 as an illustration). This means that it is the contraction of admissible step-size that is problematic with Muon. We also prove in Proposition B.19 that Adam could behave more like GD in the sense that it does not require a shrinking step-size of $\mathcal{O}(a)$ in the final phase, which means that Adam and its variants could be a better choice for late-stage convergence. This serves as a theoretical footnote as to why manual LR decay is necessary for Muon. However, a manually prescribed decay schedule could create mismatch with the actual optimization progress, calling for more principled approaches (see Appendix B.4.2 for details). We believe an alternative, or improvement to explicit scheduling is to allow for a two-stage approach where Muon is only used in the first phase, followed by a residual-agnostic refinement optimizer. To make our arguments more precise, we now present the late-stage behavior of Muon as an informal theorem:

Theorem 5.2 (Late-stage Convergence on River, informal)**.** *Assume $X^\star$ is a nondegenerate local minimizer, and consider an initialization of the form $X_0 = X^\star + \mathcal{B}(ax_0)$ with $0 < a \ll 1$, so that X_0 lies on the spectral river and is already close to $X^\star$. Suppose further that the river is approximately gradient-invariant (see Lemma B.9) near $X^\star$, and that the stored momentum remains dominated by river directions. Define y_0 as per* (14) *and $\eta^{\mathrm{Muon}} = (1+\varepsilon)ay_0$ with $0 < \varepsilon < 1$. Then momentum Muon decreases the objective at the first step but overshoots by the second step:*

$$\mathcal{F}(X_1^{\mathrm{Muon}}) < \mathcal{F}(X_0) < \mathcal{F}(X_2^{\mathrm{Muon}}).$$

By contrast, starting from the same point and with the same stored momentum, momentum gradient descent admits a residual-independent (independent of a) step-size threshold $\eta_0 = O(1)$ such that for all sufficiently small $0 < \eta^{\mathrm{GD}} \le \eta_0$,

$$\mathcal{F}(X_2^{\mathrm{GD}}) \le \mathcal{F}(X_1^{\mathrm{GD}}) \le \mathcal{F}(X_0).$$

Since $\eta^{\mathrm{Muon}} = O(a)$, one has $\eta^{\mathrm{Muon}} < \eta_0$ for sufficiently small a.

Therefore, even the same numerical step size $\eta^{\mathrm{GD}} = \eta^{\mathrm{Muon}}$ is safe for GD, while Muon with step size η^{Muon} can overshoot after two persistent polar steps. In addition, GD can often take a much larger step size than Muon in the final river stage. See Appendix B.3.2 for proof and more details.

6 Empirical Evidence for Muon as an Early-stage Explorer

As discussed in Section 5, the late-stage convergence of Muon generally requires a decreasing LR schedule, because the admissible step size shrinks as the iterates approach a local solution. Although Muon can perform well with a carefully tuned LR schedule, finding such a schedule may require substantial experimentation, which becomes particularly costly at the scale of modern LLMs. Motivated by our theoretical analysis, we provide preliminary empirical evidence that Muon is most effective as

an early-stage exploration optimizer, followed by a refinement optimizer such as AdamW in the late stage. All the LLM pre-training experiments[3] are conducted by training a 250M-parameter LLaMA-style Transformer [Touvron et al., 2023] from scratch on OpenWebText2 [Gao et al., 2020]. All runs use the same compute budget and are executed for 15k or 20k iterations. Several Muon-only baselines (with linear and `cos_inf` LR schedules) are compared against two-stage hybrid methods that train with Muon using a constant LR and then switch to AdamW at different iterations. Additional details of the LLM pre-training experiments are provided in Appendix C.2.

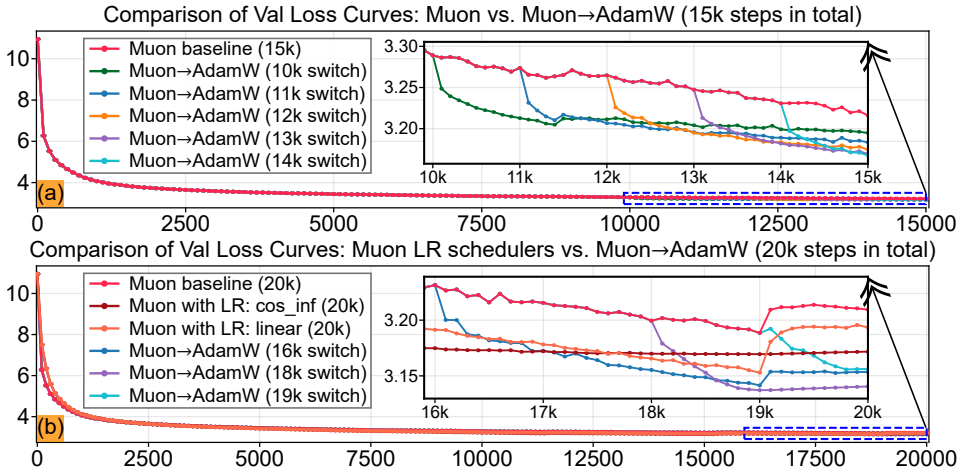

Figure 4: **Validation loss curves for LLM pre-training with a "Muon→AdamW" switch. (a)** 15k-step Muon baseline and Muon→AdamW runs that switch at steps 10k, 11k, 12k, 13k and 14k under a constant LR (no LR decay). **(b)** 20k-step Muon-only baselines with constant, `cos_inf`, and linear LR schedulers, plus Muon→AdamW runs switching at 16k, 18k, and 19k. Each hybrid matches the constant-LR Muon trajectory up to the switch step and then continues with AdamW.

Figure 4 compares Muon→AdamW runs with different switching times. Panel (a) shows that the validation loss generally decreases as the switch is delayed from 10k to 14k steps, suggesting that a later transition can be beneficial. However, the benefit of delaying the switch is expected to diminish once the switch point is pushed too close to the end of training. Panel (b) extends training to 20k iterations and shows that the hybrid runs attain lower final validation losses than the Muon-only baselines equipped with either linear or `cos_inf` LR schedulers. Overall, these results are consistent with our theory. As the iterates approach a local solution, Muon admits a smaller effective step size. Although carefully tuned decreasing LR schedules can mitigate oscillations, Muon may consequently exhibit slower late-stage convergence. By contrast, after switching to AdamW, the hybrid method follows a more stable loss trajectory and maintains a consistent downward trend, ultimately achieving a lower final validation loss than Muon even with carefully tuned LR schedules. These experiments further suggest that the hybrid strategy is relatively insensitive to the precise switching time: across a broad range of late-stage switch points, switching to AdamW continues to produce a stable decrease in validation loss, except when the switch is made extremely late.

7 Conclusion

Our work complements the growing literature on Muon by identifying not only its theoretical advantage (i.e., fast early-stage exploration) but also its limitation (i.e., poor late-stage exploitation). Our trajectory-level dynamical analysis demonstrates the utility of the river–valley perspective for disentangling signal-aligned and nuisance directions in the training landscape, while also introducing a useful Mixed-spectrum MS benchmark testbed. At the same time, the river–valley decomposition provides only one lens on modern training dynamics; integrating this view with related phenomena, such as edge-of-stability behavior, is an important direction for future work. Finally, our preliminary experiments suggest that multi-stage optimizer schedules, in which one switches optimizers during pre-training, may be a promising direction for improving training efficiency and final performance.

[3]Experimental setting adapted from `https://github.com/epfml/schedules-and-scaling`.

Acknowledgments

We thank the anonymous reviewers and area chair for their constructive comments. This work was supported by the National Natural Science Foundation of China (Grant No. 62506314), the Research Grants Council of Hong Kong (Grant No. 21208525), and City University of Hong Kong (Project No. 9382001).

References

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 21681–21716. PMLR, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

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

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Mikail Khona, Aditya Vavre, Boxiang Wang, Deyu Fu, Hao Wu, Mike Chrzanowski, Bryan Catanzaro, Dheevatsa Mudigere, Jeff Pool, Michael Lightstone, et al. Soap, muon, and beyond: Pushing llm pretraining scales. *arXiv preprint arXiv:2607.20548*, 2026.

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

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

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024.

Joel Schlotthauer, Christian Kroos, Chris Hinze, Viktor Hangya, Luzian Hahn, and Fabian Küch. Pre-training llms on a budget: A comparison of three optimizers. *arXiv preprint arXiv:2507.08472*, 2025.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650, 2019.

Martin van der Schelling, Deepesh Toshniwal, and Miguel A Bessa. Reinforcement learning to choose optimizers. *arXiv preprint arXiv:2609.01811*, 2026.

Xianbiao Qi, Marco Chen, Jiaquan Ye, Yelin He, and Rong Xiao. Delving into muon and beyond: Deep analysis and extensions. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 100249–100268. PMLR, 2026.

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.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why muon outperforms adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026a.

Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding of orthogonalization in muon. In *Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM)*, 2025.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all muon needs: How orthogonalization controls learning rate and convergence. *arXiv preprint arXiv:2605.13079*, 2026.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of muon. *arXiv preprint arXiv:2606.08388*, 2026.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Zhehang Du and Weijie Su. The newton-muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Tong Che. Second-order muon done right: A principled marriage of spectral geometry and curvature. *arXiv preprint arXiv:2608.09763*, 2026.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, volume 2026, pages 142276–142311, 2026.

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

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pages 37382–37419, 2026b.

Kaifei Wang, Binghui Li, 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*, volume 306 of *Proceedings of Machine Learning Research*, pages 127884–127936. PMLR, 2026c.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026a.

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.

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

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

Thang Do, Steffen Dereich, and Arnulf Jentzen. On muon optimization: From non-convergence to an error analysis with polar express and the newton-schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Andrej Karpathy. nanochat: The best chatgpt that $100 can buy, 2025. URL `https://github.com/karpathy/nanochat`.

Andrej Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026.

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

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 36767–36811. PMLR, 2026.

Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on muon from simple quadratics. *arXiv preprint arXiv:2602.11948*, 2026.

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

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, 2024.

Shuchen Zhu, Rizhen Hu, Mingze Wang, Mou Sun, Xue Wang, Kun Yuan, and Zaiwen Wen. Accelerating llm pre-training through flat-direction dynamics enhancement. *arXiv preprint arXiv:2602.22681*, 2026.

Zhendong Huang, Hengjie Cao, Fang Dong, Ruijun Huang, Mengyi Chen, Yifeng Yang, Xin Zhang, Anrui Chen, Mingzhi Dong, Yujiang Wang, Jinlong Hou, Qin Lv, Robert P. Dick, Yuan Cheng, Tun Lu, Fan Yang, and Li Shang. Spectra: Rethinking optimizers for LLMs under spectral anisotropy. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 46441–46456. PMLR, 2026.

Raphael Bernas, Fanny Jourdan, Antonin Poché, and Céline Hudelot. Revisiting anisotropy in language transformers: The geometry of learning dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 7741–7765. PMLR, 2026.

Joachim Weickert et al. *Anisotropic diffusion in image processing*. Teubner Stuttgart, 1998.

Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Xiaomin Yu, Yijiang Li, Yuhui Zhang, Hanzhen Zhao, Yue Yang, Hao Tang, Yue Song, Xiaobin Hu, Chengwei Qin, Shuicheng Yan, et al. Anisotropic modality align. *arXiv preprint arXiv:2605.07825*, 2026.

Hyesong Choi, Daeun Kim, Song Park, Taekyung Kim, Byeongho Heo, Sangdoo Yun, Dongbo Min, and Dongyoon Han. Isotropic embedding perturbations for robust vision language encoders. In *European Conference on Computer Vision*, pages 281–299. Springer, 2026.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Gradient descent with adaptive stepsize converges (nearly) linearly under fourth-order growth: D. davis et al. *Mathematical Programming*, pages 1–66, 2025.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, volume 2025, pages 42840–42885, 2025.

Aleksandr Dremov, Alexander Hägele, Atli Kosson, and Martin Jaggi. Training dynamics of the cooldown stage in Warmup-Stable-Decay learning rate scheduler. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=ZnSYEcZod3`.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Maosong Sun, Zhiyuan Liu, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules. In *International Conference on Learning Representations*, volume 2025, pages 39973–40017, 2025.

Miao Lu, Zeyu Bian, Kaiyue Wen, Beining Wu, Siyu Chen, Tianhao Wang, and Zhiyuan Li. Towards understanding momentum acceleration in river-valley loss landscape. In *High-dimensional Learning Dynamics 2026*, 2026.

Jueun Kim, Baekrok Shin, Jihun Yun, Beomhan Baek, Minhak Song, and Chulhee Yun. Amuse: Anytime muon with stable gradient evaluation. *arXiv preprint arXiv:2605.22432*, 2026b.

Mark A Davenport, Marco F Duarte, Yonina C Eldar, and Gitta Kutyniok. Introduction to compressed sensing., 2012.

Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013.

Hanxi Li, Chunhua Shen, and Qinfeng Shi. Real-time visual tracking using compressive sensing. In *CVPR 2011*, pages 1305–1312. IEEE, 2011.

Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

Anuj Rajani, Paritosh Mittal, Aishwarya Jain, and Angshul Majumdar. A blind compressed sensing formulation for collaborative filtering. In *2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 000438–000443. IEEE, 2014.

Ming Jin, Igor Molybog, Reza Mohammadi-Ghazi, and Javad Lavaei. Towards robust and scalable power system state estimation. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3245–3252. IEEE, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

John Neter, Michael H Kutner, Christopher J Nachtsheim, William Wasserman, et al. *Applied linear statistical models*. Irwin Chicago, 1996.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alex Krizhevsky and Geoffrey E. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 9687–9733. PMLR, 2026.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *Advances in Neural Information Processing Systems*, 38:39622–39669, 2026.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Gilbert W Stewart. Perturbation theory for the singular value decomposition. *SVD and Signal Processing II, Algorithms, Analysis and Applications*, pages 99–109, 1991.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Zhenlong Yuan, Dapeng Zhang, Zehao Li, Chengxuan Qian, Jianing Chen, Yinda Chen, Kehua Chen, Tianlu Mao, Zhaoxin Li, Hao Jiang, et al. Dvp-mvs++: Synergize depth-normal-edge and harmonized visibility prior for multi-view stereo. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al. Video-star: Reinforcing open-vocabulary action recognition with tools. In *International Conference on Learning Representations*, volume 2026, pages 51445–51468, 2026a.

Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pages 97934–97955, 2026b.

Appendix Contents

A Additional Details for Section 4 **16**

A.1 Properties of the Mixed Spectrum Sensing Model 17

A.2 Reduced Dynamics of the Mixed-spectrum Matrix Sensing Model 18

A.3 River–Valley Structure of the Reduced Mixed Spectrum Sensing Model 22

A.3.1 Additional Definitions, Assumptions, and Basic Geometry 22

A.3.2 Gradient Descent Reduced Dynamics 24

A.3.3 Simplified Muon Reduced Dynamics 27
A.3.4 Reduced Dynamics of Normalized GD and Scaled Muon in the PSD case . 31

B Additional Details for Section 5 **32**
B.1 Spectral-River Geometry 32
B.2 Early-Stage: When Muon Accelerates River Progress 34
B.3 Late-Stage: When Muon Fails in River Refinement 36
B.3.1 A toy quadratic model: explicit overshoot under `msign` 36
B.3.2 Late-stage overshoot for general smooth objectives 39
B.4 Implications for Algorithm Design 44
B.4.1 Late-stage Adam behaves more like gradient methods than Muon 44
B.4.2 Guidance for algorithm design 45

C Additional Details for Section 6 **45**
C.1 Additional Case Studies 45
C.1.1 2D Anisotropic Spectral Slice Case 45
C.1.2 Mixed-Spectrum Matrix Sensing Case 47
C.1.3 Two-Layer Neural Network Case 48
C.2 Empirical Observations on LLM Pre-training 50
C.2.1 Implementation Details for LLM Pre-training 50
C.2.2 Effects of Switching Time and Learning-Rate Scheduling 50
C.2.3 How to Switch: Warm-Starting AdamW from a Muon Checkpoint 51
C.2.4 When to Switch: A Possible Local Criterion 52

A Additional Details for Section 4

In this section, we note that Muon's `msign` operation effectively combines two distinct effects: (i) flattening the singular-value spectrum and (ii) normalizing the gradient magnitude. To disentangle these two effects, following the setup of [Vasudeva et al., 2026, Fan et al., 2026], we express the ODE flows considered below in the following unified direction–speed form:

$$\dot{X} = -s(X)d(X), \qquad \|d(X)\|_F = 1 \qquad \Longleftrightarrow \qquad \frac{dX}{dl} = -d(X), \qquad \frac{dl}{dt} = s(X),$$

where $d(X)$ denotes the update direction, $s(X)$ denotes the speed, and l denotes the arc length of the trajectory. The resulting direction–speed decompositions are summarized in Table 2. Comparing Muon with NGD, or SMuon with GD, isolates the effect of flattening the singular-value spectrum, since the two methods in each pair have the same speed but different directions. In contrast, comparing Muon with SMuon, or NGD with GD, isolates the effect of gradient magnitude normalization, since the methods in each pair have the same direction but different speeds.

Table 2: Direction–speed decomposition of the considered optimization flows.

Optimizer	**Direction** $d(X)$	**Speed** $s(X)$
GD	$\nabla h(X)/\|\nabla h(X)\|_F$	$\|\nabla h(X)\|_F$
Normalized GD (NGD)	$\nabla h(X)/\|\nabla h(X)\|_F$	1
Muon	$\mathtt{msign}(\nabla h(X))/\|\mathtt{msign}(\nabla h(X))\|_F$	1
Scaled Muon (SMuon)	$\mathtt{msign}(\nabla h(X))/\|\mathtt{msign}(\nabla h(X))\|_F$	$\|\nabla h(X)\|_F$

[1] Here, $\|\mathtt{msign}(\nabla h(X))\|_F = 1$.

Throughout this section, none of the four optimizers uses momentum, e.g., we set $\xi_t = 0$ in Muon update. Accordingly, we sometimes refer to GD as *vanilla GD* and Muon as *simplified Muon*. Nevertheless, a direct comparison between simplified Muon and vanilla GD is also meaningful, and we use this comparison when the combined effect of direction and speed is of interest.

A.1 Properties of the Mixed Spectrum Sensing Model

Definition A.1 (Gaussian Orthogonal Ensemble)**.** *A symmetric random matrix $G \in \mathbb{R}^{n\times n}$ is said to follow the normalized Gaussian Orthogonal Ensemble, denoted by $G \sim \mathrm{GOE}(n)$, if*

$$G = \frac{1}{\sqrt{n}} \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix}, \qquad g_{ij} = g_{ji},$$

where the upper-triangular entries $\{g_{ij} : 1 \le i \le j \le n\}$ are independent and satisfy

$$g_{ij} \sim \mathcal{N}(0,1) \quad (i<j), \qquad g_{ii} \sim \mathcal{N}(0,2).$$

Equivalently,

$$\mathbb{E}[G_{ij}] = 0, \qquad \mathbb{E}[G_{ij}^2] = \frac{1}{n} \quad (i \neq j), \qquad \mathbb{E}[G_{ii}^2] = \frac{2}{n}.$$

In particular, for every fixed symmetric matrix $M \in \mathbb{R}^{n\times n}$,

$$\mathbb{E}\big[\langle G, M\rangle^2\big] = \frac{2}{n}\|M\|_F^2.$$

When $\{G_i\}_{i=1}^m$ are used as sensing matrices, they are taken to be independent copies of G.

Lemma A.2 (Population loss and its gradients)**.** *Consider the mixed spectrum sensing model*

$$A = G + \pi_\star P_\star + \pi_\diamond P_\diamond + \pi_\circ P_\circ,$$

where $G \sim \mathrm{GOE}(n)$, $\pi = (\pi_\star, \pi_\diamond, \pi_\circ)^\top \sim \mathcal{N}(0, \Sigma)$, and G is independent of π. Based on our construction, $P_\star, P_\diamond, P_\circ$ are mutually orthogonal projection matrices. Let

$$\mathcal{L}(X) := \mathcal{F}(XX^\top) := \frac{1}{2}\mathbb{E}\Big[\big\langle A, XX^\top - P_\star\big\rangle^2\Big], \qquad \mathcal{F}(M) = \mathcal{F}(XX^\top),$$

with square matrix variable $M := XX^\top \in \mathbb{R}^{n\times n}$ (due to the BM-factorized MS form). Then

$$\mathcal{F}(M) = \frac{1}{n}\|M - P_\star\|_F^2 + \frac{1}{2} q(M)^\top \Sigma q(M), \tag{15}$$

where

$$q(M) := \begin{bmatrix} \langle P_\star, M - P_\star\rangle \\ \langle P_\diamond, M\rangle \\ \langle P_\circ, M\rangle \end{bmatrix}.$$

Moreover, the Frobenius gradient of $\mathcal{F}$ with respect to M is

$$G(M) := \nabla_M \mathcal{F}(M) = \frac{2}{n}(M - P_\star) + g_\star P_\star + g_\diamond P_\diamond + g_\circ P_\circ, \tag{16}$$

where

$$g(M) := \begin{bmatrix} g_\star \\ g_\diamond \\ g_\circ \end{bmatrix} = \Sigma q(M) \in \mathbb{R}^3.$$

Consequently, under the standard Frobenius-gradient convention,

$$\nabla_X \mathcal{L}(X) = 2\, G(XX^\top)X.$$

Proof. Let $\Delta := M - P_\star$. Since G and π are independent and centered,

$$\mathcal{F}(M) = \frac{1}{2}\mathbb{E}\left[\big(\langle G, \Delta\rangle + \pi^\top q(M)\big)^2\right] = \frac{1}{2}\mathbb{E}\big[\langle G, \Delta\rangle^2\big] + \frac{1}{2}\mathbb{E}\big[(\pi^\top q(M))^2\big].$$

Because Δ is symmetric, the GOE covariance identity gives

$$\mathbb{E}\big[\langle G, \Delta\rangle^2\big] = \frac{2}{n}\|\Delta\|_F^2.$$

Since $\pi \sim \mathcal{N}(0, \Sigma)$,

$$\mathbb{E}\big[(\pi^\top q(M))^2\big] = q(M)^\top \Sigma q(M).$$

Combining the two displays yields (15).

It remains to compute the gradient. For any symmetric perturbation ΔM, we have

$$\mathrm{d}q(M)[\Delta M] = \begin{bmatrix}\langle P_\star, \Delta M\rangle \\ \langle P_\diamond, \Delta M\rangle \\ \langle P_\circ, \Delta M\rangle\end{bmatrix}.$$

Using the symmetry of Σ,

$$\begin{aligned}\mathrm{d}\mathcal{F}(M)[\Delta M] &= \frac{2}{n}\langle M - P_\star, \Delta M\rangle + (\Sigma q(M))^\top \mathrm{d}q(M)[\Delta M] \\ &= \left\langle \frac{2}{n}(M - P_\star) + g_\star P_\star + g_\diamond P_\diamond + g_\circ P_\circ, \Delta M \right\rangle,\end{aligned}$$

where $g(M) = \Sigma q(M)$. This proves (16).

Finally, since $M = XX^\top$, for any perturbation ΔX of X,

$$\mathrm{d}M[\Delta X] = X(\Delta X)^\top + (\Delta X)X^\top.$$

Therefore,

$$\mathrm{d}\mathcal{L}(X)[\Delta X] = \left\langle G(M), X(\Delta X)^\top + (\Delta X)X^\top \right\rangle = 2\left\langle G(M)X, \Delta X\right\rangle,$$

where we used that $G(M) = \nabla_M \mathcal{F}(M)$ is symmetric. Hence

$$\nabla_X \mathcal{L}(X) = 2\,\nabla_M \mathcal{F}(XX^\top)X = 2\,G(XX^\top)X.$$

This completes the proof. □

A.2 Reduced Dynamics of the Mixed-spectrum Matrix Sensing Model

In the reduced dynamics, we write $\kappa := \hat{K}/r = 2I_3/n + r\Sigma$ whenever an explicit diagonal formula is given. Using $\hat{K}$ instead only rescales time and the stepsize by the fixed rank r.

Lemma A.3 (Vanilla GD dynamics on the invariant manifold $\mathcal{M}$)**.** *Let*

$$\mathcal{M} := \{M = \alpha P_\star + \beta P_\diamond + \gamma P_\circ : \alpha, \beta, \gamma \in \mathbb{R}\}.$$

If $M \in \mathcal{M}$ with $M = \alpha P_\star + \beta P_\diamond + \gamma P_\circ$, then the population matrix gradient satisfies

$$G(M) = uP_\star + vP_\diamond + wP_\circ \in \mathcal{M},$$

where

$$\begin{bmatrix}u \\ v \\ w\end{bmatrix} = \frac{2}{n}\begin{bmatrix}\alpha - 1 \\ \beta \\ \gamma\end{bmatrix} + \Sigma\begin{bmatrix}r(\alpha - 1) \\ r\beta \\ r\gamma\end{bmatrix}.$$

Moreover, the gradient flow induced by the Burer-Monteiro factorization satisfies

$$\dot{M} = -2\left[G(M)M + M\,G(M)\right],$$

and $\mathcal{M}$ is invariant under this flow.

More explicitly, we have the following ODE system if $\alpha = a^2$, $\beta = b^2$, and $\gamma = c^2$:

$$\begin{cases}\dot{\alpha} = -4u\alpha, \\ \dot{\beta} = -4v\beta, \\ \dot{\gamma} = -4w\gamma.\end{cases} \quad \Rightarrow \quad \begin{cases}\dot{a} = -2ua, \\ \dot{b} = -2vb, \\ \dot{c} = -2wc.\end{cases}$$

Proof. By the expression of the population gradient (16), where

$$\begin{bmatrix} g_\star \\ g_\diamond \\ g_\circ \end{bmatrix} = \Sigma \begin{bmatrix} \langle P_\star, M - P_\star \rangle \\ \langle P_\diamond, M \rangle \\ \langle P_\circ, M \rangle \end{bmatrix}.$$

Since $P_\star, P_\diamond, P_\circ$ are mutually orthogonal projections with ranks r, r, r, we have

$$\langle P_\star, M - P_\star \rangle = r(\alpha - 1), \qquad \langle P_\diamond, M \rangle = r\beta, \qquad \langle P_\circ, M \rangle = r\gamma.$$

Substituting these identities into the gradient formula gives

$$G(M) = uP_\star + vP_\diamond + wP_\circ,$$

with

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{2}{n} \begin{bmatrix} \alpha - 1 \\ \beta \\ \gamma \end{bmatrix} + \Sigma \begin{bmatrix} r(\alpha - 1) \\ r\beta \\ r\gamma \end{bmatrix}.$$

Thus $G(M) \in \mathcal{M}$.

It remains to derive the closed dynamics. The gradient flow in the factor variable $X \in \mathbb{R}^{n \times r_{\text{search}}}$ is

$$\dot{X} = -2\, G(M)X, \qquad M = XX^\top.$$

Therefore,

$$\dot{M} = \dot{X}X^\top + X\dot{X}^\top = -2\, G(M)XX^\top - 2\, XX^\top G(M) = -2\left[G(M)M + M\, G(M)\right],$$

where we used the symmetry of $G(M)$.

Using the projection identities, i.e., $P_i P_j = 0$ for any $i \neq j$, and $P_i^2 = P_i$ for any $i, j \in \{\star, \diamond, \circ\}$, we obtain the exchangeability:

$$G(M)M = u\alpha P_\star + v\beta P_\diamond + w\gamma P_\circ = M\, G(M).$$

Hence

$$\dot{M} = -4G(M)M = -4u\alpha P_\star - 4v\beta P_\diamond - 4w\gamma P_\circ.$$

Comparing this with $\dot{M} = \dot{\alpha}P_\star + \dot{\beta}P_\diamond + \dot{\gamma}P_\circ$ yields

$$\dot{\alpha} = -4u\alpha, \qquad \dot{\beta} = -4v\beta, \qquad \dot{\gamma} = -4w\gamma.$$

This proves that the trajectory remains in $\mathcal{M}$.

Finally, if $\alpha = a^2$, $\beta = b^2$ and $\gamma = c^2$, then $\dot{\alpha} = 2a\dot{a}$, $\dot{\beta} = 2b\dot{b}$ and $\dot{\gamma} = 2c\dot{c}$. Thus the above equations are equivalently written as

$$\dot{a} = -2ua, \qquad \dot{b} = -2vb, \qquad \dot{c} = -2wc.$$

The proof is complete. □

Lemma A.4 (Simplified Muon dynamics on the invariant manifold $\mathcal{X}$)**.** *Let*

$$\mathcal{X} := \left\{ X = \begin{bmatrix} aU & bV & cW \end{bmatrix} : a, b, c \in \mathbb{R} \right\},$$

where the columns of U, V, W are orthonormal and mutually orthogonal. For any $X \in \mathcal{X}$, we have

$$M := XX^\top = a^2 P_\star + b^2 P_\diamond + c^2 P_\circ \in \mathcal{M}.$$

Therefore,

$$G(M) = uP_\star + vP_\diamond + wP_\circ,$$

where

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{2}{n} \begin{bmatrix} a^2 - 1 \\ b^2 \\ c^2 \end{bmatrix} + \Sigma \begin{bmatrix} r(a^2 - 1) \\ rb^2 \\ rc^2 \end{bmatrix}.$$

Assume that $G(M)X$ has full column rank along the trajectory. Then the simplified Muon flow

$$\dot{X} = -\texttt{msign}\left[G(M)X\right]$$

leaves $\mathcal{X}$ invariant. More explicitly, the dynamics on $\mathcal{X}$ are given by

$$\begin{cases} \dot{a} = -\texttt{sign}(ua), \\ \dot{b} = -\texttt{sign}(vb), \\ \dot{c} = -\texttt{sign}(wc). \end{cases}$$

Proof. Let $X = [aU \;\; bV \;\; cW] \in \mathcal{X}$. Then, by orthogonality of the three subspaces,

$$M = XX^\top = a^2 P_\star + b^2 P_\diamond + c^2 P_\circ \in \mathcal{M}.$$

By the previous lemma A.3 on the invariant manifold $\mathcal{M}$, the matrix gradient satisfies

$$G(M) = uP_\star + vP_\diamond + wP_\circ,$$

with

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{2}{n} \begin{bmatrix} a^2 - 1 \\ b^2 \\ c^2 \end{bmatrix} + \Sigma \begin{bmatrix} r(a^2-1) \\ rb^2 \\ rc^2 \end{bmatrix}.$$

Using $P_\star U = U, P_\diamond V = V, P_\circ W = W$, and the mutual orthogonality of these subspaces, we obtain

$$G(M)X = (uP_\star + vP_\diamond + wP_\circ)\big[aU \;\; bV \;\; cW\big] = \big[uaU \;\; vbV \;\; wcW\big].$$

Thus $G(M)X$ lies in the tangent space of $\mathcal{X}$. Since $G(M)X$ is assumed to have full column rank,

$$\texttt{msign}\,(G(M)X) = G(M)X\left[X^\top G(M)G(M)X\right]^{-1/2}.$$

Moreover,

$$\big[X^\top G(M)G(M)X\big] = (ua)^2 I \oplus (vb)^2 I \oplus (wc)^2 I.$$

where $\oplus$ denotes a block-diagonal concatenation matching their column-blocks. Hence

$$\big[X^\top G(M)G(M)X\big]^{-1/2} = \frac{1}{|ua|} I \oplus \frac{1}{|vb|} I \oplus \frac{1}{|wc|} I,$$

and therefore

$$\begin{aligned} \texttt{msign}\,(G(M)X) &= \big[uaU \;\; vbV \;\; wcW\big]\left(\frac{1}{|ua|} I \oplus \frac{1}{|vb|} I \oplus \frac{1}{|wc|} I\right) \\ &= \big[\texttt{sign}(ua)U \;\; \texttt{sign}(vb)V \;\; \texttt{sign}(wc)W\big]. \end{aligned}$$

We thus can find the trajectory stays in $\mathcal{X}$ due to

$$\dot{X} = -\texttt{msign}(G(M)X).$$

Combining this with

$$X(t) = \big[a(t)U \;\; b(t)V \;\; c(t)W\big] \quad \Rightarrow \quad \dot{X}(t) = [\dot{a}(t)U \;\; \dot{b}(t)V \;\; \dot{c}(t)W]$$

and matching the coefficients along the mutually orthogonal blocks U, V, W, we get

$$\dot{a} = -\texttt{sign}(ua), \qquad \dot{b} = -\texttt{sign}(vb), \qquad \dot{c} = -\texttt{sign}(wc).$$

Therefore, $\dot{X}$ remains tangent to $\mathcal{X}$. The proof is complete. □

Lemma A.5 (Normalized GD and scaled Muon dynamics on $\mathcal{X}$)**.** *Let $X = [aU \; bV \; cW] \in \mathcal{X}$ and let u, v, w be as in Lemma A.4. Define the reduced factor-gradient and its Euclidean norm by*

$$z := [a, b, c]^\top, \qquad q(z) := [ua, vb, wc]^\top, \qquad \chi(z) := 2\|q(z)\|_2,$$

and define $s(z) := [\texttt{sign}(ua), \texttt{sign}(vb), \texttt{sign}(wc)]^\top$ and $m(z) := \|s(z)\|_2 = \sqrt{3}$. On every open sign cell where $\chi(z) > 0$ and $m(z) > 0$, the four direction–speed flows in Table 2 reduce, up to fixed rank factors absorbed into the time or step scale, to

$$\begin{aligned} &\text{GD} : \dot{z} = -2q(z), \qquad &&\text{NGD} : \dot{z} = -{}^{2q(z)}/_{\chi(z)}, \\ &\text{Muon} : \dot{z} = -s(z), \qquad &&\text{SMuon} : \dot{z} = -\left[{}^{\chi(z)}/_{m(z)}\right] s(z). \end{aligned} \tag{17}$$

Consequently, with τ denoting the time variable of the corresponding unscaled flow,

$$\frac{d\tau_{\text{NGD}}}{dt} = \frac{1}{\chi(z(t))}, \qquad \frac{d\tau_{\text{SMuon}}}{dt} = \frac{\chi(z(t))}{m(z(t))}, \tag{18}$$

and the continuous-time NGD trajectory is the GD trajectory reparameterized by time, whereas the continuous-time SMuon trajectory is the Muon trajectory reparameterized by time. At a switching surface, these identities hold in the Filippov sense.

Proof. The factor gradient is $\nabla_X \mathcal{L} = 2G(M)X$ and, on $\mathcal{X}$, its three blocks are $2uaU$, $2vbV$, and $2wcW$. Thus its reduced Euclidean coordinate vector is $2q(z)$ and its Frobenius norm is $\sqrt{r}\,\chi(z)$; the fixed factor $\sqrt{r}$ is absorbed into the stepsize. This gives the GD and NGD rows of (17). Lemma A.4 gives the Muon row. Normalizing the Muon polar factor gives the fixed factor $1/m(z)$, and multiplying it by the gradient norm gives the SMuon row. Each new vector field is a nonnegative scalar multiple of its parent vector field, which proves (18) and the claimed continuous-time trajectory equivalence. At $\chi = 0$ we use the standard zero-update convention; the normalized field is otherwise understood only on the open cells specified above. □

Remark A.6 (Diagonal special case: closed-form solutions of vanilla GD/simplified Muon)**.** *Recall that we have defined $\kappa := \frac{1}{r}\hat{K} = \frac{2}{n}I_3 + r\Sigma$. Assume in this remark that Σ is diagonal. Then $\hat{K}$ and κ are also diagonal; write $\kappa = \mathtt{Diag}(\kappa_{11}, \kappa_{22}, \kappa_{33})$. Since Σ is a covariance matrix in our model, its diagonal entries are nonnegative, and hence $\kappa_{11}, \kappa_{22}, \kappa_{33} > 0$. In this case,*

$$u = \kappa_{11}(\alpha - 1), \qquad v = \kappa_{22}\beta, \qquad w = \kappa_{33}\gamma.$$

Therefore the three coordinates are completely decoupled.

For vanilla GD, the reduced ODE becomes

$$\begin{cases} \dot{\alpha} = 4\kappa_{11}\alpha(1-\alpha), \\ \dot{\beta} = -4\kappa_{22}\beta^2, \\ \dot{\gamma} = -4\kappa_{33}\gamma^2. \end{cases} \tag{19}$$

For any initialization $\alpha_0, \beta_0, \gamma_0 > 0$, the solution is

$$\begin{cases} \alpha_{\mathrm{GD}}(t) = \frac{1}{1+((1-\alpha_0)/\alpha_0)e^{-4\kappa_{11}t}}, \\ \beta_{\mathrm{GD}}(t) = \frac{\beta_0}{1+4\kappa_{22}\beta_0 t}, \\ \gamma_{\mathrm{GD}}(t) = \frac{\gamma_0}{1+4\kappa_{33}\gamma_0 t}. \end{cases} \tag{20}$$

Therefore, the signal coordinate α follows a logistic-type trajectory and converges exponentially to 1, whereas the nuisance coordinates β and γ decay polynomially.

For simplified Muon, the reduced differential inclusion becomes

$$\begin{cases} \dot{\alpha} \in 2\sqrt{\alpha}\,\mathtt{sign}(\kappa_{11}(1-\alpha)) = -2\sqrt{\alpha}\,\mathtt{sign}(\alpha - 1), \\ \dot{\beta} \in -2\sqrt{\beta}\,\mathtt{sign}(\kappa_{22}\beta) = -2\sqrt{\beta}\,\mathtt{sign}(\beta), \\ \dot{\gamma} \in -2\sqrt{\gamma}\,\mathtt{sign}(\kappa_{33}\gamma) = -2\sqrt{\gamma}\,\mathtt{sign}(\gamma), \end{cases} \tag{21}$$

if $\alpha = a^2$, $\beta = b^2$ and $\gamma = c^2$. Under the natural sticking selection at the switching surfaces $\alpha = 1$, $\beta = 0$, and $\gamma = 0$, if $\alpha_0 > 1$ and $\beta_0, \gamma_0 > 0$, then

$$\begin{cases} \alpha_{\mathrm{Muon}}(t) = \left(\max\left\{\sqrt{\alpha_0} - t,\, 1\right\}\right)^2, \\ \beta_{\mathrm{Muon}}(t) = \left(\max\left\{\sqrt{\beta_0} - t,\, 0\right\}\right)^2, \\ \gamma_{\mathrm{Muon}}(t) = \left(\max\left\{\sqrt{\gamma_0} - t,\, 0\right\}\right)^2. \end{cases} \tag{22}$$

If instead $0 < \alpha_0 < 1$, the first line is replaced by

$$\alpha_{\mathrm{Muon}}(t) = \left(\min\left\{\sqrt{\alpha_0} + t,\, 1\right\}\right)^2.$$

Therefore, Muon reaches the target river coordinate $\alpha = 1$ and removes the nuisance coordinates β and γ in finite continuous time with linear speed.

Remark A.7 (Diagonal special case: NGD and SMuon)**.** *In the diagonal case of Remark A.6, let*

$$\chi(\alpha, \beta, \gamma) := 2\sqrt{\kappa_{11}^2\alpha(\alpha-1)^2 + \kappa_{22}^2\beta^3 + \kappa_{33}^2\gamma^3}.$$

For positive coordinates, the normalized GD flow is

$$\begin{cases} \dot{\alpha}_{\mathrm{NGD}} = -4\kappa_{11}\alpha(\alpha-1)/\chi(\alpha,\beta,\gamma), \\ \dot{\beta}_{\mathrm{NGD}} = -4\kappa_{22}\beta^2/\chi(\alpha,\beta,\gamma), \\ \dot{\gamma}_{\mathrm{NGD}} = -4\kappa_{33}\gamma^2/\chi(\alpha,\beta,\gamma). \end{cases} \tag{23}$$

Its fixed-step factor update is

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix}_{k+1}^{\mathrm{NGD}} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}_{k}^{\mathrm{NGD}} - \frac{2\eta}{\chi_k} \begin{bmatrix} u_k a_k \\ v_k b_k \\ w_k c_k \end{bmatrix}, \qquad \chi_k = \chi(\alpha_k, \beta_k, \gamma_k).$$

Hence, if $(\alpha_{\mathrm{GD}}(\tau), \beta_{\mathrm{GD}}(\tau), \gamma_{\mathrm{GD}}(\tau))$ *denotes the closed-form GD solution in* (20)*, then*

$$(\alpha_{\mathrm{NGD}}(t), \beta_{\mathrm{NGD}}(t), \gamma_{\mathrm{NGD}}(t)) = (\alpha_{\mathrm{GD}}(\tau(t)), \beta_{\mathrm{GD}}(\tau(t)), \gamma_{\mathrm{GD}}(\tau(t))),$$
$$t = \int_0^{\tau(t)} \chi(\alpha_{\mathrm{GD}}(s), \beta_{\mathrm{GD}}(s), \gamma_{\mathrm{GD}}(s))\, ds.$$

Therefore, the coordinate path is identical to vanilla GD. Its unit-speed parameterization removes the vanishing gradient scale: the arc length to the solution is finite, and a fixed discrete step can overshoot once the residual gradient is small. In particular, NGD has the vanilla GD geometry but the Muon-like constant-speed late-stage sensitivity.

For SMuon, we have $m(z) = m(a, b, c) = \sqrt{3}$ *on a sign cell where all three blocks are active (the harmless rank-dependent normalization is absorbed into* η*). The reduced flow is*

$$\begin{cases} \dot{\alpha}_{\mathrm{SMuon}} = & -\frac{2\chi(\alpha,\beta,\gamma)}{m(z)} \sqrt{\alpha}\, \mathtt{sign}(\kappa_{11}(\alpha - 1)), \\ \dot{\beta}_{\mathrm{SMuon}} = & -\frac{2\chi(\alpha,\beta,\gamma)}{m(z)} \sqrt{\beta}\, \mathtt{sign}(\kappa_{22}\beta), \\ \dot{\gamma}_{\mathrm{SMuon}} = & -\frac{2\chi(\alpha,\beta,\gamma)}{m(z)} \sqrt{\gamma}\, \mathtt{sign}(\kappa_{33}\gamma). \end{cases} \tag{24}$$

Its fixed-step factor update is

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix}_{k+1}^{\mathrm{SMuon}} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}_{k}^{\mathrm{SMuon}} - \eta \frac{\chi_k}{m_k} \begin{bmatrix} \mathtt{sign}(u_k a_k) \\ \mathtt{sign}(v_k b_k) \\ \mathtt{sign}(w_k c_k) \end{bmatrix}.$$

Equivalently, with the Muon closed-form path from (22)*,*

$$(\alpha_{\mathrm{SMuon}}(t), \beta_{\mathrm{SMuon}}(t), \gamma_{\mathrm{SMuon}}(t)) = (\alpha_{\mathrm{Muon}}(\tau(t)), \beta_{\mathrm{Muon}}(\tau(t)), \gamma_{\mathrm{Muon}}(\tau(t))),$$
$$\frac{dt}{d\tau} = \frac{\sqrt{3}}{\chi(\alpha_{\mathrm{Muon}}(\tau), \beta_{\mathrm{Muon}}(\tau), \gamma_{\mathrm{Muon}}(\tau))}. \tag{25}$$

Hence, SMuon keeps simplified Muon's straight, equal-coordinate path while restoring a residual-dependent speed. The Muon path reaches $(1, 0, 0)$ *in finite* τ*, but* $\chi \to 0$ *there and the physical SMuon time diverges (logarithmically under the generic signal-dominated approach), so the continuous flow converges asymptotically rather than by a finite jump. This is the diagonal manifestation of the direction–speed decomposition: NGD changes only the clock of GD, whereas SMuon changes only the clock of Muon and supplies late-stage damping.*

A.3 River–Valley Structure of the Reduced Mixed Spectrum Sensing Model

A.3.1 Additional Definitions, Assumptions, and Basic Geometry

Under the MSMS construction, the scalar dynamics induced on the invariant manifolds $\mathcal{M}$ and $\mathcal{X}$ admit an exact river–valley decomposition [Wen et al., 2025]. We later define the associated river and hill quantities, including coordinates, forces, and energies based on a reduced river $\mathcal{R}$.

Definition A.8 (River-hill coordinates)**.** *We reparameterize the reduced coordinates* (α, β, γ) *by*

$$\theta := \alpha - 1 \in \mathbb{R}, \qquad \omega := [\beta, \gamma]^\top \in \mathbb{R}^2. \tag{26}$$

Here, θ *is the river variable, measuring progress toward the target signal component, while* ω *collects the two hill variables, corresponding to the transverse spike and bulk components.*

Lemma A.9 (Reduced loss on invariant manifold)**.** *Let*

$$\mathcal{F}(M) = \frac{1}{n} \|M - P_\star\|_F^2 + \frac{1}{2} \begin{bmatrix} \langle P_\star, M - P_\star \rangle \\ \langle P_\diamond, M \rangle \\ \langle P_\circ, M \rangle \end{bmatrix}^\top \Sigma \begin{bmatrix} \langle P_\star, M - P_\star \rangle \\ \langle P_\diamond, M \rangle \\ \langle P_\circ, M \rangle \end{bmatrix},$$

and restrict M to the invariant manifold, i.e., $M = \alpha P_\star + \beta P_\diamond + \gamma P_\circ$. Then the reduced loss is

$$f(\alpha,\beta,\gamma) = \frac{r}{n}\left[(\alpha-1)^2+\beta^2+\gamma^2\right] + \frac{r^2}{2}\begin{bmatrix}\alpha-1\\ \beta\\ \gamma\end{bmatrix}^\top \Sigma \begin{bmatrix}\alpha-1\\ \beta\\ \gamma\end{bmatrix}$$
$$= \frac{1}{2}\begin{bmatrix}\alpha-1\\ \beta\\ \gamma\end{bmatrix}^\top \hat{K} \begin{bmatrix}\alpha-1\\ \beta\\ \gamma\end{bmatrix}, \qquad \textit{where } \hat{K} := \frac{2r}{n} I_3 + r^2\Sigma.$$

Definition A.10 (River-hill splitting of the reduced curvature)**.** *The above coordinate decomposition induces a natural block splitting of the reduced curvature matrix*

$$\hat{K} := \frac{2r}{n} I_3 + r^2\Sigma = \begin{bmatrix}\hat{a} & \hat{b}^\top\\ \hat{c} & \hat{D}\end{bmatrix} \in \mathbb{S}^3, \tag{27}$$

where $\hat{a} \in \mathbb{R}$, $\hat{b}, \hat{c} \in \mathbb{R}^2$, and $\hat{D} \in \mathbb{S}^2$. Since $\hat{K}$ is symmetric, $\hat{b} = \hat{c}$. Moreover, $\hat{D}$ is a principal submatrix of the positive semidefinite matrix $\hat{K}$, and hence $\hat{D} \succeq 0$. Throughout this section, we assume that the hill block is nondegenerate:

$$\hat{D} \succ 0. \tag{28}$$

We denote its extremal eigenvalues by

$$\mu := \lambda_{\min}(\hat{D}) > 0, \qquad \nu := \lambda_{\max}(\hat{D}) > 0. \tag{29}$$

In river-hill coordinates, the reduced loss $f(\alpha,\beta,\gamma)$ becomes

$$f(\theta,\omega) = \frac{1}{2}\begin{bmatrix}\theta\\ \omega\end{bmatrix}^\top \begin{bmatrix}\hat{a} & \hat{c}^\top\\ \hat{c} & \hat{D}\end{bmatrix}\begin{bmatrix}\theta\\ \omega\end{bmatrix}.$$

Definition A.11 (River-hill forces)**.** *For any reduced coordinates $(\theta,\omega) \in \mathbb{R}\times\mathbb{R}^2$, define*

$$\phi := \hat{a}\theta + \hat{b}^\top\omega \in \mathbb{R}, \qquad \psi := \hat{c}\theta + \hat{D}\omega \in \mathbb{R}^2, \tag{30}$$

as the river force and hill force, respectively.

Equivalently, (ϕ,ψ) is the block decomposition of the reduced gradient:

$$\nabla f(\theta,\omega) = \begin{bmatrix}\phi\\ \psi\end{bmatrix}.$$

Since $\hat{D} \succ 0$, the hill variable can be expressed in terms of the hill force as

$$\omega = -\theta\hat{D}^{-1}\hat{c} + \hat{D}^{-1}\psi. \tag{31}$$

Definition A.12 (River-hill energies)**.** *The reduced loss can be written as*

$$f(\theta,\omega) = \frac{1}{2}\left(\hat{a}\theta^2 + 2\theta\hat{c}^\top\omega + \omega^\top\hat{D}\omega\right).$$

Using the hill force $\psi = \hat{c}\theta + \hat{D}\omega$, we have

$$\psi^\top\hat{D}^{-1}\psi = \theta^2\hat{c}^\top\hat{D}^{-1}\hat{c} + 2\theta\hat{c}^\top\omega + \omega^\top\hat{D}\omega.$$

Therefore,

$$f(\theta,\omega) = \frac{1}{2}\left[\hat{a}_{\text{eff}}\theta^2 + \psi^\top\hat{D}^{-1}\psi\right],$$

where the effective river curvature is defined by

$$\hat{a}_{\text{eff}} := \hat{a} - \hat{c}^\top\hat{D}^{-1}\hat{c}. \tag{32}$$

We assume throughout that

$$\hat{a}_{\text{eff}} > 0. \tag{33}$$

This exact decomposition motivates the river and hill energies

$$E_k := \hat{a}_{\text{eff}}\theta_k^2, \qquad H_k := \psi_k^\top\hat{D}^{-1}\psi_k, \tag{34}$$

hence $H_k = O(\eta^2)$ means that the iterate lies inside an $\psi_k = O(\eta)$ vicinity (so we name it as $O(\eta)$ hill-force tube) around the river.

Definition A.13 (Reduced river)**.** *The reduced river is the zero-hill-energy (zero-hill-force) manifold*

$$\mathcal{R} := \{(\theta, \omega) : H = 0\} = \{(\theta, \omega) : \psi = 0\} = \{(\theta, \omega) : \omega = -\hat{D}^{-1}\hat{c}\,\theta\}. \tag{35}$$

Furthermore, since $\hat{b} = \hat{c}$, the river force satisfies

$$\phi = \hat{a}\theta + \hat{c}^\top\omega = \hat{a}_{\text{eff}}\theta + \hat{c}^\top\hat{D}^{-1}\psi. \tag{36}$$

Consequently, along the river manifold $\mathcal{R}$, where $\psi = 0$, we have $\phi = \hat{a}_{\text{eff}}\theta$. Thus, once the hill force has vanished, the reduced dynamics are governed solely by the effective river curvature $\hat{a}_{\text{eff}}$.

Remark A.14 (A brief notation summary)**.** *We summarize the notation used in this section and their relationships in Table 3. We measure river progress by the coordinate θ, because after eliminating the transverse variables, θ parametrizes the zero-hill-force valley floor and the remaining loss is $\hat{a}_{\text{eff}}\theta^2$ (see* (34)*). In contrast, the raw hill variable ω is not a genuine transverse error when $\hat{c} \neq 0$, since the coupled river itself has $\omega = -\hat{D}^{-1}\hat{c}\theta$; hence the intrinsic hill energy must be measured by the hill force $\psi = \hat{c}\theta + \hat{D}\omega$, or equivalently by the squared $\hat{D}$-metric distance to the zero-hill-force manifold.*

Table 3: Variables, forces, energies, and reduced discrete iterations in the general PSD case.

River Item	River Notation	Hill Item	Hill Notation
River variable	$\theta_k = \alpha_k - 1$	Hill variable	$\omega_k = [\beta_k, \gamma_k]^\top$
River force	$\phi_k = \hat{a}\theta_k + \hat{b}^\top\omega_k$	Hill force	$\psi_k = \hat{c}\theta_k + \hat{D}\omega_k$
River energy	$E_k = \hat{a}_{\text{eff}}\theta_k^2$	Hill energy	$H_k = \psi_k^\top\hat{D}^{-1}\psi_k$

For the subsequent local estimates, we restrict the dynamics to the compact box

$$K_{\underline{\alpha},\overline{\alpha},\underline{\omega},\overline{\omega}} := \left\{(\theta, \omega) \in \mathbb{R} \times \mathbb{R}^2 : \underline{\alpha} \leq 1 + \theta \leq \overline{\alpha},\ \underline{\omega}\mathbf{1}_2 \preceq \omega \preceq \overline{\omega}\mathbf{1}_2\right\}, \tag{37}$$

where $\mathbf{1}_2 = (1, 1)^\top$, and $\preceq$ denotes componentwise comparison.

A.3.2 Gradient Descent Reduced Dynamics

The reduced GD iteration follows readily from applying an Euler discretization to Lemma A.3:

$$\begin{cases}\theta_{k+1} = \theta_k - 4\eta(1 + \theta_k)\phi_k, \\ \omega_{k+1} = \omega_k - 4\eta\texttt{Diag}(\omega_k)\psi_k.\end{cases} \tag{38}$$

Proposition A.15 (Positivity preservation for reduced GD)**.** *Suppose that at iteration k we have $1 + \theta_k > 0$ and $\omega_k \succ 0$. Assume further that the step size $\eta > 0$ is chosen sufficiently small so that, for all k,*

$$4\eta|\phi_k| < 1, \qquad 4\eta\psi_k \prec \begin{bmatrix}1 \\ 1\end{bmatrix}.$$

Then, at iteration $k + 1$, it holds that $1 + \theta_{k+1} > 0$ and $\omega_{k+1} \succ 0$.

Proof. From 38, the river-coordinate update can be written as

$$1 + \theta_{k+1} = (1 + \theta_k)(1 - 4\eta\phi_k).$$

Since $1 + \theta_k > 0$ and $4\eta|\phi_k| < 1$, we have $1 - 4\eta\phi_k > 0$, and hence the right-hand side is positive. For each hill coordinate $i \in \{1, 2\}$,

$$(\omega_{k+1})_i = (\omega_k)_i\big(1 - 4\eta(\psi_k)_i\big), \qquad i = 1, 2.$$

Because $(\omega_k)_i > 0$ and $4\eta(\psi_k)_i < 1$, it follows that $1 - 4\eta(\psi_k)_i > 0$, and therefore $(\omega_{k+1})_i > 0$ for $i = 1, 2$. □

Proposition A.16 (Hill-energy estimate for reduced GD)**.** *Assume that the GD trajectory remains in the compact box $K_{\underline{\alpha},\overline{\alpha},\underline{\omega},\overline{\omega}}$. If $0 < \eta \leq \frac{\underline{\omega}}{8\nu\overline{\omega}^2}$, then there exists a constant $C_{\text{GD}} > 0$, depending only on $\underline{\alpha}, \overline{\alpha}, \underline{\omega}, \overline{\omega}, \hat{D}, \hat{c}$ and η, such that*

$$H_{k+1} \leq (1 - 2\underline{\omega}\mu\eta)H_k + C_{\text{GD}}\eta\,\phi_k^2. \tag{39}$$

Moreover, one admissible choice is

$$C_{\text{GD}} = \frac{8\overline{\alpha}^2\|\hat{D}^{-1}\hat{c}\|_2^2}{\underline{\omega}} + 32\overline{\alpha}^2\eta\,\hat{c}^\top\hat{D}^{-1}\hat{c}.$$

Proof. By definition, $\psi_{k+1} = \hat{c}\theta_{k+1} + \hat{D}\omega_{k+1}$. Using 38, we obtain

$$\begin{aligned}\psi_{k+1} &= \hat{c}\big[\theta_k - 4\eta(1+\theta_k)\phi_k\big] + \hat{D}\big[\omega_k - 4\eta\mathtt{Diag}(\omega_k)\psi_k\big] \\ &= \psi_k - 4\eta\left[\hat{D}\mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\hat{c}\,\phi_k\right].\end{aligned}$$

Define $r_k := \hat{D}\mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\phi_k\hat{c}$. Then $\psi_{k+1} = \psi_k - 4\eta r_k$, and therefore

$$H_{k+1} - H_k = -8\eta\,\psi_k^\top \hat{D}^{-1} r_k + 16\eta^2 r_k^\top \hat{D}^{-1} r_k.$$

For the linear term, we have

$$\psi_k^\top \hat{D}^{-1} r_k = \psi_k^\top \mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\phi_k\,\psi_k^\top \hat{D}^{-1}\hat{c} \geq \underline{\omega}\|\psi_k\|_2^2 - |1+\theta_k|\,|\phi_k|\,|\psi_k^\top \hat{D}^{-1}\hat{c}|,$$

and hence

$$-8\eta\,\psi_k^\top \hat{D}^{-1} r_k \leq -8\underline{\omega}\eta\|\psi_k\|_2^2 + 8\eta|1+\theta_k|\,|\phi_k|\,|\psi_k^\top \hat{D}^{-1}\hat{c}|.$$

For the quadratic term, note that

$$\begin{aligned}r_k^\top \hat{D}^{-1} r_k =& \big[\hat{D}\mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\hat{c}\,\phi_k\big]^\top \hat{D}^{-1}\big[\hat{D}\mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\hat{c}\,\phi_k\big] \\ =& \big\|\hat{D}^{1/2}\mathtt{Diag}(\omega_k)\psi_k + (1+\theta_k)\phi_k\,\hat{D}^{-1/2}\hat{c}\big\|_2^2 \\ \leq& 2\|\hat{D}^{1/2}\mathtt{Diag}(\omega_k)\psi_k\|_2^2 + 2(1+\theta_k)^2\phi_k^2\|\hat{D}^{-1/2}\hat{c}\|_2^2 \\ =& 2\psi_k^\top \mathtt{Diag}(\omega_k)\hat{D}\mathtt{Diag}(\omega_k)\psi_k + 2(1+\theta_k)^2\phi_k^2\,\hat{c}^\top \hat{D}^{-1}\hat{c} \\ \leq& 2\nu\overline{\omega}^2\|\psi_k\|_2^2 + 2\overline{\alpha}^2\phi_k^2\,\hat{c}^\top \hat{D}^{-1}\hat{c}.\end{aligned}$$

Consequently,

$$16\eta^2 r_k^\top \hat{D}^{-1} r_k \leq 32\nu\overline{\omega}^2\eta^2\|\psi_k\|_2^2 + 32\overline{\alpha}^2\eta^2\hat{c}^\top \hat{D}^{-1}\hat{c}\,\phi_k^2.$$

Applying Young's inequality to the mixed term yields

$$8\eta|1+\theta_k|\,|\phi_k|\,|\psi_k^\top \hat{D}^{-1}\hat{c}| \leq 2\underline{\omega}\eta\|\psi_k\|_2^2 + \frac{8\overline{\alpha}^2\|\hat{D}^{-1}\hat{c}\|_2^2}{\underline{\omega}}\eta\phi_k^2.$$

Combining these estimates we get

$$H_{k+1} - H_k \leq \Big(-8\underline{\omega}\eta + 2\underline{\omega}\eta + 32\nu\overline{\omega}^2\eta^2\Big)\|\psi_k\|_2^2 + \left(\frac{8\overline{\alpha}^2\|\hat{D}^{-1}\hat{c}\|_2^2}{\underline{\omega}}\eta + 32\overline{\alpha}^2\eta^2\hat{c}^\top \hat{D}^{-1}\hat{c}\right)\phi_k^2.$$

Since $0 < \eta \leq \underline{\omega}/(8\nu\overline{\omega}^2)$, we have $32\nu\overline{\omega}^2\eta^2 \leq 4\underline{\omega}\eta$, and thus

$$H_{k+1} - H_k \leq -2\underline{\omega}\eta\|\psi_k\|_2^2 + C_{\mathrm{GD}}\eta\,\phi_k^2,$$

where one may take

$$C_{\mathrm{GD}} = \frac{8\overline{\alpha}^2\|\hat{D}^{-1}\hat{c}\|_2^2}{\underline{\omega}} + 32\overline{\alpha}^2\eta\,\hat{c}^\top \hat{D}^{-1}\hat{c}.$$

Finally, using $\|\psi_k\|_2^2 \geq \mu H_k$, we conclude that

$$H_{k+1} \leq (1 - 2\underline{\omega}\mu\eta)H_k + C_{\mathrm{GD}}\eta\,\phi_k^2.$$

The proof is complete. □

Corollary A.17 (Tail bound for the GD hill energy)**.** *Under the assumptions of the preceding proposition, define $\rho_{\mathrm{GD}} := 1 - 2\underline{\omega}\mu\eta \in (0,1)$. For any $N \geq 0$, let $\Phi_N := \sup_{j\geq N}|\phi_j|$. Then, for every $k \geq N$,*

$$H_k \leq \rho_{\mathrm{GD}}^{k-N} H_N + \frac{C_{\mathrm{GD}}}{2\underline{\omega}\mu}\Phi_N^2. \tag{40}$$

In particular, if $\phi_k \to 0$, then $H_k \to 0$ and hence $\psi_k \to 0$.

Proof. Iterating (39) from N to $k-1$ yields

$$H_k \le \rho_{\mathrm{GD}}^{k-N} H_N + C_{\mathrm{GD}}\eta \sum_{j=N}^{k-1} \rho_{\mathrm{GD}}^{k-1-j}\phi_j^2 \le \rho_{\mathrm{GD}}^{k-N} H_N + C_{\mathrm{GD}}\eta\Phi_N^2 \sum_{\iota=0}^{k-1-N} \rho_{\mathrm{GD}}^{\iota}$$
$$\le \rho_{\mathrm{GD}}^{k-N} H_N + \frac{C_{\mathrm{GD}}\eta}{1-\rho_{\mathrm{GD}}}\Phi_N^2.$$

Since $1-\rho_{\mathrm{GD}} = 2\underline{\omega}\mu\eta$, this implies (40). Finally, if $\phi_k \to 0$, then $\Phi_N \to 0$ as $N \to \infty$, and therefore $H_k \to 0$. □

Remark A.18 (A Lyapunov perspective on GD)**.** *The GD dynamics admit a natural interpretation from the viewpoint of Lyapunov dissipation. In the region $\alpha > 0$, $\beta \ge 0$, and $\gamma \ge 0$, consider*

$$V_{\mathrm{GD}}(\alpha,\beta,\gamma) := r(\alpha - 1 - \ln\alpha) + r\beta + r\gamma. \tag{41}$$

Defining $x := [\alpha,\beta,\gamma]^\top$, $e_1 := [1,0,0]^\top$, $\hat{k} := \hat{K}e_1$, a direct differentiation along the GD flow gives

$$\dot{V}_{\mathrm{GD}} = r\left(1-\frac{1}{\alpha}\right)\dot{\alpha} + r\dot{\beta} + r\dot{\gamma} = 4(x-e_1)^\top \mathtt{Diag}(r,r,r)\,(\hat{k}-\hat{K}x)$$
$$= -4(x-e_1)^\top \mathtt{Diag}(r,r,r)\hat{K}(x-e_1) \le 0.$$

Thus V_{GD} is a smooth Lyapunov function for the GD continuous-time dynamics, and its decay is governed by a quadratic dissipation mechanism around the equilibrium $x = e_1$, under the corresponding positivity assumptions on the above quadratic form.

This provides a useful conceptual contrast with the Muon dynamics. For GD, the evolution is compatible with a smooth energy landscape and a correspondingly smooth dissipation law. By contrast, the sign-based Muon update is intrinsically nonsmooth, and the quantity (47) *does not play the role of a comparably smooth Lyapunov function. This helps explain why GD typically admits cleaner convergence behavior, whereas Muon more naturally exhibits switching effects and $O(\eta)$-scale residual oscillations under a constant step size.*

Proposition A.19 (GD river progress after hill reduction)**.** *Assume that there exist $N_{\mathrm{GD}} \ge 0$ and $R_{\mathrm{GD}} > 0$ such that $H_k \le R_{\mathrm{GD}}^2\eta^2$ for all $k \ge N_{\mathrm{GD}}$. In particular, the hill forcing ψ_k is of order $O(\eta)$ for all $k \ge N_{\mathrm{GD}}$. Suppose further that $\eta \le \frac{1}{6\overline{\alpha}\hat{a}_{\mathrm{eff}}}$, and that η is sufficiently small so that for all $k \ge N_{\mathrm{GD}}$,*

$$|\theta_k| \ge \max\left\{\frac{2(\hat{c}^\top\hat{D}^{-1}\hat{c})^{1/2}}{\hat{a}_{\mathrm{eff}}}, \sqrt{\frac{\Gamma_{\mathrm{GD}}}{\hat{a}_{\mathrm{eff}}}}\right\} R_{\mathrm{GD}}\eta, \qquad \Gamma_{\mathrm{GD}} \gg 1. \tag{42}$$

Then the river variable θ_k contracts geometrically:

$$|\theta_{k+1}| \le (1 - 2\underline{\alpha}\hat{a}_{\mathrm{eff}}\eta)|\theta_k|. \tag{43}$$

Proof. By (36), we have $\phi_k - \hat{a}_{\mathrm{eff}}\theta_k = \hat{c}^\top\hat{D}^{-1}\psi_k$. Applying the Cauchy-Schwarz inequality in the $\hat{D}^{-1}$-inner product yields

$$|\hat{c}^\top\hat{D}^{-1}\psi_k| \le \left(\hat{c}^\top\hat{D}^{-1}\hat{c}\right)^{1/2}\left(\psi_k^\top\hat{D}^{-1}\psi_k\right)^{1/2} = \left(\hat{c}^\top\hat{D}^{-1}\hat{c}\right)^{1/2}\sqrt{H_k}.$$

Using $H_k \le R_{\mathrm{GD}}^2\eta^2$, we obtain $|\phi_k - \hat{a}_{\mathrm{eff}}\theta_k| \le (\hat{c}^\top\hat{D}^{-1}\hat{c})^{1/2}R_{\mathrm{GD}}\eta$. If (42) holds, then

$$|\phi_k - \hat{a}_{\mathrm{eff}}\theta_k| \le \frac{\hat{a}_{\mathrm{eff}}}{2}|\theta_k| \qquad \text{and} \qquad E_k = \hat{a}_{\mathrm{eff}}\theta_k^2 \ge \Gamma_{\mathrm{GD}}R_{\mathrm{GD}}^2\eta^2 \gg H_k.$$

Consequently, the iterate is in a river-dominant phase, ϕ_k has the same sign as θ_k, and

$$\frac{\hat{a}_{\mathrm{eff}}}{2}|\theta_k| \le |\phi_k| \le \frac{3\hat{a}_{\mathrm{eff}}}{2}|\theta_k|. \tag{44}$$

Recall that the GD river update is $\theta_{k+1} = \theta_k - 4\eta(1+\theta_k)\phi_k$. Using $1+\theta_k \ge \underline{\alpha}$ and the lower bound in (44), the update decreases $|\theta_k|$ by at least $2\underline{\alpha}\hat{a}_{\mathrm{eff}}\eta|\theta_k|$, provided that no overshoot occurs. To preclude overshoot, we use the upper bound in (44) and $\eta \le \frac{1}{6\overline{\alpha}\hat{a}_{\mathrm{eff}}}$ to obtain

$$4\eta(1+\theta_k)|\phi_k| \le 6\eta\overline{\alpha}\hat{a}_{\mathrm{eff}}|\theta_k| \le |\theta_k|.$$

Hence overshoot does not occur, and (43) follows. □

Remark A.20. *Proposition A.19 shows that once the hill forcing ψ_k is reduced to order $O(\eta)$ (e.g., via (39)) and river energy dominates hill energy, the GD river variable θ_k contracts geometrically according to (43) until it reaches an $O(\eta)$-neighborhood of the river equilibrium $\theta = 0$. Consequently, the number of GD iterations required to reduce $|\theta_k|$ from a constant scale to an $O(\eta)$ scale is*

$$O\left(\frac{1}{\eta}\log\frac{1}{\eta}\right).$$

Moreover, in practice GD may drive the hill forcing ψ_k below the $O(\eta)$ level (i.e., achieve higher accuracy), whereas the corresponding simplified Muon analysis typically guarantees only an $O(\eta)$-level transient bound. We therefore focus on the $O(\eta)$ regime to enable a like-for-like comparison with simplified Muon.

A.3.3 Simplified Muon Reduced Dynamics

The reduced Muon iteration follows readily from applying an Euler discretization to Lemma A.4:

$$\begin{cases}\theta_{k+1} = \theta_k - 2\eta\sqrt{1+\theta_k}\, s_k, & s_k \in \mathtt{sign}(\phi_k),\\ \omega_{k+1} = \omega_k - 2\eta\mathtt{Diag}(\omega_k)^{1/2} t_k, & t_k \in \mathtt{sign}(\psi_k).\end{cases} \tag{45}$$

Proposition A.21 (Positivity preservation for reduced Muon)**.** *Assume that $\eta > 0$ is sufficiently small such that, for all k,*

$$1+\theta_k > 4\eta^2, \qquad \omega_k \succ 4\eta^2 \begin{bmatrix}1\\1\end{bmatrix}.$$

Then, for any choices $s_k \in \mathtt{Sign}(\phi_k)$ and $t_k \in \mathtt{Sign}(\psi_k)$, it holds that

$$1+\theta_{k+1} > 0, \qquad \omega_{k+1} \succ 0.$$

Proof. Since $s_k \in [-1,1]$, we have

$$\begin{aligned}1+\theta_{k+1} = 1+\theta_k - 2\eta\sqrt{1+\theta_k}\, s_k =& \sqrt{1+\theta_k}(\sqrt{1+\theta_k} - 2\eta s_k)\\ \geq& \sqrt{1+\theta_k}(\sqrt{1+\theta_k} - 2\eta) > 0.\end{aligned}$$

Similarly, for each $i \in \{1,2\}$,

$$\begin{aligned}(\omega_{k+1})_i = (\omega_k)_i - 2\eta\sqrt{(\omega_k)_i}\,(t_k)_i =& \sqrt{(\omega_k)_i}\left[\sqrt{(\omega_k)_i} - 2\eta(t_k)_i\right]\\ \geq& \sqrt{(\omega_k)_i}\left[\sqrt{(\omega_k)_i} - 2\eta\right] > 0.\end{aligned}$$

This completes the proof. □

Proposition A.22 (Hill-energy estimate for reduced Muon)**.** *Assume that the Muon orbit remains in the compact box $K_{\underline{\alpha},\overline{\alpha},\underline{\omega},\overline{\omega}}$. Define the constants $B_0 := \sqrt{\underline{\omega}\underline{\mu}} - \sqrt{\overline{\alpha}}\,(\hat{c}^\top \hat{D}^{-1}\hat{c})^{1/2}$ and $B_1 := 8\overline{\alpha}\hat{c}^\top\hat{D}^{-1}\hat{c} + 16\nu\overline{\omega}$. Then*

$$H_{k+1} \leq H_k - 4B_0\eta\sqrt{H_k} + B_1\eta^2. \tag{46}$$

Proof. By definition, $\psi_{k+1} = \hat{c}\theta_{k+1} + \hat{D}\omega_{k+1}$. Using 45, we obtain

$$\begin{aligned}\psi_{k+1} &= \hat{c}\left[\theta_k - 2\eta\sqrt{1+\theta_k}\, s_k\right] + \hat{D}\left[\omega_k - 2\eta\mathtt{Diag}(\omega_k)^{1/2}t_k\right]\\ &= \psi_k - 2\eta\left[\hat{c}\sqrt{1+\theta_k}\, s_k + \hat{D}\mathtt{Diag}(\omega_k)^{1/2}t_k\right].\end{aligned}$$

Let $r_k := \hat{c}\sqrt{1+\theta_k}\, s_k + \hat{D}\mathtt{Diag}(\omega_k)^{1/2}t_k$. Then $\psi_{k+1} = \psi_k - 2\eta r_k$, and hence

$$H_{k+1} - H_k = -4\eta\,\psi_k^\top\hat{D}^{-1}r_k + 4\eta^2 r_k^\top\hat{D}^{-1}r_k.$$

For the linear term, we can write

$$\begin{aligned}&\psi_k^\top\hat{D}^{-1}r_k = \sqrt{1+\theta_k}\, s_k\psi_k^\top\hat{D}^{-1}\hat{c} + \psi_k^\top\mathtt{Diag}(\omega_k)^{1/2}t_k\\ =&\sqrt{1+\theta_k}\, s_k\psi_k^\top\hat{D}^{-1}\hat{c} + \sum_{i=1}^{2}\sqrt{(\omega_k)_i}\,|(\psi_k)_i| \geq -\sqrt{\overline{\alpha}}\left(\hat{c}^\top\hat{D}^{-1}\hat{c}\right)^{1/2}\sqrt{H_k} + \sqrt{\underline{\omega}}\|\psi_k\|_2.\end{aligned}$$

Since $\|\psi_k\|_2^2 \geq \mu H_k$, it follows that $\psi_k^\top \hat{D}^{-1} r_k \geq B_0 \sqrt{H_k}$.

For the quadratic term, using $(u+v)^\top \hat{D}^{-1}(u+v) \leq 2u^\top \hat{D}^{-1}u + 2v^\top \hat{D}^{-1}v$, we obtain

$$\begin{aligned}
&r_k^\top \hat{D}^{-1} r_k = \big[\hat{c}\sqrt{1+\theta_k}\, s_k + \hat{D}\mathtt{Diag}(\omega_k)^{1/2} t_k\big]^\top \hat{D}^{-1}\big[\hat{c}\sqrt{1+\theta_k}\, s_k + \hat{D}\mathtt{Diag}(\omega_k)^{1/2} t_k\big] \\
\leq& 2\big(\hat{c}\sqrt{1+\theta_k}\, s_k\big)^\top \hat{D}^{-1}\big(\hat{c}\sqrt{1+\theta_k}\, s_k\big) + 2\big(\hat{D}\mathtt{Diag}(\omega_k)^{1/2} t_k\big)^\top \hat{D}^{-1}\big(\hat{D}\mathtt{Diag}(\omega_k)^{1/2} t_k\big) \\
=& 2(1+\theta_k) s_k^2\, \hat{c}^\top \hat{D}^{-1}\hat{c} + 2\, t_k^\top \mathtt{Diag}(\omega_k)^{1/2} \hat{D} \mathtt{Diag}(\omega_k)^{1/2} t_k \\
\leq& 2\overline{\alpha}\hat{c}^\top \hat{D}^{-1}\hat{c} + 2\nu \|\mathtt{Diag}(\omega_k)^{1/2} t_k\|_2^2 \\
=& 2\overline{\alpha}\hat{c}^\top \hat{D}^{-1}\hat{c} + 2\nu \sum_{i=1}^{2} (\omega_k)_i (t_k)_i^2 \leq 2\overline{\alpha}\hat{c}^\top \hat{D}^{-1}\hat{c} + 4\nu\overline{\omega}.
\end{aligned}$$

Multiplying by $4\eta^2$ yields

$$4\eta^2 r_k^\top \hat{D}^{-1} r_k \leq 8\overline{\alpha}\hat{c}^\top \hat{D}^{-1}\hat{c}\,\eta^2 + 16\nu\overline{\omega}\,\eta^2 = B_1\eta^2.$$

Combining the bounds on the linear and quadratic terms gives (46). □

Corollary A.23 (Eventual $O(\eta)$ hill tube for reduced Muon)**.** *Under the assumptions of the preceding proposition, define*

$$R_{\mathrm{Muon}} := \max\left\{\frac{B_1}{2B_0}, \sqrt{B_1}\right\}.$$

If $B_0 > 0$ and $\sqrt{H_k} \geq R_{\mathrm{Muon}}\eta$, then

$$H_{k+1} \leq H_k - B_1\eta^2.$$

Moreover, if $H_k \leq R_{\mathrm{Muon}}^2\eta^2$, then $H_{k+1} \leq R_{\mathrm{Muon}}^2\eta^2$. Consequently, the reduced Muon orbit enters and thereafter remains in the $O(\eta)$ hill tube, i.e., there exists N_{Muon} such that, for all $k \geq N_{\mathrm{Muon}}$,

$$\|\psi_k\|_2 \leq \sqrt{\nu} R_{\mathrm{Muon}}\eta.$$

Proof. From (46), we have $H_{k+1} \leq H_k - 4B_0\eta\sqrt{H_k} + B_1\eta^2$.

If $B_0 > 0$ and $\sqrt{H_k} \geq R_{\mathrm{Muon}}\eta$, then $R_{\mathrm{Muon}} \geq B_1/(2B_0)$ implies

$$4B_0\eta\sqrt{H_k} \geq 2B_1\eta^2,$$

and therefore $H_{k+1} \leq H_k - B_1\eta^2$.

To prove invariance of the tube, suppose $H_k \leq R_{\mathrm{Muon}}^2\eta^2$ and consider

$$f(y) = y - 4B_0\eta\sqrt{y} + B_1\eta^2.$$

This function is concave on $[0,\infty)$, hence it attains its maximum over $[0, R_{\mathrm{Muon}}^2\eta^2]$ at an endpoint. Using $R_{\mathrm{Muon}} \geq \sqrt{B_1}$ and $R_{\mathrm{Muon}} \geq B_1/(2B_0)$, one can check that $f(R_{\mathrm{Muon}}^2\eta^2) \leq R_{\mathrm{Muon}}^2\eta^2$ and $f(0) \leq R_{\mathrm{Muon}}^2\eta^2$, which yields $H_{k+1} \leq R_{\mathrm{Muon}}^2\eta^2$. Thus, the tube is forward invariant.

Finally, outside the tube the energy decreases by at least $B_1\eta^2$ per step, which implies a finite hitting-time bound with

$$N_{\mathrm{Muon}} = \left\lceil \frac{\max\{H_0 - R_{\mathrm{Muon}}^2\eta^2, 0\}}{B_1\eta^2} \right\rceil.$$

The proof is complete. □

Remark A.24 (A Lyapunov perspective on Simplified Muon)**.** *The conclusion of Corollary A.23 depends essentially on the sign of B_0. If $B_0 < 0$, then Proposition A.22 yields*

$$H_{k+1} \leq H_k + 4|B_0|\eta\sqrt{H_k} + B_1\eta^2,$$

which no longer implies a dissipative drift for the reduced hill energy H_k. This is consistent with the optimization intuition behind Muon: sign-based dynamics can exhibit nonsmooth switching and

transient oscillations in the river-valley coordinates, even when the full dynamics remain energy-dissipative. Writing $x = [\alpha, \beta, \gamma]^\top$, and let $\hat{k}$ denote the first column of $\hat{K}$. The Muon flow admits the Filippov signum form

$$\dot{x} \in 2\mathtt{Diag}(\sqrt{x})\,\mathtt{Sign}(\hat{k} - \hat{K}x) = -2\mathtt{Diag}(\sqrt{x})\,\mathtt{Sign}\big(\hat{K}(x - e_1)\big),$$

A natural Lyapunov candidate is the ℓ_1-potential of the residual,

$$V_{\mathrm{Muon}}(x) := \|\hat{K}(x - e_1)\|_1. \tag{47}$$

Since V_{Muon} is convex and locally Lipschitz, its Clarke generalized gradient coincides with its convex subdifferential, and

$$\partial V_{\mathrm{Muon}}(x) = \hat{K}^\top\,\mathtt{Sign}\big(\hat{K}(x - e_1)\big).$$

Hence, for almost every t, one may choose $\sigma(t) \in \mathtt{Sign}(\hat{K}(x(t) - e_1))$ such that

$$\dot{x}(t) = -2\mathtt{Diag}(\sqrt{x(t)})\,\sigma(t), \qquad g(t) := \hat{K}^\top\sigma(t) \in \partial V_{\mathrm{Muon}}(x(t)).$$

By the Clarke chain rule for locally Lipschitz Lyapunov functions along differential inclusions,

$$\dot{V}_{\mathrm{Muon}}\big(x(t); \dot{x}(t)\big) \le g(t)^\top \dot{x}(t) = -2\,\sigma(t)^\top \hat{K}\,\mathtt{Diag}(\sqrt{x(t)})\,\sigma(t) \le 0 \qquad \textit{for a.e. } t,$$

Therefore, although the reduced hill energy H_k need not be monotone when $B_0 < 0$, the full Muon dynamics still admit a natural nonincreasing energy functional. In this sense, the condition $B_0 > 0$ should be viewed as a technical assumption that excludes hill-direction excitation at the level of the reduced recursion, thereby yielding a clean one-step contraction estimate.

Proposition A.25 (Muon river progress after hill reduction)**.** *Assume there exist $N_{\mathrm{Muon}} \ge 0$ and $R_{\mathrm{Muon}} > 0$ such that $H_k \le R_{\mathrm{Muon}}^2\eta^2$ for all $k \ge N_{\mathrm{Muon}}$. In particular, the hill forcing ψ_k is $O(\eta)$ for all $k \ge N_{\mathrm{Muon}}$. Suppose further that η is sufficiently small so that, for all $k \ge N_{\mathrm{Muon}}$,*

$$|\theta_k| \ge \max\left\{ \frac{2\big(\hat{c}^\top\hat{D}^{-1}\hat{c}\big)^{1/2}R_{\mathrm{Muon}}}{\hat{a}_{\mathrm{eff}}}\eta,\; 2\sqrt{\overline{\alpha}}\,\eta,\; \sqrt{\frac{\Gamma_{\mathrm{Muon}}}{\hat{a}_{\mathrm{eff}}}}R_{\mathrm{Muon}}\eta \right\}, \qquad \Gamma_{\mathrm{Muon}} \gg 1. \tag{48}$$

Then the river variable θ_k contracts linearly in the sense that

$$|\theta_{k+1}| \le |\theta_k| - 2\sqrt{\underline{\alpha}}\,\eta. \tag{49}$$

Proof. By (36), we have $\phi_k - \hat{a}_{\mathrm{eff}}\theta_k = \hat{c}^\top\hat{D}^{-1}\psi_k$. Applying the Cauchy-Schwarz inequality in the $\hat{D}^{-1}$-inner product yields

$$|\hat{c}^\top\hat{D}^{-1}\psi_k| \le \big(\hat{c}^\top\hat{D}^{-1}\hat{c}\big)^{1/2}\big(\psi_k^\top\hat{D}^{-1}\psi_k\big)^{1/2} = \big(\hat{c}^\top\hat{D}^{-1}\hat{c}\big)^{1/2}\sqrt{H_k}.$$

For $k \ge N_{\mathrm{Muon}}$, Corollary A.23 gives $H_k \le R_{\mathrm{Muon}}^2\eta^2$, and hence

$$|\phi_k - \hat{a}_{\mathrm{eff}}\theta_k| \le \big(\hat{c}^\top\hat{D}^{-1}\hat{c}\big)^{1/2}R_{\mathrm{Muon}}\eta.$$

If (48) holds, then

$$|\phi_k - \hat{a}_{\mathrm{eff}}\theta_k| \le \frac{\hat{a}_{\mathrm{eff}}}{2}|\theta_k| \qquad \text{and} \qquad E_k = \hat{a}_{\mathrm{eff}}\theta_k^2 \ge \Gamma_{\mathrm{Muon}}R_{\mathrm{Muon}}^2\eta^2 \gg H_k.$$

Consequently, the iterate is in a river-dominant phase, ϕ_k has the same sign as θ_k, and we have $|\phi_k| \ge \frac{\hat{a}_{\mathrm{eff}}}{2}|\theta_k|$.

Moreover, the second term in (48) implies (48) gives $|\theta_k| \ge 2\sqrt{\overline{\alpha}}\,\eta$. Since $1 + \theta_k \le \overline{\alpha}$, we obtain $2\eta\sqrt{1+\theta_k} \le 2\eta\sqrt{\overline{\alpha}} \le |\theta_k|$. Thus, the river update does not cross the origin. Using the river recursion together with $\mathtt{sign}(\phi_k) = \mathtt{sign}(\theta_k)$, we obtain

$$|\theta_{k+1}| = \Big|\theta_k - 2\eta\sqrt{1+\theta_k}\,\mathtt{sign}(\phi_k)\Big| = \Big|\theta_k - 2\eta\sqrt{1+\theta_k}\,\mathtt{sign}(\theta_k)\Big| = |\theta_k| - 2\eta\sqrt{1+\theta_k}.$$

Finally, since $1 + \theta_k \ge \underline{\alpha}$, (49) follows. □

Remark A.26. *Proposition A.25 shows that once the hill forcing ψ_k is reduced to order $O(\eta)$ (e.g., via* (46)*) and river energy dominates hill energy, the Muon river variable θ_k contracts linearly according to* (49) *until it reaches an $O(\eta)$-neighborhood of the river equilibrium $\theta = 0$. Consequently, the number of Muon iterations required to reduce $|\theta_k|$ from a constant scale to an $O(\eta)$ scale is*

$$O\left(\frac{1}{\eta}\right).$$

We emphasize that this analysis under a constant step size typically yields at best an $O(\eta)$-accuracy guarantee. This behavior is consistent with the sign-based Muon update: the induced nonsmooth switching can produce persistent $O(\eta)$-scale oscillations in both the hill and river directions. Thus, one should not generally expect the discrete iterates to converge exactly to the equilibrium.

Proposition A.27 (Late-stage behavior in a relative river tube)**.** *Assume that the river curvature $\hat{a}_{\text{eff}} > 0$. Fix a constant $\bar{\rho} \in (0, 1)$. Suppose that, on the time interval under consideration, the iterates satisfy $0 < \underline{\alpha} \le 1 + \theta_k \le \overline{\alpha}$ and remain in the following relative river tube:*

$$\sqrt{\bar{q} H_k} \le \bar{\rho} \hat{a}_{\text{eff}} |\theta_k|. \tag{50}$$

Then the following statements hold.

(i) Vanilla GD naturally refines. *If the step size satisfies $0 < \eta \le \min\{\frac{1}{4\overline{\alpha}(1+\bar{\rho})\hat{a}_{\text{eff}}}, \frac{1}{4\underline{\alpha}(1-\bar{\rho})\hat{a}_{\text{eff}}}\}$, then vanilla GD can refine the river variable from $O(\eta)$ to any target accuracy $\varepsilon \ll \eta$.*

(ii) Fixed-step simplified Muon overshoots near the river equilibrium. *Under the relative tube condition, if $0 < |\theta_k| < \eta\sqrt{1 + \theta_k}$, then the river energy increases after one Muon step: $E_{k+1} > E_k$.*

Proof. Recall that (36) is $\phi_k = \hat{a}_{\text{eff}}\theta_k + \hat{c}^\top \hat{D}^{-1}\psi_k$. By Cauchy-Schwarz in the $\hat{D}^{-1}$-inner product,

$$|\hat{c}^\top \hat{D}^{-1}\psi_k|^2 \le (\hat{c}^\top \hat{D}^{-1}\hat{c})(\psi_k^\top \hat{D}^{-1}\psi_k) = \bar{q} H_k.$$

Therefore, $|\phi_k - \hat{a}_{\text{eff}}\theta_k| \le \sqrt{\bar{q} H_k}$. Using the relative tube condition (13), we obtain

$$|\phi_k - \hat{a}_{\text{eff}}\theta_k| \le \bar{\rho}\hat{a}_{\text{eff}}|\theta_k|.$$

Since $\bar{\rho} \in (0, 1)$ and $\hat{a}_{\text{eff}} > 0$, this implies that ϕ_k has the same sign as θ_k whenever $\theta_k \neq 0$. It also implies the two-sided bound

$$(1 - \bar{\rho})\hat{a}_{\text{eff}}|\theta_k| \le |\phi_k| \le (1 + \bar{\rho})\hat{a}_{\text{eff}}|\theta_k|. \tag{51}$$

We now prove the vanilla GD refinement statement. The vanilla GD river iteration is

$$\theta_{k+1} = \theta_k - 4\eta(1 + \theta_k)\phi_k.$$

Since ϕ_k and θ_k have the same sign, the update moves toward zero in the river coordinate. Using the upper bound in (51) and $1 + \theta_k \le \overline{\alpha}$,

$$4\eta(1 + \theta_k)|\phi_k| \le 4\eta\overline{\alpha}(1 + \bar{\rho})\hat{a}_{\text{eff}}|\theta_k|.$$

By the step-size condition $0 < \eta \le \frac{1}{4\overline{\alpha}(1+\bar{\rho})\hat{a}_{\text{eff}}}$, we have $4\eta(1 + \theta_k)|\phi_k| \le |\theta_k|$. Hence GD does not overshoot zero, and

$$|\theta_{k+1}| \le |\theta_k| - 4\eta\underline{\alpha}(1 - \bar{\rho})\hat{a}_{\text{eff}}|\theta_k| = [1 - 4\eta\underline{\alpha}(1 - \bar{\rho})\hat{a}_{\text{eff}}]\,|\theta_k|.$$

Iterating the contraction gives

$$|\theta_{k_0+t}| \le [1 - 4\eta\underline{\alpha}(1 - \bar{\rho})\hat{a}_{\text{eff}}]^t\,|\theta_{k_0}| \le \varepsilon.$$

Thus it is enough to take

$$t \ge \frac{1}{4\eta\underline{\alpha}(1 - \bar{\rho})\hat{a}_{\text{eff}}} \log \frac{|\theta_{k_0}|}{\varepsilon}.$$

We next prove the simplified Muon overshoot statement. Under the relative tube condition, the simplified Muon river iteration becomes

$$\theta_{k+1} = \theta_k - 2\eta\sqrt{1 + \theta_k}\,\texttt{sign}(\phi_k) \tag{52}$$

$$= \theta_k - 2\eta\sqrt{1 + \theta_k}\,\texttt{sign}(\theta_k) \tag{53}$$

$$= \texttt{sign}(\theta_k)\left(|\theta_k| - 2\eta\sqrt{1 + \theta_k}\right).$$

θ_{k+1} has the opposite sign from θ_k if $|\theta_k| < 2\eta\sqrt{1+\theta_k}$. This proves $\theta_k\theta_{k+1} < 0$.

It remains to prove the one-step river-energy increase. Since $E_k = \hat{a}_{\text{eff}}\theta_k^2$, we only need to compare $|\theta_{k+1}|$ and $|\theta_k|$. We already have

$$|\theta_{k+1}| = ||\theta_k| - 2\eta\sqrt{1+\theta_k}|.$$

If $0 < |\theta_k| < \frac{2\eta\sqrt{1+\theta_k}}{2}$, then

$$|\theta_{k+1}| = ||\theta_k| - 2\eta\sqrt{1+\theta_k}| = 2\eta\sqrt{1+\theta_k} - |\theta_k| > |\theta_k|.$$

Therefore

$$E_{k+1} = \hat{a}_{\text{eff}}\theta_{k+1}^2 > \hat{a}_{\text{eff}}\theta_k^2 = E_k.$$

Consequently, whenever a late-stage iterate enters the interval $0 < |\theta_k| < \eta\sqrt{1+\theta_k}$, the fixed-step update increases the river energy. This is the local instability mechanism near the equilibrium; exact entry into the interval depends on the preceding discrete orbit. This completes the proof. □

Remark A.28 (Why the relative river-tube condition is natural)**.** *The relative river-tube condition*

$$\sqrt{\bar{q}H_k} \le \bar{\rho}\hat{a}_{\text{eff}}|\theta_k|, \qquad 0 < \bar{\rho} < 1, \tag{54}$$

should be interpreted as the late-stage version of river dominance. Indeed, the river force admits the decomposition $\phi_k = \hat{a}_{\text{eff}}\theta_k + \hat{c}^\top\hat{D}^{-1}\psi_k$, and the perturbation caused by the residual hill force satisfies

$$\left|\hat{c}^\top\hat{D}^{-1}\psi_k\right| \le \sqrt{\bar{q}H_k}, \qquad \bar{q} := \hat{c}^\top\hat{D}^{-1}\hat{c}.$$

Therefore the relative tube condition is exactly the requirement that the hill-induced perturbation of the river force is at most a ρ-fraction of the intrinsic river force $\hat{a}_{\text{eff}}\theta_k$. Equivalently, it guarantees that ϕ_k has the same sign as θ_k, so the dynamics can still be interpreted as refinement along the river.

This condition is automatic on the exact reduced river $\psi_k = 0$, and it also holds in any sufficiently thin $O(|\theta_k|)$-neighborhood (now θ_k enters an $O(\eta)$ neighborhood of equilibrium) of that river:

$$H_k \le \frac{\bar{\rho}^2\hat{a}_{\text{eff}}^2}{\bar{q}}\,\theta_k^2 \quad \Rightarrow \quad \sqrt{\bar{q}H_k} \le \bar{\rho}\hat{a}_{\text{eff}}|\theta_k|.$$

Thus the condition is not an additional geometric object; it is a quantitative way of saying that, even in the coupled PSD case, the iterate is close enough to the tilted river for the remaining force to be river-dominated.

In the next Appendix A.3.4, we can see the normalized variants have the same continuous river–hill curves but a different clock. More precisely, under the hypotheses of the preceding theorem and the effective-step conditions, NGD inherits the GD hill reduction and has an $O(\eta)$ additive river decrement in the late tube. Thus it reaches an $O(\eta)$ neighborhood in $O(1/\eta)$ iterations, but its unit-speed update need not vanish as the residual goes to zero and fixed-step iterates may overshoot. SMuon inherits the Muon path and has an $O(\eta|\theta_k|)$ river decrement when $\chi_k \asymp |\theta_k|$; its gradient-dependent speed vanishes at the solution, so it supports GD-like refinement below the $O(\eta)$ scale. The precise recursions, hill-energy estimates, and constants are given in Proposition A.29.

A.3.4 Reduced Dynamics of Normalized GD and Scaled Muon in the PSD case

Let

$$\chi_k := 2\sqrt{u_k^2\alpha_k + v_k^2\beta_k + w_k^2\gamma_k}, \qquad m_k := \left\| \begin{bmatrix} \mathtt{sign}(u_k a_k) \\ \mathtt{sign}(v_k b_k) \\ \mathtt{sign}(w_k c_k) \end{bmatrix} \right\|_2,$$

where fixed factors depending on r are absorbed into LR. At $\chi_k = 0$ we use the zero-update convention; otherwise the normalized updates below are defined on the active sign cells. The reduced NGD and SMuon Euler iterations are the GD and Muon iterations with state-dependent effective steps

$$\eta_k^{\text{NGD}} := \frac{\eta}{\chi_k}, \qquad \eta_k^{\text{SMuon}} := \frac{\eta\chi_k}{m_k}. \tag{55}$$

Explicitly,

$$\begin{cases}\theta_{k+1} = \theta_k - 4\eta_k^{\mathrm{NGD}}(1+\theta_k)\phi_k, \\ \omega_{k+1} = \omega_k - 4\eta_k^{\mathrm{NGD}}\mathtt{Diag}(\omega_k)\psi_k,\end{cases} \tag{56}$$

and

$$\begin{cases}\theta_{k+1} = \theta_k - 2\eta_k^{\mathrm{SMuon}}\sqrt{1+\theta_k}\, s_k, \\ \omega_{k+1} = \omega_k - 2\eta_k^{\mathrm{SMuon}}\mathtt{Diag}(\omega_k)^{1/2} t_k,\end{cases} \tag{57}$$

with $s_k \in \mathtt{sign}(\phi_k)$ and $t_k \in \mathtt{sign}(\psi_k)$. Thus the GD hill-energy and river-progress estimates apply to NGD after replacing η by η_k^{NGD}, while the Muon estimates apply to SMuon after replacing η by η_k^{SMuon}. In continuous time, NGD is exactly the GD path with clock $d\tau/dt = 1/\chi$, and SMuon is exactly the Muon path with clock $d\tau/dt = \chi/m$. Here $C_{\mathrm{GD}}(h)$ denotes the GD estimate constant from Proposition A.16 with step size set to h. In particular, the inherited one-step hill estimates are

$$H_{k+1}^{\mathrm{NGD}} \le (1 - 2\underline{\omega}\mu\eta_k^{\mathrm{NGD}})H_k^{\mathrm{NGD}} + C_{\mathrm{GD}}(\eta_k^{\mathrm{NGD}})\eta_k^{\mathrm{NGD}}\phi_k^2, \tag{58}$$

$$H_{k+1}^{\mathrm{SMuon}} \le H_k^{\mathrm{SMuon}} - 4B_0\eta_k^{\mathrm{SMuon}}\sqrt{H_k^{\mathrm{SMuon}}} + B_1(\eta_k^{\mathrm{SMuon}})^2, \tag{59}$$

whenever the corresponding parent estimate and effective-step smallness conditions hold. Therefore NGD has geometric hill contraction in its gradient-time variable, while SMuon has the same square-root drift toward an $O(\eta_k^{\mathrm{SMuon}})$ hill tube as Muon.

Proposition A.29 (River-valley consequences for NGD and SMuon)**.** *Assume $\hat{a}_{\mathrm{eff}} > 0$, the iterates remain in the compact box, and the relative river-tube condition 13 holds. In continuous time, NGD has the same river/hill path as GD and SMuon has the same path as Muon. In a region where χ_k is bounded away from zero, their progress rates are the corresponding GD/Muon rates up to a constant time rescaling. In late river-dominant region, if $c_-|\theta_k| \le \chi_k \le c_+|\theta_k|$ for fixed $c_-, c_+ > 0$, then*

$$|\theta_{k+1}^{\mathrm{NGD}}| \le |\theta_k^{\mathrm{NGD}}| - c_{\mathrm{NGD}}\eta, \qquad |\theta_{k+1}^{\mathrm{SMuon}}| \le (1 - c_{\mathrm{SMuon}}\eta)|\theta_k^{\mathrm{SMuon}}|, \tag{60}$$

outside an $O(\eta)$ tube, for constants $c_{\mathrm{NGD}}, c_{\mathrm{SMuon}} > 0$. If C_{NGD} and C_{SMuon} denote the corresponding tube constants, the displayed inequalities imply the hitting-time bounds

$$T_{\mathrm{NGD}} - N_{\mathrm{NGD}} \le \left\lceil \frac{(|\theta_{N_{\mathrm{NGD}}}| - C_{\mathrm{NGD}}\eta)_+}{c_{\mathrm{NGD}}\eta} \right\rceil = O(1/\eta),$$

$$T_{\mathrm{SMuon}} - N_{\mathrm{SMuon}} \le \left\lceil \frac{\log(|\theta_{N_{\mathrm{SMuon}}}|/(C_{\mathrm{SMuon}}\eta))}{-\log(1 - c_{\mathrm{SMuon}}\eta)} \right\rceil = O\left(\frac{1}{\eta}\log\frac{1}{\eta}\right).$$

whenever the logarithm is positive and $0 < c_{\mathrm{SMuon}}\eta < 1$. Hence, NGD retains GD's geometry and $O(1/\eta)$ late-stage hitting time but can overshoot because its unit-speed update does not vanish at the solution; SMuon retains Muon's straight, uniform exploration path while its gradient-dependent speed vanishes with residual and permits asymptotic GD-like refinement below the $O(\eta)$ scale.

Proof. Equation 56 and 57 is obtained by substituting the direction–speed rows of Table 2 into the reduced GD and Muon recursions. The path identities follow from Lemma A.5. In the relative tube, ϕ_k has the sign of θ_k and is comparable to $\hat{a}_{\mathrm{eff}}\theta_k$ by (51). Since $\chi_k \asymp |\theta_k|$, division by χ_k turns the GD decrement into an $O(\eta)$ decrement (NGD), whereas multiplication by χ_k turns the Muon decrement into an $O(\eta|\theta_k|)$ decrement (SMuon), proving (60). The hill estimates inherit the parent ones whenever the effective-step smallness conditions hold. □

B Additional Details for Section 5

B.1 Spectral-River Geometry

In this section, we study the optimization of a general objective function $\mathcal{F}(X) : \mathbb{R}^{n\times n} \to \mathbb{R}$ with the goal of finding a critical point $X^\star \in \mathbb{R}^{n\times n}$, i.e., $\nabla\mathcal{F}(X^\star) = 0$. We assume that the target matrix $X^\star$ (also called the ground truth) admits the singular value decomposition (SVD)

$$X^\star = U_\star\Sigma_\star V_\star^\top, \qquad U_\star^\top U_\star = I_n, \qquad V_\star^\top V_\star = I_n, \qquad \Sigma_\star = \mathtt{Diag}(\sigma_1, \dots, \sigma_n),$$

with $\sigma_i > 0$. The rank-deficient case follows by replacing n with the signal rank r and restricting attention to the corresponding signal block. Therefore, we focus on the square case for simplicity, though the optimization variable $X \in \mathbb{R}^{n\times n}$ can be generalized to a rectangular matrix.

Definition B.1 (Spectral embedding and spectral river)**.** *We define the spectral embedding $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ by $\mathcal{B}(x) := U_\star \mathtt{Diag}(x) V_\star^\top$ for any $x \in \mathbb{R}^n$. The spectral river through $X^\star$ is the affine set*

$$\mathcal{R}^\star := \{X^\star + \mathcal{B}(x) \mid x \in \mathbb{R}^n\} \subset \mathbb{R}^{n\times n}. \tag{61}$$

Since $X^\star = \mathcal{B}(\sigma)$ with $\sigma := [\sigma_1, \ldots, \sigma_n]^\top$, any $Y \in \mathcal{R}^\star$ can be written as $Y = \mathcal{B}(y)$ for some $y \in \mathbb{R}^n$, equivalently $Y = U_\star \mathtt{Diag}(y) V_\star^\top$. Therefore, any point Y on the river shares the left/right singular vector subspaces $(U_\star, V_\star)$ with $X^\star$; only the singular-value magnitudes vary along the river. Moreover, $\mathcal{B}(\mathbf{1}) = U_\star V_\star^\top$ and $\|\mathcal{B}(x)\|_F = \|x\|_2$.

Definition B.2 (River-restricted objective)**.** *We define the objective restricted to the spectral river (i.e., the pullback of $\mathcal{F}$ under the map $x \mapsto X^\star + \mathcal{B}(x)$) as*

$$f(x) := \mathcal{F}\big(X^\star + \mathcal{B}(x)\big) - \mathcal{F}(X^\star), \qquad x \in \mathbb{R}^n. \tag{62}$$

Definition B.3 (Positive chamber)**.** *We define the positive chamber of the spectral river as*

$$\mathcal{R}^\star_+ := \{X^\star + \mathcal{B}(x) \mid \sigma + x \succ \mathbf{0}\} = \{U_\star \mathtt{Diag}(y) V_\star^\top \mid y \succ \mathbf{0}\}. \tag{63}$$

Within this chamber, the polar factor is constant and equals $\mathtt{msign}(Y) = \mathcal{B}(\mathbf{1})$ for all $Y \in \mathcal{R}^\star_+$. Moreover, if some entries of y are negative (so that $Y = U_\star \mathtt{Diag}(y) V_\star^\top$ has mixed signs), then the polar factor becomes $U_\star \mathtt{Diag}(\mathtt{sign}(y)) V_\star^\top$.

Definition B.4 (Orthogonal basis and projection operator)**.** *For each $i = 1, \cdots, n$, define the basis matrices for the river direction space $\mathtt{span}\{B_1, \ldots, B_n\}$ using the i-th standard basis vector $e_i \in \mathbb{R}^n$:*

$$B_i := U_\star e_i e_i^\top V_\star^\top. \tag{64}$$

These matrices form an orthonormal set under the Frobenius inner product, i.e., $\langle B_i, B_j \rangle = \delta_{ij}$. Consequently, $\mathcal{B}(x) = \sum_{i=1}^n x_i B_i$ and $\langle \mathcal{B}(x), \mathcal{B}(y) \rangle = x^\top y$. Moreover, for any $Y \in \mathbb{R}^{n\times n}$, the (Frobenius) orthogonal projection onto $\mathtt{span}\{B_1, \cdots, B_n\}$ is

$$\mathtt{Proj}_{\mathcal{R}}(Y) = \sum\nolimits_{i=1}^n \langle B_i, Y \rangle B_i = U_\star \mathtt{Diag}\big(\mathtt{diag}(U_\star^\top Y V_\star)\big) V_\star^\top. \tag{65}$$

Definition B.5 (Spectral-river tube)**.** *Fix a scale $\delta > 0$. Define the δ-tube around the spectral river $\mathcal{R}^\star := X^\star + \mathtt{span}\{B_1, \cdots, B_n\}$ by*

$$\mathcal{T}_\delta(\mathcal{R}^\star) := \left\{ X \in \mathbb{R}^{n\times n} \,\middle|\, \mathtt{dist}_F(X, \mathcal{R}^\star) \le \delta \right\}, \tag{66}$$

where $\mathtt{dist}_F(X, \mathcal{R}^\star) := \inf_{Y \in \mathcal{R}^\star} \|X - Y\|_F$. Equivalently, each X admits a unique orthogonal decomposition

$$X = X^\star + \mathcal{B}(x) + E^\perp, \qquad E^\perp \in \mathtt{span}\{B_1, \cdots, B_n\}^\perp,$$

and $X \in \mathcal{T}_\delta(\mathcal{R}^\star)$ if and only if $\|E^\perp\|_F \le \delta$.

Definition B.6 (Hill component and river-dominance)**.** *For any $Y \in \mathbb{R}^{n\times n}$, we define its hill component as $\mathtt{Proj}_{\mathcal{R}^\perp}(Y)$, where $\mathtt{Proj}_{\mathcal{R}^\perp}(\cdot)$ denotes the Frobenius-orthogonal projections onto $\mathtt{span}\{B_1, \cdots, B_n\}$'s orthogonal complement, i.e., $\mathtt{span}\{B_1, \cdots, B_n\}^\perp$. We say that Y is river-dominated if its hill component is small relative to its river component.*

Lemma B.7 (Hessian projection)**.** *Let $\mathcal{H}_\star := \nabla^2 \mathcal{F}(X^\star)$, viewed as a linear operator acting on matrices. If the river residual is $\mathcal{B}(x)$, then its Hessian image projected back onto the river direction space satisfies*

$$\begin{aligned} \mathtt{Proj}_{\mathcal{R}}\big(\mathcal{H}_\star[\mathcal{B}(x)]\big) &= \sum\nolimits_{i=1}^n \langle B_i, \mathcal{H}_\star[\mathcal{B}(x)] \rangle B_i = \sum\nolimits_{i=1}^n \Big\langle B_i, \sum\nolimits_{j=1}^n x_j \mathcal{H}_\star[B_j] \Big\rangle B_i \\ &= \sum\nolimits_{i=1}^n \Big(\sum\nolimits_{j=1}^n x_j \langle B_i, \mathcal{H}_\star[B_j] \rangle \Big) B_i = \sum\nolimits_{i=1}^n \Big(\sum\nolimits_{j=1}^n K_{ij} x_j \Big) B_i = \mathcal{R}(Kx), \end{aligned} \tag{67}$$

where $K \in \mathbb{R}^{n\times n}$ is the coordinate matrix of the projected Hessian action in the basis $\{B_i\}_{i=1}^n$, defined by

$$K_{ij} := \langle B_i, \mathcal{H}_\star[B_j] \rangle. \tag{68}$$

If $\mathcal{F}$ is twice continuously differentiable, then $\mathcal{H}_\star$ is self-adjoint w.r.t. the Frobenius inner product, hence K is symmetric. Moreover, near the critical point $X^\star$ (as $a \to 0$), a second-order expansion yields

$$\nabla \mathcal{F}(X^\star + a\mathcal{B}(x)) = a\, \mathcal{H}_\star[\mathcal{B}(x)] + O(a^2).$$

If the off-river component of $\mathcal{H}_\star[\mathcal{B}(x)]$ is negligible so that $\mathcal{H}_\star[\mathcal{B}(x)] \approx \mathtt{Proj}_{\mathcal{R}}\big(\mathcal{H}_\star[\mathcal{B}(x)]\big) = \mathcal{B}(Kx)$ and $Kx \succ \mathbf{0}$, then the resulting gradient direction lies in the positive chamber, and hence $\mathtt{msign}(\mathcal{B}(Kx)) = \mathcal{B}(\mathbf{1})$. Note that Muon observes $\mathcal{H}_\star[\mathcal{B}(x)]$ rather than $\mathcal{B}(x)$ itself; however, in later analysis we will assume that the river component dominates.

B.2 Early-Stage: When Muon Accelerates River Progress

This subsection establishes an early-stage regime in which Muon accelerates progress along the spectral river. Our goal is to isolate conditions under which the effective curvature restricted to the river is substantially smaller than the ambient curvature in the full matrix space. In such a regime, a Muon step that is aligned with the river geometry can admit a provably larger decrease than a conservative gradient step. We adopt the following local smoothness assumption.

Assumption B.8 (Local smoothness separation for $\mathcal{F}$ and f)**.** *Fix a neighborhood of interest around the current iterate(s). There exist constants $L_{\mathrm{full}}, L_{\mathrm{riv}} > 0$ with $L_{\mathrm{riv}} \ll L_{\mathrm{full}}$ such that, for all X in this neighborhood and all sufficiently small perturbations $\Delta X \in \mathbb{R}^{n\times n}$,*

$$\|\nabla\mathcal{F}(X+\Delta X) - \nabla\mathcal{F}(X)\|_F \le L_{\mathrm{full}}\,\|\Delta X\|_F. \tag{69}$$

Moreover, for all x in the corresponding river neighborhood and all sufficiently small $\Delta x \in \mathbb{R}^n$, the river-restricted objective f satisfies

$$f(x+\Delta x) \le f(x) + \langle \nabla f(x), \Delta x\rangle + \frac{L_{\mathrm{riv}}}{2}\|\Delta x\|_2^2. \tag{70}$$

Here, (69) *is the standard Lipschitz continuity of $\nabla\mathcal{F}$ (in Frobenius norm), while* (70) *is the usual quadratic upper bound (descent lemma Nesterov et al. [2018]) for f, which is implied by ∇f being L_{riv}-Lipschitz on the neighborhood.*

Because L_{full} controls smoothness in all directions, it can be dominated by high-curvature components (e.g., "spike" directions). In contrast, L_{riv} quantifies curvature only along the signal-aligned, singular-value directions parametrized by the river. This separation generalizes the mixed-spiked MS dynamics developed in this paper: river coefficients may remain well-conditioned even when transverse "hill" directions are stiff and therefore force conservative gradient steps. Consequently, once the iterates lie on or near the river, the effective constant L_{riv} governing Muon's motion can be substantially smaller than the global constant L_{full} that limits vanilla gradient methods.

Early-stage: a one-step comparison. We formalize a simple one-step manifestation of the early-stage speedup. Fix a point on the spectral river

$$X_0 \;=\; X^\star + \mathcal{B}(x_0).$$

(Here "0" indicates the point at which we start the analysis, not necessarily the initialization of the algorithm.) We compare the certified one-step decrease obtained by a momentum-Muon step aligned with the river geometry versus a conservative gradient step whose step size is limited by the ambient smoothness.

- **One-step momentum-Muon (ambient form).** Given a stored momentum matrix $M_0^{\mathrm{Muon}} \in \mathbb{R}^{n\times n}$ and momentum parameter $\xi \ge 0$, define

$$M_1^{\mathrm{Muon}} = \xi M_0^{\mathrm{Muon}} + \nabla\mathcal{F}(X_0), \qquad X_1^{\mathrm{Muon}} = X_0 - \eta^{\mathrm{Muon}}\,\mathtt{msign}(M_1^{\mathrm{Muon}}).$$

- **One-step vanilla gradient descent (ambient form).** With step size $\eta^{\mathrm{GD}} > 0$,

$$X_1^{\mathrm{GD}} \;=\; X_0 - \eta^{\mathrm{GD}}\nabla\mathcal{F}(X_0).$$

The key technical simplification is that, under a (local) *gradient-invariance* condition, the ambient gradients and momenta can be represented in the river coordinates, allowing us to reduce the comparison from $\mathcal{F}$ on matrices to f on $\mathbb{R}^n$.

Lemma B.9 (River reduction under gradient invariance)**.** *Suppose the Muon momentum is river-aligned in the sense that*

$$M_0^{\mathrm{Muon}} = \mathcal{B}(m_0) \quad \textit{and} \quad M_1^{\mathrm{Muon}} = \mathcal{B}(m_1) \qquad \textit{for some } m_0, m_1 \in \mathbb{R}^n.$$

Assume moreover that the river is gradient-invariant at $X_0 = X^\star + \mathcal{B}(x_0)$, namely

$$\nabla\mathcal{F}(X_0) = \mathcal{B}\big(\nabla f(x_0)\big). \tag{71}$$

Then the momentum update in the ambient space is equivalent to the river-coordinate update

$$M_1^{\mathrm{Muon}} = \xi M_0^{\mathrm{Muon}} + \nabla\mathcal{F}(X_0) \quad\Longleftrightarrow\quad m_1 = \xi m_0 + \nabla f(x_0).$$

In particular, the corresponding sign directions satisfy $\texttt{msign}(M_1^{\mathrm{Muon}}) = \mathcal{B}(\texttt{sign}(m_1))$.

Furthermore, if m_0 *is coordinate-wise sign-aligned with* $\nabla f(x_0)$*, then*

$$\langle \nabla f(x_0), \texttt{sign}(m_1) \rangle = \|\nabla f(x_0)\|_1.$$

More generally, without requiring sign alignment, define

$$\rho \;:=\; \frac{\langle \nabla f(x_0), \texttt{sign}(m_1) \rangle}{\|\nabla f(x_0)\|_1},$$

whenever $\nabla f(x_0) \neq 0$*. Then* $\rho \in (0, 1]$ *precisely captures the degree of sign agreement and yields*

$$\langle \nabla f(x_0), \texttt{sign}(m_1) \rangle \;\geq\; \rho \, \|\nabla f(x_0)\|_1. \tag{72}$$

Proof. The equivalence follows by substituting $M_0^{\mathrm{Muon}} = \mathcal{B}(m_0)$ and (71) into Muon's momentum update and using linearity of $\mathcal{B}$. The statement about sign directions follows from the definition of $\texttt{msign}(\cdot)$ and the coordinate-wise nature of $\texttt{sign}(\cdot)$. If m_0 and $\nabla f(x_0)$ are coordinate-wise sign-aligned, then $m_1 = \xi m_0 + \nabla f(x_0)$ has the same sign pattern as $\nabla f(x_0)$, hence $\langle \nabla f(x_0), \texttt{sign}(m_1) \rangle = \|\nabla f(x_0)\|_1$. Equation 72 is immediate from the definition of ρ. □

By Lemma B.9, it suffices to compare the following *river-coordinate* one-step updates:

- **One-step momentum-Muon (river form).**

$$m_1 = \xi m_0 + \nabla f(x_0), \qquad x_1^{\mathrm{Muon}} = x_0 - \eta^{\mathrm{Muon}} \, \texttt{sign}(m_1).$$

- **One-step vanilla gradient descent (river form).**

$$x_1^{\mathrm{GD}} = x_0 - \eta^{\mathrm{GD}} \nabla f(x_0).$$

Theorem B.10 (Early-stage Muon acceleration along the spectral river (one-step analysis))**.** *Assume the local smoothness conditions in Assumption B.8 hold and that the gradient-invariance condition 71 holds at* x_0*. Fix* $\rho \in (0, 1]$ *as in* (72)*, and assume* $\nabla f(x_0) \neq 0$*.*

- ***(Muon decrease along the sign direction.)*** *Let* $\eta^{\mathrm{Muon}} = \langle \nabla f(x_0), \texttt{sign}(m_1) \rangle / (L_{\mathrm{riv}} n)$*, then the one-step update* $x_1^{\mathrm{Muon}} = x_0 - \eta^{\mathrm{Muon}} \texttt{sign}(m_1)$ *satisfies*

$$f(x_1^{\mathrm{Muon}}) \leq f(x_0) - \frac{\langle \nabla f(x_0), \texttt{sign}(m_1) \rangle^2}{2L_{\mathrm{riv}} \, \|\texttt{sign}(m_1)\|_2^2} \leq f(x_0) - \frac{\rho^2 \|\nabla f(x_0)\|_1^2}{2L_{\mathrm{riv}} \, n},$$

  where we used $\|\texttt{sign}(m_1)\|_2^2 = n$ *(assume no ties so that* $\texttt{sign}(m_1) \in \{\pm 1\}^n$*) and* (72)*.*

- ***(Conservative GD decrease under ambient step size.)*** *With the conservative choice* $\eta^{\mathrm{GD}} = 1/L_{\mathrm{full}}$*, the GD update* $x_1^{\mathrm{GD}} = x_0 - \eta^{\mathrm{GD}} \nabla f(x_0)$ *satisfies*

$$f(x_1^{\mathrm{GD}}) \leq f(x_0) - \frac{\|\nabla f(x_0)\|_2^2}{2L_{\mathrm{full}}},$$

  whenever $L_{\mathrm{full}} \geq L_{\mathrm{riv}}$*.*

Consequently, the certified Muon decrease is larger than the certified conservative GD decrease whenever

$$\rho^2 \frac{L_{\mathrm{full}}}{L_{\mathrm{riv}}} \cdot \frac{\|\nabla f(x_0)\|_1^2}{n \|\nabla f(x_0)\|_2^2} \;>\; 1.$$

Equivalently, the certified speedup factor satisfies

$$\frac{\text{Muon decrease}}{\text{GD decrease}} \geq \rho^2 \frac{L_{\mathrm{full}}}{L_{\mathrm{riv}}} \cdot \frac{\|\nabla f(x_0)\|_1^2}{n \|\nabla f(x_0)\|_2^2}.$$

Proof. For the momentum-Muon step, apply the descent lemma Nesterov et al. [2018] for the L_{riv}-smooth function f with direction $\texttt{sign}(m_1)$:

$$f(x_0 - \eta \texttt{sign}(m_1)) \leq f(x_0) - \eta \langle \nabla f(x_0), \texttt{sign}(m_1) \rangle + \frac{L_{\mathrm{riv}}}{2} \eta^2 \|\texttt{sign}(m_1)\|_2^2.$$

The right-hand side is minimized at $\eta = \langle \nabla f(x_0), \mathtt{sign}(m_1)\rangle / (L_{\mathrm{riv}} \|\mathtt{sign}(m_1)\|_2^2)$, which yields

$$f(x_0 - \eta^{\mathrm{Muon}} \mathtt{sign}(m_1)) \le f(x_0) - \frac{\langle \nabla f(x_0), \mathtt{sign}(m_1)\rangle^2}{2L_{\mathrm{riv}} \|\mathtt{sign}(m_1)\|_2^2}.$$

Using $\|\mathtt{sign}(m_1)\|_2^2 = n$ together with (72) gives the claimed bound.

For gradient descent, apply the same descent lemma with direction $\nabla f(x_0)$ and step size $\eta^{\mathrm{GD}} = 1/L_{\mathrm{full}}$:

$$f\big(x_0 - \eta^{\mathrm{GD}} \nabla f(x_0)\big) \le f(x_0) - \frac{1}{L_{\mathrm{full}}} \|\nabla f(x_0)\|_2^2 + \frac{L_{\mathrm{riv}}}{2L_{\mathrm{full}}^2} \|\nabla f(x_0)\|_2^2.$$

If $L_{\mathrm{full}} \ge L_{\mathrm{riv}}$, then the last two terms combine to $-\|\nabla f(x_0)\|_2^2/(2L_{\mathrm{full}})$, which completes the proof. □

Remark B.11 (When can $L_{\mathrm{riv}} \ll L_{\mathrm{full}}$ occur?)**.** *The inequality $L_{\mathrm{riv}} \le L_{\mathrm{full}}$ is natural: L_{riv} quantifies smoothness only along the river directions, while L_{full} controls smoothness over all ambient directions. A pronounced gap arises when the local Hessian is much flatter on the river than on its transverse complement. Concretely, with respect to the orthogonal decomposition $\mathcal{V} = \mathcal{R}_\star \oplus \mathcal{R}_\star^\perp$, if the Hessian admits an (approximately) block-structured form*

$$\nabla^2 \mathcal{F}(X) \approx \begin{pmatrix} K_\star & K_\circ^\top \\ K_\circ & K_\diamond \end{pmatrix},$$

where the cross-coupling is small (e.g., $\|K_\circ\|_{\mathrm{op}}$ is small, corresponds to "bulk") and the transverse hill block is much stiffer than the river block (i.e., $\|K_\diamond\|_{\mathrm{op}} \gg \|K_\star\|_{\mathrm{op}}$, "spike" $\gg$ "signal"), then one expects

$$L_{\mathrm{riv}} \approx \|K_\star\|_{\mathrm{op}} \qquad \textit{and} \qquad L_{\mathrm{full}} \approx \|\nabla^2 \mathcal{F}(X)\|_{\mathrm{op}} \approx \|K_\diamond\|_{\mathrm{op}}.$$

In this regime, conservative gradient steps are limited by high-curvature off-river ("hill") directions, whereas Muon can exploit the milder geometry along the information-bearing spectral river.

B.3 Late-Stage: When Muon Fails in River Refinement

We now turn to the late-stage regime, in which the iterate is already close to the optimum and the residual magnitude is small. In this regime, the spectral polar operator $\mathtt{msign}(\cdot)$—which discards singular-value magnitudes by keeping only their signs—can become harmful. Intuitively, the same scale-invariance that helps Muon rapidly align with the river direction early on may prevent the algorithm from making progressively smaller, well-calibrated corrections near the solution. As a result, Muon can exhibit a coarse "quantized" motion that overshoots the minimizer, whereas a suitably small gradient(-momentum) step continues to refine.

The next subsection isolates this phenomenon in an exactly solvable quadratic model, where all statements can be proved without approximation.

B.3.1 A toy quadratic model: explicit overshoot under `msign`

We strip away nonessential nonlinear effects and retain only a strictly quadratic geometry:

$$\mathcal{F}_{\mathrm{quad}}(X) := \frac{1}{2} \|X - X^\star\|_F^2, \qquad X \in \mathbb{R}^{n \times n}. \tag{73}$$

For (73), one has

$$\nabla \mathcal{F}_{\mathrm{quad}}(X) = X - X^\star, \qquad \nabla^2 \mathcal{F}_{\mathrm{quad}}(X) = I,$$

so the gradient is exactly the residual. In particular, at any river point $X = X^\star + \mathcal{B}(x)$,

$$\nabla \mathcal{F}_{\mathrm{quad}}(X) = \mathcal{B}(x),$$

hence the gradient-invariance property holds automatically (no additional assumption is needed).

The river-restricted objective induced by (73) is

$$f_{\mathrm{quad}}(x) := \mathcal{F}_{\mathrm{quad}}(X^\star + \mathcal{B}(x)) - \mathcal{F}_{\mathrm{quad}}(X^\star) = \frac{1}{2} \|\mathcal{B}(x)\|_F^2 = \frac{1}{2} \|x\|_2^2, \tag{74}$$

and therefore $\nabla f_{\mathrm{quad}}(x) = x$ and $\nabla^2 f_{\mathrm{quad}}(x) = I_n$.

Update rules (river form). In this subsection we work directly in river coordinates, using (74). We consider a momentum-Muon update that moves along the sign direction (`msign` becomes `sign`):

$$m_{t+1}^{\mathrm{Muon}} = \xi m_t^{\mathrm{Muon}} + x_t^{\mathrm{Muon}}, \qquad x_{t+1}^{\mathrm{Muon}} = x_t^{\mathrm{Muon}} - \eta^{\mathrm{Muon}}\, \mathtt{sign}(m_{t+1}^{\mathrm{Muon}}),$$

and we compare it against a (heavy-ball style) momentum gradient method that uses the full momentum magnitude (i.e., a kind of momentum-GD):

$$m_{t+1}^{\mathrm{GD}} = \xi m_t^{\mathrm{GD}} + x_t^{\mathrm{GD}}, \qquad x_{t+1}^{\mathrm{GD}} = x_t^{\mathrm{GD}} - \eta^{\mathrm{GD}} m_{t+1}^{\mathrm{GD}}.$$

Both are initialized from the same (x_0, m_0).

Proposition B.12 (Exact quadratic separation: Muon overshoots while momentum GD refines)**.** *Fix $\xi \geq 0$ and consider* (73)*. Let $X_0 = X^\star + a\,\mathcal{B}(x_0)$ with $a \in (0,1)$ and $x_0 \succ \mathbf{0}$. Assume the stored momentum is river-aligned and small:*

$$M_0 = \mathcal{B}(m_0), \qquad m_0 \succ \mathbf{0}, \qquad \|m_0\|_2 \leq C_m\, a,$$

for some constant C_m independent of a. Define the coordinate average $\bar{x}_0 := \frac{1}{n}\mathbf{1}^\top x_0$.

Choose a Muon step size of the form

$$\eta^{\mathrm{Muon}} = (1+\varepsilon)\, a\, \bar{x}_0, \qquad \textit{for some } \varepsilon \in (0,1).$$

Assume moreover that the second Muon momentum remains in the positive chamber:

$$\xi^2 m_0 + a\,[(1+\xi)x_0 - (1+\varepsilon)\bar{x}_0\,\mathbf{1}] \succ \mathbf{0}. \tag{75}$$

Then the Muon iterates satisfy a strict two-step overshoot:

$$\mathcal{F}_{\mathrm{quad}}\big(X_1^{\mathrm{Muon}}\big) < \mathcal{F}_{\mathrm{quad}}(X_0) < \mathcal{F}_{\mathrm{quad}}\big(X_2^{\mathrm{Muon}}\big)\,. \tag{76}$$

Moreover, there exists a constant $\eta_0 > 0$, depending only on (ξ, x_0, C_m) but not on a, such that the momentum-GD iterates initialized from the same (X_0, M_0) satisfy

$$\mathcal{F}_{\mathrm{quad}}\big(X_2^{\mathrm{GD}}\big) < \mathcal{F}_{\mathrm{quad}}(X_0) \tag{77}$$

for all $0 < \eta^{\mathrm{GD}} \leq \eta_0$. In particular, since $\eta^{\mathrm{Muon}} = O(a)$ while $\eta_0 = O(1)$, we have $\eta^{\mathrm{Muon}} < \eta_0$ for all sufficiently small a.

Proof. Throughout the proof, we use the river objective 74, i.e., $f_{\mathrm{quad}}(x) = \frac{1}{2}\|x\|_2^2$, and the identity $\mathcal{F}_{\mathrm{quad}}(X^\star + \mathcal{B}(x)) - \mathcal{F}_{\mathrm{quad}}(X^\star) = f_{\mathrm{quad}}(x)$. Thus it suffices to compare the squared norms of the corresponding river coordinates.

Step 1: Muon decreases at the first step. At $X_0 = X^\star + a\mathcal{B}(x_0)$ we have $\nabla\mathcal{F}_{\mathrm{quad}}(X_0) = a\mathcal{B}(x_0)$, hence the first Muon momentum is

$$M_1^{\mathrm{Muon}} = \xi M_0 + \nabla\mathcal{F}_{\mathrm{quad}}(X_0) = \mathcal{B}(\xi m_0 + a x_0).$$

Since $m_0 \succ \mathbf{0}$ and $x_0 \succ \mathbf{0}$, we obtain $\xi m_0 + a x_0 \succ \mathbf{0}$, and therefore $\mathtt{sign}(\xi m_0 + a x_0) = \mathbf{1}$. Consequently, the first Muon iterate in river coordinates is

$$x_1^{\mathrm{Muon}} = a x_0 - \eta^{\mathrm{Muon}}\mathbf{1}.$$

A direct expansion yields

$$\begin{aligned}
&f_{\mathrm{quad}}(x_1^{\mathrm{Muon}}) - f_{\mathrm{quad}}(a x_0) \\
=&\frac{1}{2}\|a x_0 - \eta^{\mathrm{Muon}}\mathbf{1}\|_2^2 - \frac{1}{2}\|a x_0\|_2^2 = -\eta^{\mathrm{Muon}}\, a\, \mathbf{1}^\top x_0 + \frac{(\eta^{\mathrm{Muon}})^2}{2}\,\|\mathbf{1}\|_2^2 \\
=&-\eta^{\mathrm{Muon}}\, a\, n\bar{x}_0 + \frac{(\eta^{\mathrm{Muon}})^2}{2}\, n = \frac{n}{2}\,\eta^{\mathrm{Muon}}\left(\eta^{\mathrm{Muon}} - 2a\bar{x}_0\right).
\end{aligned}$$

With $\eta^{\mathrm{Muon}} = (1+\varepsilon)a\bar{x}_0$ and $\varepsilon \in (0,1)$, we have $\eta^{\mathrm{Muon}} - 2a\bar{x}_0 = (\varepsilon - 1)a\bar{x}_0 < 0$, hence $f_{\mathrm{quad}}(x_1^{\mathrm{Muon}}) < f_{\mathrm{quad}}(a x_0)$, which implies $\mathcal{F}_{\mathrm{quad}}(X_1^{\mathrm{Muon}}) < \mathcal{F}_{\mathrm{quad}}(X_0)$.

Step 2: Muon overshoots after two steps. The second Muon momentum in river coordinates equals

$$\begin{aligned} m_2^{\mathrm{Muon}} &= \xi(\xi m_0 + a x_0) + x_1^{\mathrm{Muon}} = \xi^2 m_0 + a(1+\xi)x_0 - \eta^{\mathrm{Muon}}\mathbf{1} \\ &= \xi^2 m_0 + a\left[(1+\xi)x_0 - (1+\varepsilon)\bar{x}_0\,\mathbf{1}\right]. \end{aligned}$$

By assumption 75, $m_2^{\mathrm{Muon}} \succ \mathbf{0}$, hence again $\mathtt{sign}(m_2^{\mathrm{Muon}}) = \mathbf{1}$ and

$$x_2^{\mathrm{Muon}} = x_1^{\mathrm{Muon}} - \eta^{\mathrm{Muon}}\mathbf{1} = a x_0 - 2\eta^{\mathrm{Muon}}\mathbf{1}.$$

Expanding as before,

$$\begin{aligned} & f_{\mathrm{quad}}(x_2^{\mathrm{Muon}}) - f_{\mathrm{quad}}(a x_0) \\ =& \frac{1}{2}\|a x_0 - 2\eta^{\mathrm{Muon}}\mathbf{1}\|_2^2 - \frac{1}{2}\|a x_0\|_2^2 = -2\eta^{\mathrm{Muon}}\, a\, \mathbf{1}^\top x_0 + \frac{(2\eta^{\mathrm{Muon}})^2}{2}\,\|\mathbf{1}\|_2^2 \\ =& -2\eta^{\mathrm{Muon}}\, a\, n\bar{x}_0 + 2(\eta^{\mathrm{Muon}})^2 n = 2n\eta^{\mathrm{Muon}} a\bar{x}_0 \left(\frac{\eta^{\mathrm{Muon}}}{a\bar{x}_0} - 1\right). \end{aligned}$$

Because $\eta^{\mathrm{Muon}} = (1+\varepsilon)a\bar{x}_0 > a\bar{x}_0$, the last factor is positive; thus $f_{\mathrm{quad}}(x_2^{\mathrm{Muon}}) > f_{\mathrm{quad}}(a x_0)$, which yields the overshoot claim 76.

Step 3: two-step decrease for momentum GD with a small constant step size. Now consider the momentum-GD recursion

$$\begin{aligned} m_1^{\mathrm{GD}} &= \xi m_0 + a x_0, & x_1^{\mathrm{GD}} &= a x_0 - \eta^{\mathrm{GD}} m_1^{\mathrm{GD}}, \\ m_2^{\mathrm{GD}} &= \xi m_1^{\mathrm{GD}} + x_1^{\mathrm{GD}}, & x_2^{\mathrm{GD}} &= x_1^{\mathrm{GD}} - \eta^{\mathrm{GD}} m_2^{\mathrm{GD}}. \end{aligned}$$

A straightforward elimination gives

$$x_2^{\mathrm{GD}} = a x_0 - \eta^{\mathrm{GD}}\Big[\xi(1+\xi)m_0 + (2+\xi)a x_0\Big] + (\eta^{\mathrm{GD}})^2(\xi m_0 + a x_0). \tag{78}$$

Define the a-independent vectors $u := x_0$ and $v := m_0/a$, so that $\|v\|_2 \le C_m$ and (78) can be rewritten as

$$x_2^{\mathrm{GD}} = a\left[u - \eta^{\mathrm{GD}}\big(\xi(1+\xi)v + (2+\xi)u\big) + (\eta^{\mathrm{GD}})^2(\xi v + u)\right].$$

Let $q(\eta) := u - \eta\big(\xi(1+\xi)v + (2+\xi)u\big) + \eta^2(\xi v + u)$, so that $x_2^{\mathrm{GD}} = a\, q(\eta^{\mathrm{GD}})$ and

$$f_{\mathrm{quad}}(x_2^{\mathrm{GD}}) - f_{\mathrm{quad}}(a x_0) = \frac{a^2}{2}\Big(\|q(\eta^{\mathrm{GD}})\|_2^2 - \|u\|_2^2\Big).$$

Expanding $\|q(\eta)\|_2^2$ around $\eta = 0$ gives

$$\|q(\eta)\|_2^2 = \|u\|_2^2 - 2\eta\,\langle u, \xi(1+\xi)v + (2+\xi)u\rangle + \eta^2\, R(\eta),$$

where $R(\eta)$ is a polynomial in η whose coefficients are bounded in terms of $(\xi, \|u\|_2, \|v\|_2)$. Using $u = x_0 \succ \mathbf{0}$ and $v = m_0/a \succ \mathbf{0}$, we have

$$\langle u, \xi(1+\xi)v + (2+\xi)u\rangle \ge (2+\xi)\|u\|_2^2.$$

Moreover, there exists a constant $\bar{C}_m = \bar{C}_m(\xi, \|u\|_2, C_m)$ such that $|R(\eta)| \le \bar{C}_m$ for all $\eta \in [0,1]$ (e.g., by bounding $\|q(\eta)\|_2$ and its coefficients via triangle inequality). Therefore, for every $\eta \in (0,1]$,

$$\|q(\eta)\|_2^2 - \|u\|_2^2 \le -2(2+\xi)\eta\|u\|_2^2 + \bar{C}_m\eta^2.$$

Choose

$$\eta_0 := \min\left\{1,\ \frac{(2+\xi)\|u\|_2^2}{\bar{C}_m}\right\} = \min\left\{1,\ \frac{(2+\xi)\|x_0\|_2^2}{\bar{C}_m}\right\},$$

then for any $0 < \eta^{\mathrm{GD}} \le \eta_0$ we have $\|q(\eta^{\mathrm{GD}})\|_2^2 - \|u\|_2^2 < 0$, which implies $f_{\mathrm{quad}}(x_2^{\mathrm{GD}}) < f_{\mathrm{quad}}(a x_0)$ and hence (77). □

B.3.2 Late-stage overshoot for general smooth objectives

We now move from the exact quadratic toy model to a local theorem for general smooth objectives. The proof is carried out inside a late-stage spectral-river tube, where the iterate, gradient, and momentum admit a river-hill decomposition and the hill component remains small. In this regime, Muon's `msign` update behaves as an almost-constant river sign step, so the two-step overshoot mechanism from the quadratic model persists up to perturbative errors.

River-dominated momentum. We first explain why a river-dominated momentum state is natural in the late stage. Intuitively, once the iterates enter a spectral-river tube, the gradients become predominantly supported on the river subspace, with only small hill components. Since the stored momentum is formed as an exponential moving average (EMA) of recent gradients, it should therefore inherit the same river–hill structure. The following lemma makes this inheritance precise.

Lemma B.13 (EMA momentum remains river-dominated in a tube)**.** *Suppose that at a reference time* 0 *the stored momentum is an exponential moving average (EMA) over a window of length* L*:*

$$M_0 = \sum_{j=1}^{L} \xi^{j-1} \nabla \mathcal{F}(X_{-j}) + \xi^L M_{-L}, \qquad 0 \le \xi < 1.$$

Assume that along this window the gradients admit decompositions

$$\nabla \mathcal{F}(X_{-j}) = \mathcal{B}(g_j) + E_j^g, \quad g_j \succ \mathbf{0}, \quad \|g_j\|_2 \le \overline{C}_g a, \quad \min_i [g_j]_i \ge \underline{C}_g a, \quad \|E_j^g\|_F \le \tau_g a,$$

for constants $\overline{C}_g, \underline{C}_g > 0$ *and* $\tau_g \ge 0$ *independent of* a*. Assume also that the old momentum satisfies*

$$\|M_{-L}\|_F \le C_M a, \qquad \|\mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L})\|_F \le \tau_m a,$$

for constants $C_M < \infty$ *and* $\tau_m \ge 0$ *independent of* a*. Then* M_0 *admits a decomposition* $M_0 = \mathcal{B}(m_0) + E_0^m$ *with* $E_0^m \in \mathtt{span}\{B_1, \cdots, B_n\}^\perp$*, where*

$$m_0 = \sum_{j=1}^{L} \xi^{j-1} g_j + \xi^L \, \mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big), \tag{79}$$

$$E_0^m = \sum_{j=1}^{L} \xi^{j-1} E_j^g + \xi^L \, \mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L}). \tag{80}$$

Here, $\mathcal{B}^{-1}$ *denotes the inverse of* $\mathcal{B}$ *restricted to the river subspace. Consequently,*

$$\|m_0\|_2 \le \overline{C}_g a \frac{1-\xi^L}{1-\xi} + C_M \xi^L a, \qquad \min_i [m_0]_i \ge \underline{C}_g a \frac{1-\xi^L}{1-\xi} - C_M \xi^L a,$$

and

$$\|E_0^m\|_F \le \tau_g a \frac{1-\xi^L}{1-\xi} + \tau_m \xi^L a.$$

Proof. We decompose every term in the EMA formula into its river component and its hill component. For each $j = 1, \cdots, L$, the assumed gradient decomposition is

$$\nabla \mathcal{F}(X_{-j}) = \mathcal{B}(g_j) + E_j^g, \qquad E_j^g \in \mathtt{span}\{B_1, \cdots, B_n\}^\perp.$$

Similarly, decompose the old momentum as

$$M_{-L} = \mathtt{Proj}_{\mathcal{R}}(M_{-L}) + \mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L}),$$

where $\mathtt{Proj}_{\mathcal{R}}(M_{-L}) \in \mathtt{span}\{B_1, \cdots, B_n\}$ and $\mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L}) \in \mathtt{span}\{B_1, \cdots, B_n\}^\perp$.

Substituting these decompositions into the EMA representation of M_0 and using the linearity of $\mathcal{B}$ and of the orthogonal projections, we obtain

$$\begin{aligned} M_0 &= \sum_{j=1}^{L} \xi^{j-1} \big(\mathcal{B}(g_j) + E_j^g\big) + \xi^L \Big(\mathtt{Proj}_{\mathcal{R}}(M_{-L}) + \mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L})\Big) \\ &= \mathcal{B}\left(\sum_{j=1}^{L} \xi^{j-1} g_j + \xi^L \, \mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big)\right) + \left(\sum_{j=1}^{L} \xi^{j-1} E_j^g + \xi^L \, \mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L})\right). \end{aligned}$$

Therefore M_0 admits the decomposition $M_0 = \mathcal{B}(m_0) + E_0^m$, where m_0 is (79) and E_0^m is (80). Because each E_j^g and $\mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L})$ lies in $\mathtt{span}\{B_1, \cdots, B_n\}^\perp$, the same is true for their weighted sum, i.e., $E_0^m \in \mathtt{span}\{B_1, \cdots, B_n\}^\perp$ as well.

It remains to verify the stated bounds.

First, since $\mathcal{B}$ is an isometric identification between $\mathbb{R}^n$ and the river subspace, we have

$$\left\|\mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(Y)\big)\right\|_2 = \|\mathtt{Proj}_{\mathcal{R}}(Y)\|_F \le \|Y\|_F \qquad \text{for every matrix } Y \in \mathbb{R}^{n\times n}.$$

Hence, using the triangle inequality together with the assumptions on g_j and M_{-L},

$$\begin{aligned}\|m_0\|_2 &\le \sum_{j=1}^{L} \xi^{j-1}\|g_j\|_2 + \xi^L\left\|\mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big)\right\|_2 \le \sum_{j=1}^{L} \xi^{j-1}\overline{C}_g a + \xi^L\|M_{-L}\|_F \\ &\le \overline{C}_g a \sum_{j=1}^{L}\xi^{j-1} + C_M\xi^L a = \overline{C}_g a\frac{1-\xi^L}{1-\xi} + C_M\xi^L a.\end{aligned}$$

Next, we prove the coordinatewise lower bound. Fix any coordinate i. By the definition of m_0,

$$[m_0]_i = \sum_{j=1}^{L}\xi^{j-1}[g_j]_i + \xi^L\left[\mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big)\right]_i.$$

Using the lower bound on $[g_j]_i$ and the estimate

$$\left|\left[\mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big)\right]_i\right| \le \left\|\mathcal{B}^{-1}\big(\mathtt{Proj}_{\mathcal{R}}(M_{-L})\big)\right\|_2 \le \|M_{-L}\|_F \le C_M a,$$

we obtain

$$[m_0]_i \ge \underline{C}_g a\sum_{j=1}^{L}\xi^{j-1} - C_M\xi^L a = \underline{C}_g a\frac{1-\xi^L}{1-\xi} - C_M\xi^L a.$$

Since this bound is uniform in i, it follows that

$$\min_i [m_0]_i \ge \underline{C}_g a\frac{1-\xi^L}{1-\xi} - C_M\xi^L a.$$

Finally, for the off-river hill remainder, the triangle inequality gives

$$\begin{aligned}\|E_0^m\|_F &\le \sum_{j=1}^{L}\xi^{j-1}\|E_j^g\|_F + \xi^L\|\mathtt{Proj}_{\mathcal{R}^\perp}(M_{-L})\|_F \\ &\le \tau_g a\sum_{j=1}^{L}\xi^{j-1} + \tau_m\xi^L a = \tau_g a\frac{1-\xi^L}{1-\xi} + \tau_m\xi^L a.\end{aligned}$$

This proves the stated bound on E_0^m and completes the proof. □

Structural hypotheses. We now state the structural hypotheses used in Theorem B.18.

The first assumption is standard: it requires a nondegenerate local minimizer together with quantitative Taylor control near $X^\star$.

Assumption B.14 (Local quadratic model with third-order control)**.** *The objective $\mathcal{F} : \mathbb{R}^{n\times n} \to \mathbb{R}$ is C^3 in a neighborhood of $X^\star$, and $\nabla\mathcal{F}(X^\star) = 0$. Write $H_\star := \nabla^2\mathcal{F}(X^\star)$. There exist constants $0 < \underline{L} \le \overline{L} < \infty$ such that*

$$\underline{L}\|Y\|_F^2 \le \langle Y, H_\star[Y]\rangle \le \overline{L}\|Y\|_F^2 \qquad \textit{for all } Y \in \mathbb{R}^{n\times n}. \tag{81}$$

Moreover, for some constant $C_3 < \infty$ and all sufficiently small Y,

$$\|\nabla\mathcal{F}(X^\star + Y) - H_\star[Y]\|_F \le C_3\|Y\|_F^2,$$

and

$$|\mathcal{F}(X^\star + Y) - \mathcal{F}(X^\star) - \langle Y, H_\star[Y]\rangle/2| \le C_3\|Y\|_F^3.$$

Assumption B.15 (Hessian river-to-hill leakage bound)**.** *Let K be the matrix introduced earlier (see B.7) and assume $K \succ 0$. There exist constants $C_y < \infty$ and $\tau_h \geq 0$ such that for every $y \in \mathbb{R}^n$ with $\|y\|_2 \leq C_y$,*

$$H_\star[\mathcal{B}(y)] = \mathcal{B}(Ky) + E^h(y), \qquad \texttt{Proj}_{\mathcal{R}}(E^h(y)) = 0, \qquad \|E^h(y)\|_F \leq \tau_h \|y\|_2. \tag{82}$$

Assumption B.16 (Late-stage initialization inside a tube)**.** *Let $a > 0$. Assume the initial iterate decomposes as*

$$X_0 = X^\star + \mathcal{B}(ax_0) + E_0^x, \qquad \texttt{Proj}_{\mathcal{R}}(E_0^x) = 0, \qquad \|E_0^x\|_F \leq \tau_x a, \tag{83}$$

for some $x_0 \in \mathbb{R}^n$ satisfying $Kx_0 \succ \mathbf{0}$. Assume the stored momentum decomposes as

$$M_0 = \mathcal{B}(m_0) + E_0^m, \qquad \|m_0\|_2 \leq C_m a, \qquad \|E_0^m\|_F \leq \tau_m a, \tag{84}$$

for constants $C_m < \infty$ and $\tau_m \geq 0$ independent of a.

Lemma B.17 (Polar stability near a positive river matrix)**.** *Let $y \succ \mathbf{0}$ and define $Y = \mathcal{B}(y) + E^y$. If $\|E^y\|_2 \leq \frac{1}{2} \min_i [y]_i$, then Y has full column rank, and*

$$\|\texttt{msign}(Y) - \texttt{msign}(\mathcal{B}(y))\|_F \leq \frac{4}{3 \min_i [y]_i} \|E^y\|_F. \tag{85}$$

Proof. The unperturbed matrix $\mathcal{B}(y) = U_\star \texttt{Diag}(y) V_\star^\top$ has singular values exactly $[y]_1, \cdots, [y]_n$, and hence $\sigma_{\min}(\mathcal{B}(y)) = \min_i [y]_i$. By Weyl's inequality Stewart [1991],

$$\sigma_{\min}(Y) \geq \sigma_{\min}(\mathcal{B}(y)) - \|E^y\|_2 \geq \frac{1}{2} \min_i [y]_i > 0,$$

so Y has full column rank.

We now apply the standard perturbation bound for the unit polar factor of full-column-rank matrices:

$$\|\texttt{msign}(Y) - \texttt{msign}(\mathcal{B}(y))\|_F \leq \frac{2}{\sigma_{\min}(Y) + \sigma_{\min}(\mathcal{B}(y))} \|E^y\|_F.$$

Using the lower bounds above gives

$$\frac{2}{\sigma_{\min}(Y) + \sigma_{\min}(\mathcal{B}(y))} \leq \frac{2}{\frac{1}{2} \min_i [y]_i + \min_i [y]_i} = \frac{4}{3 \min_i [y]_i},$$

and therefore

$$\|\texttt{msign}(Y) - \texttt{msign}(\mathcal{B}(y))\|_F \leq \frac{4}{3 \min_i [y]_i} \|E^y\|_F.$$

This proves the claim. □

Local step size and positivity margins. Define

$$y_0 := \frac{\mathbf{1}^\top K x_0}{\mathbf{1}^\top K \mathbf{1}} \qquad \text{and} \qquad \eta^{\text{Muon}} := (1 + \varepsilon) a y_0, \tag{86}$$

where $0 < \varepsilon < 1$ is fixed. Define the scaled positivity margins

$$\kappa_1 := \min_i \frac{\xi [m_0]_i + a [K x_0]_i}{a}, \tag{87}$$

and

$$\kappa_2 := \min_i \left[\frac{\xi\big(\xi m_0 + aKx_0\big) + aK\big(x_0 - (1+\varepsilon) y_0 \mathbf{1}\big)}{a} \right]_i. \tag{88}$$

Theorem B.18 (Late-stage Muon overshoot along the spectral river (two-steps analysis))**.** *Assume Assumptions B.14, B.15, and B.16. Fix $\xi \in (0,1)$ and $\varepsilon \in (0,1)$, and define η^{Muon} by* (86)*. Assume that κ_1 and κ_2 defined in* (87) *and 88 are bounded below by positive constants independent of a. Then there exist constants $a_0 > 0$, $\tau_0^h > 0$, $\tau_0^m > 0$, $\tau_0^x > 0$, and $\eta_0 > 0$ (depending only on the local constants and on $(K, \xi, x_0, C_m, \varepsilon)$, but not on a) such that whenever*

$$0 < a \leq a_0, \qquad 0 \leq \tau_h \leq \tau_0^h, \qquad 0 \leq \tau_m \leq \tau_0^m, \qquad 0 \leq \tau_x \leq \tau_0^x,$$

the Muon iterates and momentum-GD iterates initialized from the same pair (X_0, M_0) satisfy

$$\mathcal{F}(X_1^{\text{Muon}}) < \mathcal{F}(X_0) < \mathcal{F}(X_2^{\text{Muon}}), \qquad \mathcal{F}(X_2^{\text{GD}}) < \mathcal{F}(X_0), \tag{89}$$

for every GD step size $0 < \eta^{\text{GD}} \leq \eta_0$. In particular, since $\eta^{\text{Muon}} = O(a)$ while $\eta_0 = O(1)$, we have $\eta^{\text{Muon}} < \eta_0$ for all sufficiently small a.

Proof. We prove (89) in three steps.

Step 1: first Muon step decreases. From (83), write $Y_0 := X_0 - X^\star = \mathcal{B}(ax_0) + E_0^x$ with $\|E_0^x\|_F \le \tau_x a$. Assumption B.14 gives

$$\nabla\mathcal{F}(X_0) = H_\star[Y_0] + E_0^g, \qquad \|E_0^g\|_F \le C_3\|Y_0\|_F^2.$$

Using (82) with $y = ax_0$, linearity of $H_\star$, and (81), we obtain

$$\nabla\mathcal{F}(X_0) = a\,\mathcal{B}(Kx_0) + E_0^\nabla, \qquad \|E_0^\nabla\|_F \le \bar{C}(\tau_h + \tau_x)a + \bar{C}a^2,$$

where $\bar{C}$ depends only on the local constants and on $\|x_0\|_2$. Throughout this proof, $\bar{C}$ denotes a generic positive constant, possibly changing from line to line, whose value depends only on fixed problem constants and is independent of a, τ_h, τ_m, and τ_x. Combining the result with (84) yields

$$M_1^{\mathrm{Muon}} = \xi M_0 + \nabla\mathcal{F}(X_0) = \mathcal{B}\big(\xi m_0 + aKx_0\big) + E_1^m, \quad \|E_1^m\|_F \le \bar{C}(\tau_m + \tau_h + \tau_x)a + \bar{C}a^2.$$

By (87),

$$\min_i [\xi m_0 + aKx_0]_i \ge a\kappa_1.$$

Thus, shrinking $a_0, \tau_0^h, \tau_0^m, \tau_0^x$ if needed so that $\|E_1^m\|_2 \le \frac{1}{2}a\kappa_1$, Lemma B.17 implies

$$\mathtt{msign}(M_1^{\mathrm{Muon}}) = \mathcal{B}(\mathbf{1}) + \Delta_1, \quad \|\Delta_1\|_F \le \bar{C}(\tau_m + \tau_h + \tau_x + a). \tag{90}$$

Hence with $\|E_1^x\|_F \le \bar{C}a(\tau_m + \tau_h + \tau_x + a)$, we have

$$X_1^{\mathrm{Muon}} = X_0 - \eta^{\mathrm{Muon}}\mathtt{msign}(M_1^{\mathrm{Muon}}) = X^\star + \mathcal{B}(ax_0 - \eta^{\mathrm{Muon}}\mathbf{1}) + E_1^x.$$

Using the loss expansion in Assumption B.14 at $X^\star$ and the identity

$$\langle \mathcal{B}(y), H_\star[\mathcal{B}(y)]\rangle = y^\top K y + \langle \mathcal{B}(y), E^h(y)\rangle = y^\top K y,$$

we obtain

$$\mathcal{F}(X_1^{\mathrm{Muon}}) - \mathcal{F}(X_0) = -\eta^{\mathrm{Muon}} a\,\mathbf{1}^\top K x_0 + \frac{1}{2}(\eta^{\mathrm{Muon}})^2 \mathbf{1}^\top K \mathbf{1} + O\big(a^2(\tau_m + \tau_h + \tau_x + a)\big).$$

Substituting $\eta^{\mathrm{Muon}} = (1+\varepsilon)ay_0$ and $\mathbf{1}^\top K x_0 = y_0\,\mathbf{1}^\top K\mathbf{1}$ yields a strictly negative leading term:

$$\mathcal{F}(X_1^{\mathrm{Muon}}) - \mathcal{F}(X_0) = -\frac{1}{2}(1-\varepsilon^2)a^2 y_0^2\,\mathbf{1}^\top K\mathbf{1} + O\big(a^2(\tau_m + \tau_h + \tau_x + a)\big).$$

Choosing $a_0, \tau_0^h, \tau_0^m, \tau_0^x$ so that the error term is dominated by the leading negative term gives

$$\mathcal{F}(X_1^{\mathrm{Muon}}) < \mathcal{F}(X_0).$$

Step 2: second Muon step overshoots. Repeat the previous gradient expansion at X_1^{Muon}. From the representation of $X_1^{\mathrm{Muon}} - X^\star$ and Assumptions B.14–B.15, we obtain

$$\nabla\mathcal{F}(X_1^{\mathrm{Muon}}) = a\,\mathcal{B}\big(K(x_0 - (1+\varepsilon)y_0\mathbf{1})\big) + E_1^\nabla, \quad \|E_1^\nabla\|_F \le \bar{C}(\tau_h + \tau_m + \tau_x)a + \bar{C}a^2.$$

Therefore

$$M_2^{\mathrm{Muon}} = \xi M_1^{\mathrm{Muon}} + \nabla\mathcal{F}(X_1^{\mathrm{Muon}}) = \mathcal{B}(\bar{m}_1) + E_2^m, \quad \|E_2^m\|_F \le \bar{C}(\tau_h + \tau_m + \tau_x)a + \bar{C}a^2,$$

where

$$\bar{m}_1 := \xi(\xi m_0 + aKx_0) + aK\big(x_0 - (1+\varepsilon)y_0\mathbf{1}\big).$$

By (88), we know $\min_i[\bar{m}_1]_i \ge a\kappa_2$. Shrinking $a_0, \tau_0^h, \tau_0^m, \tau_0^x$ if necessary so that $\|E_2^m\|_2 \le \frac{1}{2}a\kappa_2$, Lemma B.17 yields

$$\mathtt{msign}(M_2^{\mathrm{Muon}}) = \mathcal{B}(\mathbf{1}) + \Delta_2, \quad \|\Delta_2\|_F \le \bar{C}(\tau_h + \tau_m + \tau_x + a).$$

Hence with $\|E_2^x\|_F \le \bar{C}a(\tau_h + \tau_m + \tau_x + a)$, we have

$$X_2^{\mathrm{Muon}} = X_1^{\mathrm{Muon}} - \eta^{\mathrm{Muon}}\mathtt{msign}(M_2^{\mathrm{Muon}}) = X^\star + \mathcal{B}(ax_0 - 2\eta^{\mathrm{Muon}}\mathbf{1}) + E_2^x.$$

Using again the loss expansion around $X^\star$ and the leakage bound (82), we obtain

$$\mathcal{F}(X_2^{\mathrm{Muon}}) - \mathcal{F}(X_0) = -2\eta^{\mathrm{Muon}} a\,\mathbf{1}^\top K x_0 + 2(\eta^{\mathrm{Muon}})^2 \mathbf{1}^\top K\mathbf{1} + O\big(a^2(\tau_h + \tau_m + \tau_x + a)\big).$$

Substituting $\eta^{\mathrm{Muon}} = (1+\varepsilon)ay_0$ and $\mathbf{1}^\top K x_0 = y_0\,\mathbf{1}^\top K\mathbf{1}$ yields a strictly positive leading term:

$$\mathcal{F}(X_2^{\mathrm{Muon}}) - \mathcal{F}(X_0) = 2\varepsilon(1+\varepsilon)a^2 y_0^2\,\mathbf{1}^\top K\mathbf{1} + O\big(a^2(\tau_h + \tau_m + \tau_x + a)\big).$$

Choosing $a_0, \tau_0^h, \tau_0^m, \tau_0^x$ sufficiently small ensures the error term is dominated by the positive leading term, hence

$$\mathcal{F}(X_2^{\mathrm{Muon}}) > \mathcal{F}(X_0).$$

Step 3: two-step momentum GD decreases for a small constant step size. Consider momentum GD initialized from the same (X_0, M_0):

$$M_1^{\mathrm{GD}} = \xi M_0 + \nabla\mathcal{F}(X_0), \qquad X_1^{\mathrm{GD}} = X_0 - \eta^{\mathrm{GD}} M_1^{\mathrm{GD}},$$
$$M_2^{\mathrm{GD}} = \xi M_1^{\mathrm{GD}} + \nabla\mathcal{F}(X_1^{\mathrm{GD}}), \qquad X_2^{\mathrm{GD}} = X_1^{\mathrm{GD}} - \eta^{\mathrm{GD}} M_2^{\mathrm{GD}}.$$

We show that there exists $\eta_0 = O(1)$ such that $\mathcal{F}(X_2^{\mathrm{GD}}) < \mathcal{F}(X_0)$ for all $0 < \eta^{\mathrm{GD}} \le \eta_0$ and all sufficiently small $(a, \tau_h, \tau_m, \tau_x)$.

Define the quadratic model function

$$Q(Y) := \frac{1}{2}\langle Y, H_\star[Y]\rangle.$$

By Assumption B.14, for all iterates staying in the local neighborhood,

$$\mathcal{F}(X^\star + Y) = \mathcal{F}(X^\star) + Q(Y) + O(\|Y\|_F^3),$$

and

$$\nabla\mathcal{F}(X^\star + Y) = H_\star[Y] + O(\|Y\|_F^2).$$

Since $X_0 - X^\star = O(a)$ and we only take two steps with $\eta^{\mathrm{GD}} = O(1)$, by choosing a_0 sufficiently small we ensure $\|X_t^{\mathrm{GD}} - X^\star\|_F = O(a)$ for $t = 0, 1, 2$, and thus the Taylor remainders contribute at most $O(a^3)$ to function values and $O(a^2)$ to gradients.

We now compare the true two-step map with the quadratic dynamics obtained by replacing $\nabla\mathcal{F}(X)$ by $H_\star[X - X^\star]$. Let $\tilde{X}_t$ denote this quadratic two-step sequence (initialized at the same $\tilde{X}_0 = X_0$ and $\tilde{M}_0 = M_0$):

$$\tilde{M}_1 = \xi\tilde{M}_0 + H_\star[\tilde{X}_0 - X^\star], \qquad \tilde{X}_1 = \tilde{X}_0 - \eta^{\mathrm{GD}}\tilde{M}_1,$$
$$\tilde{M}_2 = \xi\tilde{M}_1 + H_\star[\tilde{X}_1 - X^\star], \qquad \tilde{X}_2 = \tilde{X}_1 - \eta^{\mathrm{GD}}\tilde{M}_2.$$

Because all operators are linear in this quadratic model, $\tilde{X}_2 - X^\star$ is an affine polynomial in η^{GD} whose coefficients depend continuously on $(\tilde{X}_0 - X^\star, \tilde{M}_0)$. In particular, the map $\eta^{\mathrm{GD}} \mapsto Q(\tilde{X}_2 - X^\star)$ is a polynomial in η^{GD}.

We claim that the directional derivative at $\eta^{\mathrm{GD}} = 0$ is strictly negative uniformly over all admissible (X_0, M_0) in Assumption B.16. At $\eta^{\mathrm{GD}} = 0$, we have $\tilde{X}_1 = \tilde{X}_0$ and hence

$$\frac{d}{d\eta^{\mathrm{GD}}} Q(\tilde{X}_2 - X^\star)\Big|_{\eta^{\mathrm{GD}}=0} = -\left\langle \tilde{X}_0 - X^\star,\ H_\star\Big[\xi(1+\xi)\tilde{M}_0 + (2+\xi)H_\star[\tilde{X}_0 - X^\star]\Big]\right\rangle.$$

Using $\tilde{X}_0 - X^\star = \mathcal{B}(ax_0) + E_0^x$ and $\tilde{M}_0 = \mathcal{B}(m_0) + E_0^m$, expand the inner products. The leading river contribution equals

$$-a^2(2+\xi)\,(Kx_0)^\top Kx_0 - a\,\xi(1+\xi)\,(Kx_0)^\top m_0,$$

which is strictly negative because $K \succ 0$, $Kx_0 \succ \mathbf{0}$, and m_0 is $O(a)$ with nonnegative coordinates in the late-stage regime (captured by the positivity margin in (87)). All remaining terms involve at least one hill component (E_0^x or E_0^m) and are therefore bounded by

$$O\big(a^2(\tau_x + \tau_m + \tau_h)\big),$$

using (81), 82, and Cauchy–Schwarz. Thus, for sufficiently small $\tau_x + \tau_m + \tau_h$, the derivative remains strictly negative uniformly over the admissible set.

Since $Q(\tilde{X}_2 - X^\star)$ is a polynomial in η^{GD} with coefficients bounded uniformly over the admissible set (because $\|X_0 - X^\star\|_F = O(a)$, $\|M_0\|_F = O(a)$, and $H_\star$ is bounded), there exists $\eta_0 > 0$ independent of a such that

$$Q(\tilde{X}_2 - X^\star) < Q(X_0 - X^\star) \qquad \text{for all } 0 < \eta^{\mathrm{GD}} \le \eta_0.$$

Finally, the difference between the true iterate X_2^{GD} and the quadratic iterate $\tilde{X}_2$ over two steps is $O(a^2)$ in norm, because each gradient differs from its linearization by $O(a^2)$ and the step size is $O(1)$. Therefore,

$$\mathcal{F}(X_2^{\mathrm{GD}}) - \mathcal{F}(X_0) = \Big(Q(\widetilde{X}_2 - X^\star) - Q(X_0 - X^\star)\Big) + O\big(a^2(\tau_h + \tau_m + \tau_x) + a^3\big),$$

and by shrinking $a_0, \tau_0^h, \tau_0^m, \tau_0^x$ if necessary, the negative quadratic decrease dominates the perturbation, giving

$$\mathcal{F}(X_2^{\mathrm{GD}}) < \mathcal{F}(X_0) \qquad \text{for all } 0 < \eta^{\mathrm{GD}} \le \eta_0.$$

Combining Steps 1–3 proves (89). □

B.4 Implications for Algorithm Design

B.4.1 Late-stage Adam behaves more like gradient methods than Muon

We finally record an Adam-like comparison. The purpose of this proposition is not to analyze the full Adam recursion, but rather to show that, once the diagonal preconditioner has stabilized, the resulting short-horizon behavior is qualitatively closer to gradient-based refinement than to Muon's sign-type dynamics. As a transparent late-stage surrogate for Adam, we consider a momentum method with a frozen positive diagonal preconditioner:

$$M_{t+1}^{\mathrm{Adam}} = \xi M_t^{\mathrm{Adam}} + \nabla \mathcal{F}(X_t^{\mathrm{Adam}}), \qquad X_{t+1}^{\mathrm{Adam}} = X_t^{\mathrm{Adam}} - \eta^{\mathrm{Adam}} P^{-1} M_{t+1}^{\mathrm{Adam}}, \tag{91}$$

where

$$P = \mathtt{Diag}(p_1, \cdots, p_{n^2}), \qquad 0 < \underline{p} \le p_i \le \overline{p}, \qquad i = 1, 2, \cdots, n^2.$$

This model is intended to approximate the late-stage behavior of Adam over a short time horizon after the second-moment denominator has effectively stabilized.

The next proposition shows that, under a frozen positive diagonal preconditioner and mild alignment and size conditions, the first two Adam-like steps are both descent steps, with an admissible step size that is independent of the residual scale a.

Proposition B.19 (Two-step descent for an Adam-like method)**.** *Assume that $\mathcal{F}$ is L-smooth on a neighborhood containing the iterates under consideration, i.e.,*

$$\|\nabla \mathcal{F}(X + \Delta X) - \nabla \mathcal{F}(X)\|_F \le L \|\Delta X\|_F \qquad \textit{for all } X, \Delta X \textit{ in the neighborhood.}$$

Let $(X_t^{\mathrm{Adam}}, M_t^{\mathrm{Adam}})$ evolve according to (91)*, and suppose that for $t = 0, 1$,*

$$\begin{aligned} \left\langle \nabla \mathcal{F}(X_t^{\mathrm{Adam}}), P^{-1} M_{t+1}^{\mathrm{Adam}} \right\rangle &\ge \underline{C}_A \, \|\nabla \mathcal{F}(X_t^{\mathrm{Adam}})\|_F^2, \\ \|M_{t+1}^{\mathrm{Adam}}\|_F &\le \overline{C}_A \, \|\nabla \mathcal{F}(X_t^{\mathrm{Adam}})\|_F, \end{aligned}$$

for some constants $\underline{C}_A, \overline{C}_A > 0$ independent of the residual scale a. Then every step size

$$0 < \eta^{\mathrm{Adam}} \le \eta_0 := \frac{2 \underline{C}_A \underline{p}^2}{L \overline{C}_A^2}$$

guarantees

$$\mathcal{F}(X_1^{\mathrm{Adam}}) < \mathcal{F}(X_0^{\mathrm{Adam}}), \qquad \mathcal{F}(X_2^{\mathrm{Adam}}) < \mathcal{F}(X_1^{\mathrm{Adam}}). \tag{92}$$

In particular, the admissible Adam-like step size is independent of the late-stage residual scale.

Proof. By the descent lemma for L-smooth functions, for each $t = 0, 1$,

$$\mathcal{F}(X_t^{\mathrm{Adam}} - \eta^{\mathrm{Adam}} D_t) \le \mathcal{F}(X_t^{\mathrm{Adam}}) - \eta^{\mathrm{Adam}} \langle \nabla \mathcal{F}(X_t^{\mathrm{Adam}}), D_t \rangle + \frac{L (\eta^{\mathrm{Adam}})^2}{2} \|D_t\|_F^2,$$

where $D_t := P^{-1} M_{t+1}^{\mathrm{Adam}}$. Since $X_{t+1}^{\mathrm{Adam}} = X_t^{\mathrm{Adam}} - \eta^{\mathrm{Adam}} D_t$, we obtain

$$\begin{aligned} &\mathcal{F}(X_{t+1}^{\mathrm{Adam}}) - \mathcal{F}(X_t^{\mathrm{Adam}}) \\ \le& - \eta^{\mathrm{Adam}} \left\langle \nabla \mathcal{F}(X_t^{\mathrm{Adam}}), P^{-1} M_{t+1}^{\mathrm{Adam}} \right\rangle + \frac{L (\eta^{\mathrm{Adam}})^2}{2} \|P^{-1} M_{t+1}^{\mathrm{Adam}}\|_F^2 \\ \le& - \eta^{\mathrm{Adam}} \underline{C}_A \|\nabla \mathcal{F}(X_t^{\mathrm{Adam}})\|_F^2 + \frac{L (\eta^{\mathrm{Adam}})^2}{2 \underline{p}^2} \|M_{t+1}^{\mathrm{Adam}}\|_F^2 \\ \le& \left(-\eta^{\mathrm{Adam}} \underline{C}_A + \frac{L (\eta^{\mathrm{Adam}})^2 \overline{C}_A^2}{2 \underline{p}^2} \right) \|\nabla \mathcal{F}(X_t^{\mathrm{Adam}})\|_F^2. \end{aligned}$$

Therefore, whenever $0 < \eta^{\mathrm{Adam}} \le \eta_0$, the coefficient in parentheses is negative, and hence

$$\mathcal{F}(X_{t+1}^{\mathrm{Adam}}) < \mathcal{F}(X_t^{\mathrm{Adam}}) \qquad \text{for } t = 0, 1.$$

This proves (92). □

B.4.2 Guidance for algorithm design

The preceding analysis suggests a simple principle for optimizer design: the update geometry that is beneficial in the early stage need not remain appropriate in the late stage.

At a river point

$$X_t = X^\star + \mathcal{B}(x_t),$$

the natural positive-chamber Muon scale induced by the local quadratic model is

$$y_t := \frac{\mathbf{1}^\top K x_t}{\mathbf{1}^\top K \mathbf{1}}. \tag{93}$$

If $x_t = a_t x_0$, then the safe displacement scale for Muon contracts proportionally to the current river residual. Consequently, a schedule depending only on time, such as cosine decay,

$$\eta_t^{\mathrm{Muon}} = \frac{\eta_{\max}}{2}\left(1 + \cos\frac{\pi t}{T}\right), \qquad 0 \le t \le T, \tag{94}$$

can become mismatched with the actual late-stage geometry in two distinct ways.

First, if η_t^{Muon} is too large relative to y_t, then Muon may step across the minimizer, after which the momentum can continue pushing in essentially the same polar direction and produce overshoot. Second, if η_t^{Muon} is too small relative to y_t, then Muon still moves by additive sign-type increments, but now does so unnecessarily slowly along the river.

This mismatch is already visible in the one-dimensional quadratic model: $\frac{1}{2}x^2$. A sign-based or Muon-like step takes the form $x \leftarrow x - \eta_t^{\mathrm{Muon}}\,\mathtt{sign}(x)$, which increases the objective whenever

$$0 < |x| < \eta_t^{\mathrm{Muon}}/2.$$

By contrast, gradient descent with step size 1 sends x to 0 in one step, and the Adam-like model above also admits an $O(1)$ admissible step size under a stabilized positive preconditioner.

Switching away from Muon is therefore not merely another learning-rate schedule; it changes the geometry of the update itself. Muon intentionally suppresses gradient magnitude information through `msign`, which is advantageous when the goal is rapid spectral alignment and coarse river tracking. However, in the final river-refinement regime, the residual magnitude contains precisely the information needed to calibrate the step and avoid overshoot. Gradient methods preserve this information through the raw gradient and momentum, while Adam-like methods preserve it through a positive diagonal preconditioner. This suggests that a Muon-to-GD or Muon-to-Adam transition can serve as a residual-adaptive alternative to relying solely on a manually prescribed decay schedule.

C Additional Details for Section 6

C.1 Additional Case Studies

C.1.1 2D Anisotropic Spectral Slice Case

We consider a simple two-dimensional anisotropic spectral landscape as a motivating example for the phenomena discussed in the main text. Specifically, we choose two orthogonal rank-one spectral directions $U = uu^\top$ and $V = vv^\top$ with $\langle U, V\rangle_F = 0$. This allows us to restrict the matrix variable to the two-dimensional spectral slice $X(a,b) = aU + bV$, where $a, b \in \mathbb{R}$ are scalar coordinates. On this slice, we define an anisotropic quadratic objective

$$f(a,b) = \frac{1}{2}\lambda_w(a-a^\star)^2 + \frac{1}{2}\lambda_s(b-b^\star)^2, \qquad \lambda_s \gg \lambda_w.$$

Equivalently, lifting the gradient back to the ambient matrix space gives

$$\nabla_X f = \lambda_w(a-a^\star)U + \lambda_s(b-b^\star)V.$$

This toy problem can be viewed as a "no-bulk" version of the rank-one mixed-spectrum matrix sensing model studied in the paper. In our experiments, we take

$$u = [1,1]^\top, \qquad v = [1,-1]^\top, \qquad \lambda_w = 0.6, \qquad \lambda_s = 10.0.$$

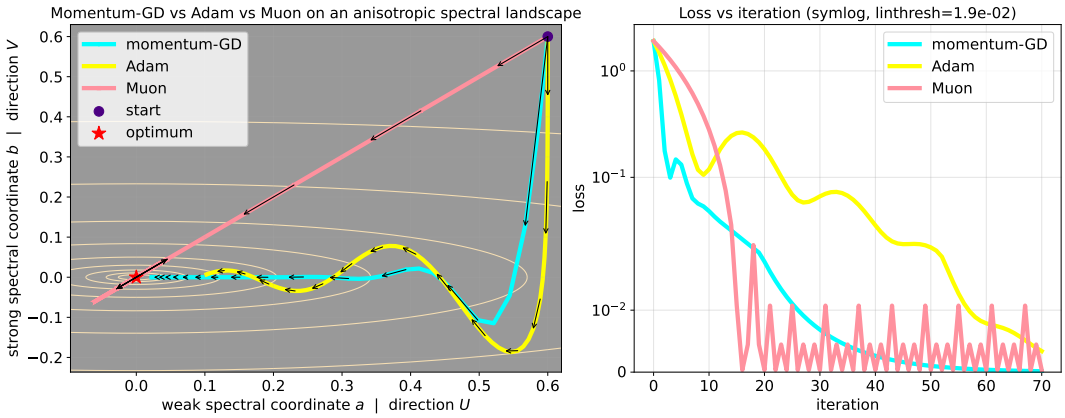

Figure 5: **2D anisotropic spectral slice case: pure optimizers.** Left: optimization trajectories of momentum-GD, Adam, and Muon under a fixed learning rate 0.037. The black arrows indicate the displacement every other iteration. Right: loss versus iteration for the three optimizers.

so that the two rank-one directions are orthogonal in Frobenius inner product. We initialize the optimization at $(a_0, b_0) = (0.6, 0.6)$, and set the ground-truth point to $(a^\star, b^\star) = (0, 0)$. Since the dynamics are fully described by the two coordinates (a, b), we can directly visualize the optimization trajectories on the two-dimensional contour plot of the objective.

Figure 5 reports the trajectories and loss curves for momentum-GD, Adam, and Muon under a fixed learning rate. We observe that momentum-GD and Adam move relatively slowly along the weak-curvature direction and approach the ground truth only gradually within 70 iterations. In contrast, Muon reaches the neighborhood of the ground truth much earlier, confirming its fast early-stage progress in anisotropic spectral landscapes (corroborate Theorem B.10). However, the loss curve reveals a complementary drawback: with a fixed learning rate, Muon repeatedly overshoots around the ground-truth point and exhibits persistent oscillations, leading to substantially worse late-stage convergence (corroborate Theorem B.18). By comparison, momentum-GD and Adam converge more accurately once they enter the local basin. This example also illustrates that, in the late stage, Adam behaves more similarly to GD-type methods than to Muon (corroborate Proposition B.19).

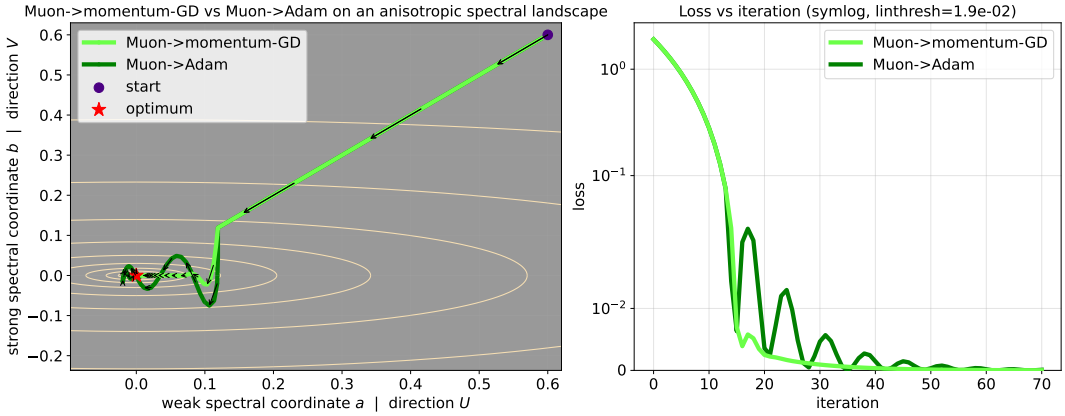

Figure 6: **2D anisotropic spectral slice case: hybrid optimizers.** Left: optimization trajectories of the two hybrid schemes under a fixed learning rate 0.037. The black arrows indicate the displacement every other iteration. Right: loss versus iteration for the pure and hybrid optimizers.

Overall, the pure-optimizer experiment is consistent with our intuition: Muon achieves rapid early-stage progress but poor final accuracy, while momentum-GD and Adam converge more slowly but more precisely in late-stage. Motivated by the intuition, we further evaluate two hybrid optimization schemes: "Muon $\to$ momentum-GD" and "Muon $\to$ Adam". In both cases, we switch from Muon to the second optimizer at iteration 13, at which point Muon has already reached a small neighborhood of the ground truth. The results are shown in Figure 6. Both hybrid methods inherit

the fast initial movement of Muon and the accurate late-stage convergence of momentum-GD or Adam. Consequently, they outperform the corresponding pure optimizers in terms of both speed and final accuracy. This toy example motivates (also corroborates) the analysis in Section 5, where the river-valley perspective is used as a tool for understanding generalized optimization settings.

C.1.2 Mixed-Spectrum Matrix Sensing Case

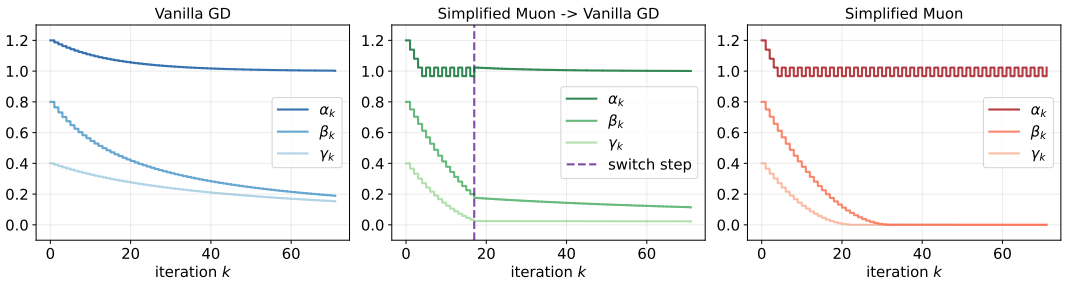

Figure 7: **Mixed-spectrum MS with a diagonal interaction matrix $\hat{K}$.** Evolution of the reduced coefficients $\alpha_k, \beta_k, \gamma_k$ versus iteration. Left: pure GD. Middle: hybrid optimization, which switches from simplified Muon to vanilla GD. Right: pure Muon.

Diagonal case. We first consider the diagonal reduced dynamics of the mixed-spectrum matrix sensing model. The coefficient iterations for this example are illustrated in Figure 7. In this experiment, we set $\hat{K} = \texttt{Diag}(0.5, 0.5, 0.5)$, and run the continuous-time dynamics up to time $T = 2.0$. The hybrid optimizer switches from Muon to GD at time $t = 0.45$, since at this point the Muon trajectory has already reached a small neighborhood of the ground-truth solution (PSD case has the same solution)

$$(\alpha^\star, \beta^\star, \gamma^\star) = (1, 0, 0).$$

The discrete trajectories $\{\alpha_k, \beta_k, \gamma_k\}$ are obtained by discretizing the reduced continuous dynamics in Proposition 4.1. We initialize the iterates at $\alpha_0 = 1.2$, $\beta_0 = 0.8$, $\gamma_0 = 0.4$ and use stepsize 0.028.

The rightmost panel shows that Muon rapidly approaches the ground truth, but its late-stage convergence is inaccurate. In particular, the signal coordinate α oscillates around the target value $\alpha^\star = 1$ (which also corresponds to the river). Note that the nuisance coordinates β and γ do not visibly oscillate below zero in the plot because, for visualization purposes, we apply a $\max\{\cdot, 0\}$ truncation (also used in PSD case) when evaluating the square-root expressions appearing in the discrete simulation and closed-form trajectories. This truncation prevents negative plotted values, but it should not be interpreted as evidence that Muon attains high-accuracy convergence in the β and γ directions.

By contrast, vanilla GD converges more slowly. In the diagonal case, the GD reduced dynamics are fully decoupled and admit the closed-form solutions in (5), which exhibit inverse-logistic or inverse-polynomial type convergence. In particular, β and γ decay only at a $1/t$-type rate. For simplified Muon, the corresponding diagonal reduced dynamics take the form in (6). Hence, in the idealized continuous-time diagonal model, Muon moves each coordinate toward its target at an approximately constant speed in the square-root scale. This explains its rapid early progress. However, under discretization with a fixed stepsize, the same aggressive update can lead to overshooting and oscillatory late-stage behavior near the optimum. The middle panel of Figure 7 shows the performance of the hybrid scheme. The hybrid method combines the fast early-stage movement of simplified Muon with the accurate late-stage convergence of vanilla GD. Consequently, it reaches the neighborhood of the ground truth quickly and then refines the solution more stably than pure Muon. In this sense, the hybrid method improves over both pure optimizers.

We emphasize that this diagonal example is fully decoupled: the river coordinate α evolves independently of the hill coordinates β and γ. As a result, the river progress and hill progress are not synchronized. It is therefore possible for the signal coordinate to reach the ground-truth neighborhood while the nuisance coordinates remain comparatively far from zero. For this reason, the diagonal case does not fall under the coupled river-hill setting covered by Theorem 4.2 (they even have different river definitions). Nevertheless, it remains a useful sanity check for the main intuition of the paper: Muon is best viewed as an early-stage exploration optimizer that rapidly moves

along some spectral directions, while GD-type methods are better suited for high-accuracy late-stage refinement.

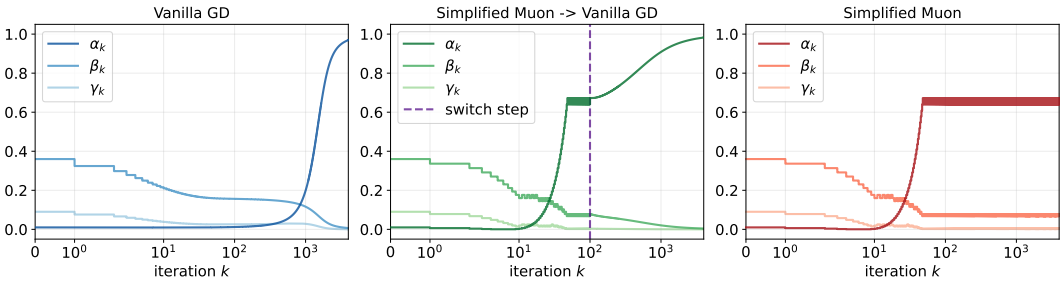

Figure 8: **Mixed-spectrum MS with a PSD interaction matrix $\hat{K}$.** Evolution of the reduced coefficients $\alpha_k, \beta_k, \gamma_k$ versus iteration (horizontal axis plotted in the symmetric-log scale). Left: pure GD. Middle: hybrid optimization, which switches from simplified Muon to vanilla GD. Right: pure Muon.

PSD case. We next consider a more general positive semidefinite interaction matrix $\hat{K}$, beyond the diagonal setting discussed above. Unlike the diagonal case, the coordinates are now coupled through $\hat{K}$, and this example falls within the setting of Theorem 4.2. In the experiment, we choose a rank-1 positive semidefinite matrix $\hat{K}$ (the outer product of $[0.1, 0.5, 0.8]$ with itself). We simulate the GD dynamics using the iteration in (38), and the Muon dynamics using the iteration in (45). Both methods use the same stepsize 0.02. The initial condition is $(\alpha_0, \beta_0, \gamma_0) = (0.01, 0.36, 0.09)$, and all methods are run for 4000 iterations. For the hybrid method, we switch from Muon to GD at iteration 100, since the Muon trajectory has already entered an oscillatory regime by that time.

The results are shown in Figure 8. We observe that pure Muon makes extremely rapid initial progress: within fewer than 100 iterations, it reaches a neighborhood of a local basin close to the ground-truth solution. However, after this fast transient phase, its trajectory begins to oscillate rather than refine. In particular, the coefficient α_k starts oscillating around approximately 0.6, and the nuisance coefficient β_k does not oscillate near zero. This indicates that, under the chosen fixed stepsize, Muon can become trapped in an oscillatory neighborhood of a local minimum: the stepsize is not large enough to escape the basin, yet not small enough to ensure accurate convergence within it. Thus, the late-stage behavior of Muon is highly sensitive to learning-rate tuning. By comparison, pure GD progresses much more slowly in the early stage. It requires more than 1000 iterations to reach a region comparable to where Muon arrives within the first 100 iterations. Nevertheless, once GD approaches the relevant basin, it continues to refine the coefficients and eventually converges more accurately toward the ground-truth solution. This behavior is consistent with the diagonal case: GD is slow but stable, whereas Muon is fast but can suffer from persistent oscillations under a fixed stepsize.

The hybrid method again combines the advantages of both optimizers. During the initial Muon phase, it rapidly approaches the relevant basin. After switching to GD at iteration 100, it avoids the oscillation of pure Muon and instead performs stable late-stage refinement. This PSD example thus reinforces the main message of the paper: Muon is particularly effective as an early-stage exploration method, while GD-type dynamics are better suited for accurate late-stage convergence.

C.1.3 Two-Layer Neural Network Case

We further evaluate the behavior of different optimizers in a standard supervised learning setting: training a two-layer neural network on the MNIST handwritten digit classification task Deng [2012]. This experiment serves as an empirical extension of the mixed-spectrum matrix sensing examples to a more conventional machine learning problem. There is also a close connection between Burer-Monteiro (BM)-factorized matrix sensing and the training of a two-layer neural network Li et al. [2018]. In BM-factorized matrix sensing, the sensing matrices $\{A_i\}_{i=1}^m$ can be interpreted as data samples, while the PSD ground-truth matrix

$$M^\star = X^\star X^{\star\top}, \qquad X^\star \in \mathbb{R}^{n \times r_\star},$$

plays a role analogous to the optimal weight representation of the model. From this perspective, the present case study can be viewed as extending the mixed-spectrum matrix sensing construction back to a general BM-factorized MS or shallow neural network setting. This extension differs from the theoretical model in two important ways. First, the sensing matrices are no longer required to follow the special mixed-spectrum construction in (2); instead, they are replaced by general data-dependent matrices induced by the MNIST samples. Second, the objective is no longer the population loss in (Mixed-spectrum MS); rather, we optimize the empirical mini-batch training loss.

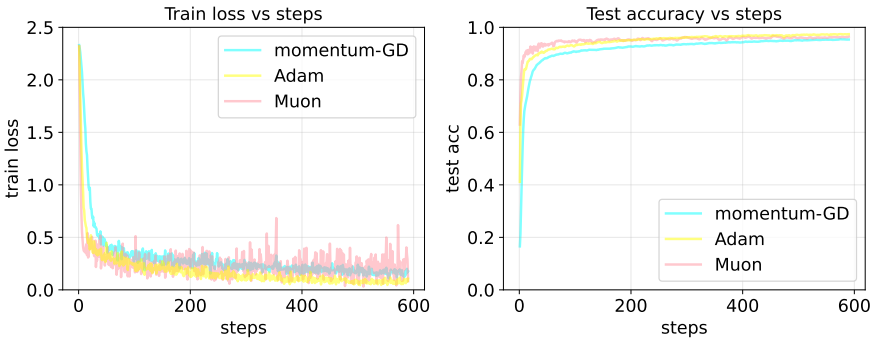

Figure 9: **Training a two-layer neural network on MNIST with pure optimizers.** Left: training loss of momentum-GD, Adam, and Muon. Right: test accuracy of the three optimizers.

Concretely, we train a two-layer neural network with input dimension $28 \times 28 = 784$, hidden width 256, and output dimension 10, corresponding to the ten MNIST digit classes. We use the ReLU activation function Nair and Hinton [2010] and train with mini-batches of size 512. Unlike the matrix sensing examples, which use full-batch optimization, the present experiment adopts stochastic mini-batch training in order to better reflect common practice in modern machine learning. Figure 9 reports the results of training for 5 epochs, corresponding to 590 iterations in total. For momentum-GD, we use learning rate 0.01 and momentum coefficient 0.9. For Muon, we use learning rate 0.1 and momentum coefficient 0.9; this larger learning rate is consistent with the intuition from our theoretical analysis that Muon can tolerate more aggressive early-stage steps (Theorem B.10). For Adam, we use learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We found that using a substantially larger learning rate for Adam degrades its performance.

The left panel of Figure 9 shows that Muon decreases the training loss fastest during the early stage, followed by Adam and then momentum-GD. However, after roughly 400 iterations, the Muon loss continues to oscillate, whereas the losses of Adam and momentum-GD keep decreasing more steadily. The right panel shows the corresponding test accuracy. Muon rapidly reaches above 90% test accuracy, but is later surpassed by Adam and approached by momentum-GD. This behavior is consistent with the phenomenon observed in the matrix sensing case studies: Muon is highly effective for fast early progress, but its fixed-stepsize dynamics may lead to inferior late-stage refinement.

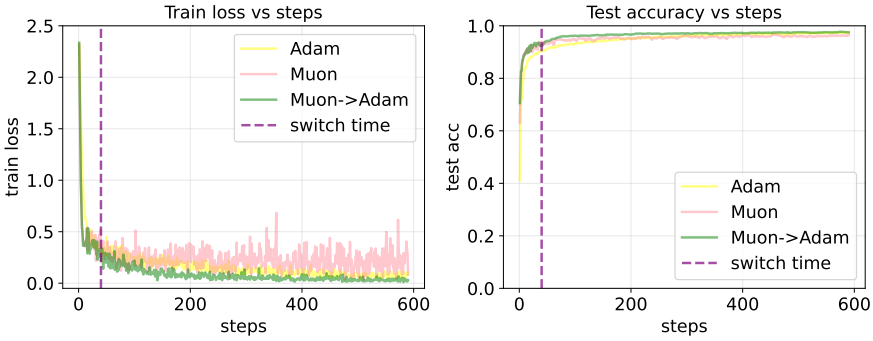

Figure 10: **Training a two-layer neural network on MNIST with a hybrid optimizer.** The optimizer first follows Muon and then switches to Adam. Left: training loss. Right: test accuracy.

Motivated by this observation, we further test a hybrid optimization strategy. In Figure 10, the optimizer starts with Muon and switches to Adam at iteration 40. During the Muon phase, we maintain

the moment estimates required by Adam, although they are not used to update the parameters before the switch. After switching, Adam inherits these accumulated first- and second-order moment estimates. The left panel of Figure 10 shows that, after the switch, the hybrid method avoids the persistent oscillations observed in pure Muon. Moreover, its training loss decreases below that of pure Adam. The right panel shows that the hybrid method also achieves the highest final test accuracy among the compared methods. We additionally tested a "Muon $\rightarrow$ momentum-GD" hybrid strategy, which exhibits qualitatively similar behavior to the "Muon $\rightarrow$ Adam" variant.

Overall, this case study provides empirical evidence, beyond the stylized theoretical MS model, for the main message of the paper: Muon is particularly effective as an early-stage exploration optimizer, while GD-type methods like Adam, are better suited for accurate late-stage convergence.

C.2 Empirical Observations on LLM Pre-training

C.2.1 Implementation Details for LLM Pre-training

We provide additional implementation details for the "Muon$\rightarrow$AdamW" pre-training experiments. For computational efficiency, these diagnostic experiments are conducted on a fixed subset of OpenWebText2 Gao et al. [2020], rather than on the full corpus. The subset contains approximately 250M tokens and is preprocessed into binary training and validation files. All optimizer variants use the same train–validation split. Mini-batches are sampled using a fixed data seed and without replacement. Together, these design choices allow us to compare optimizer and scheduler variants under a controlled data protocol.

We conduct two sets of experiments with different training horizons. In the 15K-step experiments, we compare a fixed-LR Muon baseline with Muon$\rightarrow$AdamW runs that switch at 10K, 11K, 12K, 13K, or 14K steps. In the 20K-step experiments, we compare Muon baselines with several LR schedules, including fixed, linear, cosine, and `cos_inf` schedules, against Muon$\rightarrow$AdamW runs with late-stage switches. For two-stage Muon$\rightarrow$AdamW runs, we train with Muon up to a prescribed switching step, resume from the corresponding checkpoint, and continue with AdamW for the remaining steps. At the switch, AdamW's second moment is initialized using the elementwise maximum of the calibrated squared gradients and the scaled squared Muon momentum. Its first moment is initialized by blending the Muon update direction with the calibrated gradient direction. This warm-start procedure is formalized in Section C.2.3. When a LR scheduler is used after the switch, it is applied over the remaining training horizon. All runs are trained on $4\times$ NVIDIA RTX PRO 6000 GPUs with 96GB of memory. Each process uses a batch size of 16. We evaluate every 100 steps using 32 validation batches and log training and validation metrics at the same interval. Unless otherwise stated, we use seed 0, data seed 1337, weight decay 0.1, and a gradient-clipping threshold of 1.0.

C.2.2 Effects of Switching Time and Learning-Rate Scheduling

The main LLM pre-training results in Section 6 focus primarily on the validation loss of the two-stage Muon$\rightarrow$AdamW procedure. Here, we provide additional results that report both training and validation losses, offering a more complete view of the optimization dynamics.

Figure 11 studies the effect of switching time in a 15K-step training horizon without learning-rate decay. In this setting, the Muon baseline is trained with a fixed LR of 5×10^{-3}, while the AdamW phase in the two-stage runs uses a fixed LR of 5×10^{-5}. We switch from Muon to AdamW at 10K, 11K, 12K, 13K, or 14K steps. The results reveal two main observations. First, Panel (a) shows late-stage switching from Muon to AdamW consistently outperforms the fixed-step Muon baseline across all tested switching times. This indicates that the advantage of the Muon$\rightarrow$AdamW transition is robust to the precise choice of switching point. Second, within the tested range, Panel (b) shows delaying the switch generally leads to lower validation loss. This suggests that it is beneficial to exploit Muon for the earlier exploration phase and use AdamW mainly for late-stage refinement.

Figure 12 further examines this phenomenon under LR decay over a 20K-step horizon. We compare Muon-only baselines trained with several LR schedules, including linear, cosine, and `cos_inf`, against two-stage Muon$\rightarrow$AdamW runs that switch at different training steps. For the hybrid runs, the AdamW phase uses a peak LR of 9×10^{-5} and is evaluated with the same set of schedules. For each switching time, we report the best-performing Muon$\rightarrow$AdamW configuration, which uses the `cos_inf` schedule during the AdamW phase. Compared with the scheduled Muon-only baselines,

the differences in training loss are relatively modest, excluding the fixed-LR Muon baseline, which performs substantially worse. However, the validation-loss curves show clearer benefits from the Muon→AdamW transition. In particular, switching at 17K or 18K steps yields markedly lower final validation losses than switching at 15K or 16K steps. The 19K-switch setting also decreases rapidly after the transition and appears promising. Nevertheless, because only a small number of steps remain after the switch, its validation loss does not fall below the best-performing earlier-switch configurations by the end of training. This suggests that, under a fixed training budget, switching too late may leave insufficient time for AdamW to fully realize its benefits.

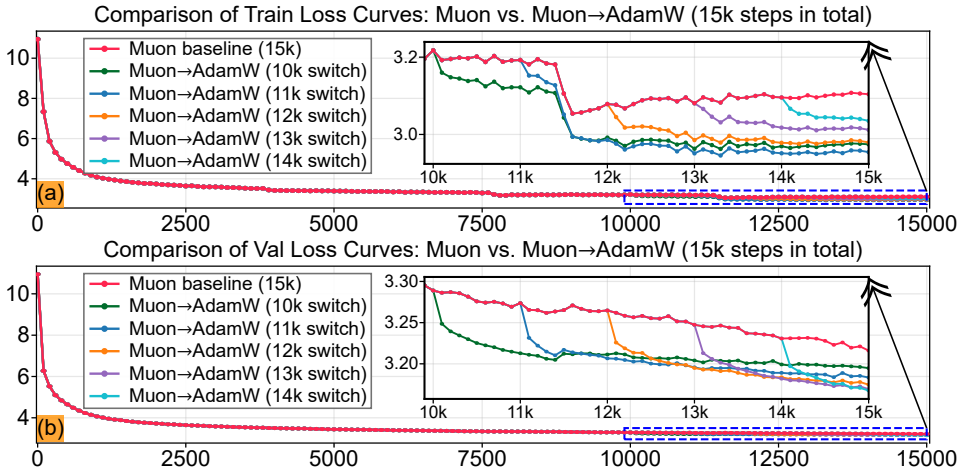

Figure 11: **Muon→AdamW switching times over a 15K-step pre-training horizon.** We compare a fixed-LR Muon baseline with two-stage Muon→AdamW runs that switch from Muon to AdamW at 10K, 11K, 12K, 13K, or 14K steps. Panel (a) shows the training loss, and Panel (b) shows the validation loss. The insets zoom in on the late-stage region where the optimizer switch occurs.

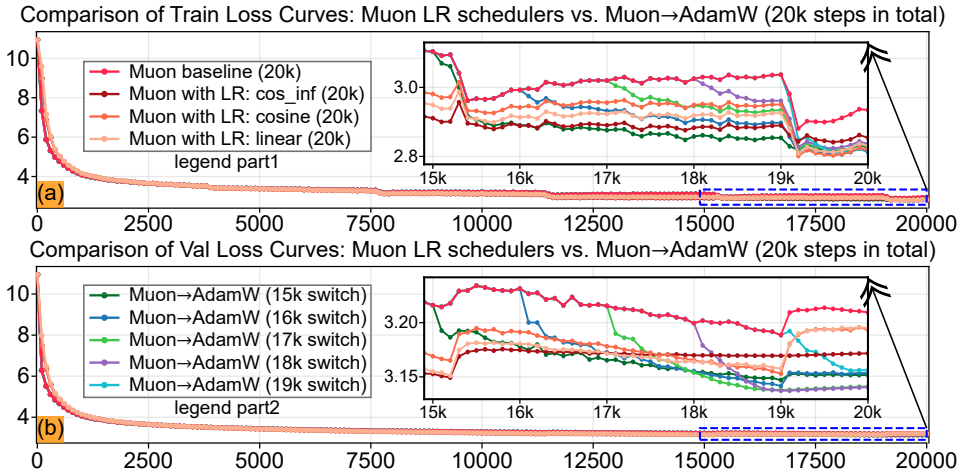

Figure 12: **Muon LR schedules versus Muon→AdamW switching over a 20K-step pre-training horizon.** We compare Muon baselines with fixed, linear, cosine, and `cos_inf` LR schedules against two-stage Muon→AdamW runs that switch from Muon to AdamW at 15K, 16K, 17K, 18K, or 19K steps. Panel (a) shows the training loss, and Panel (b) shows the validation loss. The insets zoom in on the final 5K steps, where the effects of LR scheduling and optimizer switching are most visible.

C.2.3 How to Switch: Warm-Starting AdamW from a Muon Checkpoint

At a selected switching iteration k, we restore the model parameters θ_k and the Muon optimizer state from the corresponding checkpoint, then instantiate a new AdamW optimizer at the same model parameters. Because Muon does not maintain AdamW's element-wise second moment, its state cannot

be converted exactly into an AdamW state. We therefore construct a warm start that combines two complementary sources of information: gradients evaluated at the restored model and the momentum direction accumulated by Muon. We use subscripts i, k, and ℓ to index a parameter tensor, the switching iteration, and a calibration mini-batch, respectively. Let $p_{i,k}$ denote the Muon momentum buffer for parameter tensor i stored at iteration k.

We first estimate local gradient statistics without updating the restored model. Using B calibration mini-batches, let $g_{i,k,\ell}$ be the gradient of parameter tensor i on calibration mini-batch ℓ. We compute

$$\bar{g}_{i,k}^{\mathrm{cal}} := \frac{1}{B}\sum_{\ell=1}^{B} g_{i,k,\ell}, \qquad q_{i,k}^{\mathrm{cal}} := \frac{1}{B}\sum_{\ell=1}^{B} g_{i,k,\ell} \odot g_{i,k,\ell},$$

where $\odot$ denotes element-wise multiplication. Thus, $q_{i,k}^{\mathrm{cal}}$ is the mean of the element-wise squared gradients, not the element-wise square of $\bar{g}_{i,k}^{\mathrm{cal}}$.

We next initialize AdamW's second-moment state. The calibration statistic captures the gradient scale at the switching point, whereas the squared Muon buffer retains information accumulated before the switch. We combine them as

$$\tilde{v}_{i,k} := \max\left\{c_g q_{i,k}^{\mathrm{cal}}, c_\mu (s p_{i,k})^2\right\}, \qquad \tau_{i,k} := \delta \max\left\{10^{-30}, \mathrm{mean}(\tilde{v}_{i,k})\right\},$$

where both the square and the maximum are element-wise, except that $\mathrm{mean}(\tilde{v}_{i,k})$ is a scalar. Here, s rescales the checkpoint momentum, c_g and c_μ weight the two sources, and δ sets a relative numerical floor. The initial AdamW second moment is

$$v_{i,k}^{\mathrm{init}} := \max\{\tilde{v}_{i,k}, \tau_{i,k}\},$$

with the scalar floor broadcast over the tensor.

To initialize the first moment, we blend the Muon direction with the local gradient direction in AdamW's preconditioned coordinates. Let $Q_{i,k}$ be the five-step Newton–Schulz orthogonalization of $sp_{i,k}$. A tensor with more than two dimensions is first reshaped to $d_0 \times \prod_{j>0} d_j$ and reshaped back after orthogonalization, whereas a one-dimensional tensor passes through without orthogonalization. For a matrix with dimensions $r_i \times c_i$, we additionally multiply the orthogonalized output by $\max\{\sqrt{r_i/c_i}, \sqrt{c_i/r_i}\}$, following the scaling used by the Muon implementation. Define

$$u_{i,k}^{\mathrm{grad}} := \frac{\bar{g}_{i,k}^{\mathrm{cal}}}{\sqrt{v_{i,k}^{\mathrm{init}}}}, \qquad \tilde{u}_{i,k}^{\mathrm{Muon}} := Q_{i,k}\frac{\mathrm{RMS}(u_{i,k}^{\mathrm{grad}})}{\mathrm{RMS}(Q_{i,k})},$$

where $\mathrm{RMS}(x) := \sqrt{\mathrm{mean}(x^2)}$. The first expression is the calibrated gradient direction under the initialized AdamW preconditioner; the second rescales the Muon direction to the same RMS. We then form

$$u_{i,k}^{\mathrm{init}} := \alpha \tilde{u}_{i,k}^{\mathrm{Muon}} + (1-\alpha) u_{i,k}^{\mathrm{grad}}, \qquad m_{i,k}^{\mathrm{init}} := u_{i,k}^{\mathrm{init}} \odot \sqrt{v_{i,k}^{\mathrm{init}}}.$$

Consequently, $m_{i,k}^{\mathrm{init}} / \sqrt{v_{i,k}^{\mathrm{init}}} = u_{i,k}^{\mathrm{init}}$ before AdamW applies its subsequent moment update and bias correction. If calibration gradients are unavailable for a parameter, the implementation falls back to the orthogonalized Muon direction without the gradient-based blend.

Finally, we initialize AdamW's per-parameter step counter to the checkpoint iteration k and continue with the configured (β_1, β_2) and LR schedule. This construction carries both the pre-switch Muon direction and local gradient information into the handoff, but it should not be interpreted as reconstructing the moments that AdamW would have accumulated had it been used from the beginning of training. In the 20K-step experiments, we use $B = 1024$, $s = 1$, $c_g = 0.3$, $c_\mu = 0.01$, $\delta = 10^{-5}$, and $\alpha = 0.8$. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, a peak LR of 9×10^{-5}, and zero weight decay. After computing the calibration statistics at θ_k, we reset the training-data reader to the switching point before resuming optimization.

C.2.4 When to Switch: A Possible Local Criterion

The experiments above indicate that an exact switching iteration is not essential within the late-stage window considered here. Across a broad range of late switching points, the Muon→AdamW

runs improve upon the corresponding Muon-only baselines. These observations suggest that the Muon→AdamW strategy may be useful beyond language-model pre-training, particularly in large-scale vision, video, and embodied-AI settings where training commonly involves long horizons and staged optimization pipelines [Yuan et al., 2025, 2026a,b]. The main practical restriction is that a switch made too close to the end of the training horizon leaves too few iterations for AdamW to complete the refinement phase. Thus, for the present experiments, selecting a reasonable late-stage switching window is more important than identifying a single precise iteration.

Designing an adaptive switching rule remains an interesting direction for future work. We give a brief local calculation that suggests one possible criterion, while leaving stochastic, nonconvex, and fully general variants to future study. Consider a twice-differentiable objective $\mathcal{F}$ that is smooth and locally strongly convex around the current matrix parameter X. Let

$$G := \nabla \mathcal{F}(X), \qquad \mathcal{H} := \nabla^2 \mathcal{F}(X),$$

and assume that, for every matrix perturbation Y in this neighborhood,

$$\lambda_{\min}\|Y\|_F^2 \le \langle Y, \mathcal{H}[Y]\rangle \le \lambda_{\max}\|Y\|_F^2, \qquad 0 < \lambda_{\min} \le \lambda_{\max}.$$

For a candidate descent direction $-D$, the local quadratic model gives

$$\mathcal{F}(X - \eta D) - \mathcal{F}(X) \approx -\eta\langle G, D\rangle + \frac{\eta^2}{2}\langle D, \mathcal{H}[D]\rangle.$$

Provided $\langle G, D\rangle > 0$, the minimizing coefficient and the resulting predicted change are

$$\eta_D^* = \frac{\langle G, D\rangle}{\langle D, \mathcal{H}[D]\rangle}, \qquad \Delta_{\text{quad}}(D) = -\frac{\langle G, D\rangle^2}{2\langle D, \mathcal{H}[D]\rangle}.$$

For GD, take $D_{\text{GD}} := G$. For an idealized Muon step, take $D_{\text{Muon}} := \texttt{msign}(G)$, and define $r_{\text{pol}} := \|D_{\text{Muon}}\|_F^2$. Then $\langle G, D_{\text{Muon}}\rangle = \|G\|_*$, where $\|\cdot\|_*$ denotes the nuclear norm. Under the compact-polar convention, $r_{\text{pol}} = \text{rank}(G)$; the same identity holds for the full-rank rectangular gradients typically encountered by Muon. The corresponding predicted changes are

$$\Delta_{\text{GD}} = -\frac{\|G\|_F^4}{2\langle G, \mathcal{H}[G]\rangle}, \qquad \Delta_{\text{Muon}} = -\frac{\|G\|_*^2}{2\langle D_{\text{Muon}}, \mathcal{H}[D_{\text{Muon}}]\rangle}.$$

The curvature bounds imply

$$\Delta_{\text{GD}} \le -\frac{\|G\|_F^2}{2\lambda_{\max}}, \qquad \Delta_{\text{Muon}} \ge -\frac{\|G\|_*^2}{2\lambda_{\min} r_{\text{pol}}}.$$

Therefore, a sufficient condition for the locally optimized GD step to provide at least as much predicted decrease as the locally optimized Muon step is

$$\|G\|_*^2 \le \frac{r_{\text{pol}}}{\kappa(\mathcal{H})}\|G\|_F^2, \qquad \kappa(\mathcal{H}) := \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Equivalently, defining the spectral effective rank $r_{\text{eff}}(G) := \|G\|_*^2/\|G\|_F^2$, the condition is $\kappa(\mathcal{H}) r_{\text{eff}}(G) \le r_{\text{pol}}$. This calculation suggests that a future switching rule could monitor two quantities: local conditioning and the dispersion of the gradient's singular values. However, it is only a local deterministic comparison between idealized GD and Muon directions; it does not yet account for AdamW's adaptive preconditioner, momentum, stochastic gradients, nonconvex curvature, higher-order terms, or variation across parameter blocks. Constructing reliable estimators for these quantities and turning this sufficient condition into a practical Muon→AdamW switching criterion are left for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The main claims stated in the abstract and introduction accurately reflect the paper's actual contributions and scope, and are supported by the theoretical and empirical results presented in the paper.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses limitations by explicitly stating the assumptions and scope for all theoretical results within the theorem statements, clarifying when conclusions may not hold if these assumptions are violated. In addition, any informal results in the main text are backed by corresponding formal statements and proofs (in the paper and appendix), ensuring that limitations and applicability are rigorously specified.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, we state the full set of assumptions alongside the theorem statements and provide complete proofs. Where space is limited, we include proof sketches in the main text and place full formal proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the information needed to reproduce the main experimental results that support the core claims, including the experimental protocol, model/training details, and evaluation procedure.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to our code via an anonymized repository for the review period. All datasets used are publicly available (LLM pretraining and finetuning corpora).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the key training and evaluation details needed to interpret and reproduce the results, including the data splits, model/training hyperparameters, and optimization settings. Additional implementation details and full configurations are provided in the appendix and the accompanying code release.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests because running multiple independent LLM pretraining/finetuning repeats (or extensive bootstrapping) would be prohibitively expensive in terms of compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute used to run the experiments, including the hardware configuration (4× NVIDIA RTX 5880 Ada Generation and 4× NVIDIA RTX PRO 6000 Blackwell GPUs) and the resources required to reproduce the runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics and confirm that the research complies with its requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is primarily theoretical and does not introduce a new deployable system, dataset, or application-specific pipeline, so we do not identify concrete, direct societal impacts tied to real-world deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The theoretical paper does not release high-risk artifacts. Hence, responsible-release safeguards (e.g., gated access, usage restrictions, or content filtering) are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly credit the creators of all third-party assets used (datasets, codebases, and pretrained models where applicable) and cite the original sources. We also

explicitly document the relevant licenses/terms of use (including asset versions and access links where available) and use the assets in accordance with those stated licenses and restrictions.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new assets (e.g., a new dataset, model weights, or a new benchmark). As the contributions are theoretical and we do not distribute new artifacts beyond standard implementation used for experiments, asset documentation alongside a release is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing, user studies, or any research with human subjects (i.e., no direct interaction with participants and no human annotation/compensated data collection conducted by the authors). Therefore, participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subjects research (no direct interaction with participants and no human data collection/annotation conducted by the authors). As a result, participant risk disclosure and IRB (or equivalent) approval are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing and for assisting with routine implementation of experimental code, and did not contribute to the core methodology, theoretical development, or the originality of the scientific results. Therefore, an LLM-usage declaration is not required under the checklist guidance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.