Denoise First, Orthogonalize Later: Understanding Momentum in Muon via Spectral Filtering

Xianliang Li[1,2,*] **Zihan Zhang**[3,2,*] **Weiyang Liu**[4] **Han Bao**[1,2,5,6,†]

[1]The Institute of Statistical Mathematics [2]The Graduate Institute for Advanced Studies, SOKENDAI
[3]National Institute of Informatics [4]The Chinese University of Hong Kong
[5]Tohoku University [6]RIKEN AIP

Abstract

Muon has recently demonstrated strong empirical performance in large language model training, but the theoretical role of momentum in Muon remains unclear. Existing analyses of Muon either remove momentum to study spectral updates in isolation, or retain momentum without explaining why it improves empirical performance. Our work bridges this gap by showing momentum in Muon acts as a spectral filter. Under a structured signal-plus-perturbation gradient model, we prove that momentum suppresses perturbations while preserving the dominant signal, thereby enlarging the spectral gap between them. This enlarged gap stabilizes the singular subspaces of the matrix passed to Muon's orthogonalization step, making the resulting update more reliable. We further show that applying momentum before orthogonalization achieves *provably stronger* alignment with the signal component of the gradient than either reversing this order or simply removing momentum. Experiments across diverse tasks, including LLM pretraining, support our theoretical analysis. More broadly, our theory offers a starting point for understanding the benefits of momentum in other matrix-based optimizers.

1 Introduction

Modern neural networks, especially large language models (LLMs), typically consist of billions of parameters and require substantial computational resources for training. As the model scale continues to grow, optimizer design has become increasingly important for efficient large-scale learning. In LLM training, adaptive first-order, coordinate-wise methods such as Adam [27] and AdamW [33] remain the de facto choice. More recently, matrix-based optimizers such as Shampoo [20], SOAP [57], and Scion [38] have received growing interest as alternatives that exploit matrix structure in gradient updates. Among these optimizers, Muon [25], which approximately orthogonalizes the momentum matrix using a Newton–Schulz iteration, has shown promising empirical performance in LLM pretraining [31], and recent optimizer benchmarks further identify Muon, along with matrix-based methods more broadly, as a strong alternative to AdamW [45, 61].

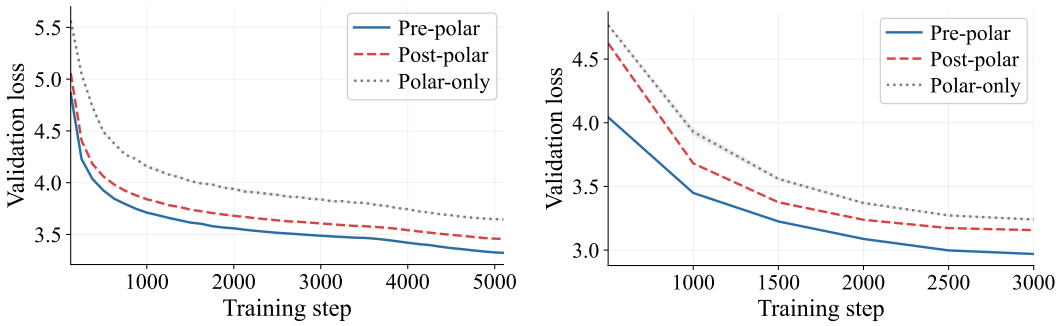
A second long-standing thread in optimizer design is the role of momentum. Classical analyses establish accelerated rates for momentum on smooth and strongly convex problems [39, 36, 52]. Cutkosky and Mehta [11] prove that momentum eliminates the need for large batches in normalized SGD, which is the vector-side analog of Muon's polar update in the sense that both retain only the direction of the gradient. The practical picture in deep learning is, however, more nuanced. Wang et al. [58] show that at small learning rates and high gradient noise, momentum offers only marginal benefit, suggesting that its provable acceleration sits in the deterministic (or large-learning-rate) regime rather than in stochastic noise reduction per se. More recent work [30] reframes momentum

*Equal contribution. †Corresponding author. Project page: yinleung.com/denoise-ortho

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

in the frequency domain as a low-pass filter on the gradient stream, suggesting that the relevant mechanism may be filtering rather than acceleration. Despite these developments, neither Muon's strong empirical performance nor the role momentum plays inside it is well understood.

In practice, Muon's pretraining performance is closely tied to momentum — see figure 1, where we see a clear trend that Polar-only pipeline (i.e., Muon without momentum) underperforms in pretraining — yet existing theory has not successfully characterized the role of momentum inside Muon. Substantial recent work analyzes Muon-style polar updates from the aspects of its spectral property, implicit bias, and feature learning, but *without momentum* [34, 51, 12, 29, 59, 26]. By construction, these analyses cannot explain momentum's benefit. A thorough review of these works is deferred to Appendix A. More recently, Chen et al. [9] establish that Muon converges to the spectrally-constrained minimizer, though its convergence rate deteriorates with heavier momentum. Kovalev [28] shows convergence of Muon to a stationary point but heavier momentum brings no further improvement. Shulgin et al. [47] analyze Muon under an inexact linear minimization oracle (LMO) that models the Newton–Schulz approximation error, yet its theoretical bound deteriorates in the heavy-momentum regime. Fan et al. [15] establish implicit bias of Muon with momentum, but still its convergence rate blows up with too heavy momentum. On the whole, theory tells us that heavy momentum brings no benefit — and may even degrade convergence — inside Muon, but this obviously contradicts what we observe in practice.



(a) NanoGPT. (b) LLaMA 350M.

Figure 1: End-to-end validation loss comparisons across (a) NanoGPT training and (b) LLaMA 350M training. The Muon Pre-polar pipeline outperforms Post-polar and Polar-only pipelines. The full experimental settings are in Appendix F.4.

The relationship between Muon's polar update and momentum is more nuanced. Whereas Orthogonal-SGDM [54], proposed prior to Muon, orthogonalizes each per-step gradient *before* momentum smoothing, Muon applies the polar update *after* momentum smoothing, significantly outperforming particularly on language model pretraining [25]. Therefore, Muon's success encourages us to closely look at (i) why momentum smoothing matters practically, and (ii) why the polar factor should be put after momentum smoothing. To this end, we compare three Muon pipelines for the weight update ΔW_t at iteration t, each combining a *momentum buffer* $M_t = \beta M_{t-1} + (1-\beta)G_t$ and the polar factor $\mathcal{O}(X) = UV^\top$ (from the thin SVD $X = U\Sigma V^\top$) on the gradient stream $\{G_t\}$ in different orders:

$$\text{Pre-polar:} \quad \Delta W_t \propto \mathcal{O}(M_t), \tag{1}$$

$$\text{Post-polar:} \quad \widetilde{M}_t = \beta \widetilde{M}_{t-1} + (1-\beta)\,\mathcal{O}(G_t), \quad \Delta W_t \propto \widetilde{M}_t, \quad \text{and} \tag{2}$$

$$\text{Polar-only:} \quad \Delta W_t \propto \mathcal{O}(G_t). \tag{3}$$

Pre-polar update equation (1) is the standard Muon update [25]. Post-polar update equation (2) formalizes the Orthogonal-SGDM variant. Polar-only update equation (3) removes momentum entirely as a no-momentum baseline. Figure 1 shows that Pre-polar dominates both Post-polar and Polar-only throughout training. Two conclusions follow. First, Pre-polar and Post-polar pipelines use the same two operations in opposite orders, so any gap between them must come from the ordering itself, because momentum and orthogonalization do not commute as numerical operations. Second, the gap from Pre-polar to Polar-only shows that momentum plays a role that orthogonalization alone cannot fulfill: momentum is an essential component of Muon, not a cosmetic acceleration trick. This raises the central question of our interest:

What role does momentum play on the gradient stream, and why does placing it before orthogonalization matter inside Muon?

This paper supplies the missing puzzle of Muon theory. Momentum first acts as a spectral filter on the gradient stream. Inside Muon, this filtering must happen before orthogonalization. The polar factor preserves singular directions but replaces all singular values with ones, so when it acts directly on a noisy gradient it erases the amplitude gap between signal and noise directions, and we shall never distinguish signal and noise even after undergoing the subsequent momentum. When momentum acts first, it filters the gradient stream to preserve the persistent signal component and attenuate the perturbation, so the polar factor operates on a matrix whose singular-value structure already separates signal from noise. The guiding principle is simple: *denoise first, then orthogonalize*.

Contributions.

1. **Spectral gap and singular-subspace reliability (Theorem 1 and Corollary 1).** Under a structured signal-plus-perturbation gradient model with a coherent signal, a bounded variance mean-zero orthogonal (BVMZOS) perturbation, and the signal persistence assumption, momentum solely opens a spectral gap between the r-th and $(r+1)$-th singular values that grows with the momentum window size T, and aligns the top-r singular subspaces of M_t with those of the coherent signal. This benefits Pre-polar because the subsequent orthogonalization operates on a reliable momentum matrix rather than on the raw gradient matrix.
2. **Pre-polar dominance and quantitative separation (Theorems 2 and 3).** For any fixed gradient signal matrix and any absolutely continuous perturbation distribution, Pre-polar pipeline provably dominates both Post-polar and Polar-only pipelines in expected signal alignment once the momentum window size T is sufficiently large. In the rank-1 spiked Gaussian model under low Signal-to-Noise Ratio (SNR) regime, this separation is quantitative: Polar-only alignment vanishes as the matrix size $m \to \infty$, while Pre-polar continues to recover the signal.
3. **Empirical validation across diverse tasks.** Synthetic spiked-perturbation models, CIFAR-10 and NanoGPT stationary and trajectory probes across layers and training checkpoints, and end-to-end NanoGPT and LLaMA 350M training support every prediction of Theorem 1, Corollary 1, and Theorem 2.

2 Preliminary and Analysis Setup

Notation. Let $G_t \in \mathbb{R}^{m\times n}$ $(m \geq n)$ denote the gradient matrix at iteration t. For the thin SVD decomposition $X = U\Sigma V^\top$ with orthogonal matrices $U \in \mathbb{R}^{m\times n}$ and $V \in \mathbb{R}^{n\times n}$, the polar factor is $\mathcal{O}(X) := UV^\top$. We write $\sigma_k(X)$ for the k-th largest singular value, $\hat{u}_k(X)$, $\hat{v}_k(X)$ for the corresponding singular vectors, $\langle A, B\rangle_F := \mathrm{tr}(A^\top B)$ for the Frobenius inner product, $\|A\|_F := \sqrt{\mathrm{tr}(A^\top A)} = \sqrt{\sum_{i,j} A_{ij}^2}$ for the Frobenius norm, and $\|\cdot\|_2$ for the operator norm.

Momentum buffer. For the momentum coefficient $\beta \in [0, 1)$, we define the *momentum buffer* via the recursion

$$M_t = \beta M_{t-1} + (1-\beta)G_t, \tag{4}$$

with the standard zero initialization $M_0 = 0$. The recursion unrolls to a convolution of the gradient stream as $M_t = \sum_{s=0}^{t-1} h_s\, G_{t-s}$ with the momentum kernel $h_s := (1-\beta)\beta^s$. The form of equation (4) is the exponential moving average (EMA) normalization [16] of the more general first-order recursion $M_t = \beta M_{t-1} + \gamma G_t$ with scalar $\gamma > 0$. Because the polar factor is scale-invariant, Pre-polar update $\mathcal{O}(M_t)$ depends only on the decay β, so the EMA buffer equation (4) and the Polyak/heavy-ball momentum ($\gamma = 1$) [52] yield the same update after being passed to the polar factor $\mathcal{O}(\cdot)$. Appendix B.2 records the derivation, and Appendix B.3 discusses arbitrary initialization. Nesterov momentum, common in Muon implementations, applies the polar factor not to the momentum buffer M_t in equation (4) but to a linear combination $\nu\, M_t + \kappa\, G_t$ of the M_t and the current gradient. This is not covered by our analysis. We record its kernel weights and effective sample size for completeness in Appendix B.4.

Effective window size. Based on the momentum buffer equation (4), the analysis uses the *effective window size*, $T \coloneqq \frac{1}{1-\beta}$, of the momentum kernel $(1-\beta)\beta^s$ [18]. As momentum becomes heavier, gradient smoothing is effective across longer horizon T.

Gradient model. Our results study the momentum buffer equation (4) applied to a gradient stream with the following *signal-plus-perturbation* structure.

Assumption 1 (Structured gradient decomposition)**.** *For each gradient-update step t,*

$$G_t = G_t^{\mathrm{sig}} + \Xi_t, \tag{5}$$

where the following two conditions are satisfied:

(a) ***(Coherent signal.)*** $G_t^{\mathrm{sig}} = \sum_{k=1}^{r} \lambda_k(t) u_k v_k^\top$*, where* $\{u_k\}_{k=1}^r \subset \mathbb{R}^m$ *and* $\{v_k\}_{k=1}^r \subset \mathbb{R}^n$ *are* temporally coherent *(i.e., deterministic and time-invariant) orthonormal families, and the* signed coordinate $\lambda_k(t)$ *of* G_t^{sig} *along the k-th signal direction* $u_k v_k^\top$ *is a real-valued random variable on the same probability space as the perturbation matrix* Ξ_t *(introduced below), with finite second moment.*

(b) ***(Bounded variance mean-zero orthogonal perturbation)*** $\{\Xi_t\}_{t\geq 0}$ *is an $m \times n$ matrix-valued* bounded variance mean-zero orthogonal sequence (BVMZOS) *: there exists $\eta > 0$ such that for every $t \geq 0$,* $\mathbb{E}[\Xi_t] = 0$ *and* $\mathbb{E}\left[\|\Xi_t\|_F^2\right] \leq \eta$*. Moreover, for any $t_1, t_2 \geq 0$ with $t_1 \neq t_2$,* $\mathbb{E}[\langle \Xi_{t_1}, \Xi_{t_2} \rangle_F] = 0$.

Assumption 1(b) is sufficiently general to cover time-dependent noise and finite-variance heavy-tailed noise, going beyond the standard i.i.d. sub-Gaussian setting. The rank-r decomposition can in fact address more general cases by absorbing higher-rank yet marginal components into the perturbation Ξ_t at the cutoff parameter r. Empirical and theoretical evidence suggests that neural network gradients maintain effectively low rank [21, 2], and on real neural network gradients r can be regarded as an effective rank where the spectrum has its dominant spikes [35, 5].

Remark 1 (Matrix sensing example)**.** *While we introduce Assumption 1 as a theoretical proxy to practically observed gradient streams in large-scale training, it is naturally satisfied by a standard matrix-sensing example. In this setting, the signal component G_t^{sig} corresponds to the full-batch gradient lying in a time-invariant low-rank subspace, while the BVMZOS perturbation Ξ_t corresponds to the mini-batch sampling noise.*
Consider the regression model $y = \langle W^*, X \rangle_F + \epsilon$, where $W^* \in \mathbb{R}^{m\times n}$ is a fixed but unknown matrix, $X \in \mathrm{span}(\{u_k v_k^\top\}_{k=1}^r)$ is the sensing matrix assumed to be low-rank, and ϵ is mean-zero noise. Write $X_i = \sum_{k=1}^r \mu_{i,k} u_k v_k^\top$ with some $\mu_{i,k} \in \mathbb{R}$. At the parameter $W = W_t$, the least-squares loss $\mathcal{L}(W) \coloneqq \frac{1}{2N}\sum_{i=1}^N (y_i - \langle W, X_i \rangle_F)^2$ has gradient $\nabla\mathcal{L}(W_t) = \sum_{k=1}^r \lambda_k(t) u_k v_k^\top$, where the signed coordinate is $\lambda_k(t) \coloneqq \frac{1}{N}\sum_{i=1}^N \mu_{i,k}(\langle W_t, X_i \rangle_F - y_i)$. Thus the full-batch gradient has exactly the coherent-signal structure. If the full-batch gradient is replaced with an unbiased mini-batch estimator, the mini-batch sampling noise satisfies Assumption 1(b), provided that the sampling is independently conditional on the past.

Signal persistence. In addition to the temporal coherency in Assumption 1, the signed coordinates $\lambda_k(t)$ are assumed to weakly drift in time. We measure this drift relative to the *root-mean-squared (RMS) coefficient scale* $\lambda_k \coloneqq \sqrt{\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \mathbb{E}[\lambda_k(t)^2]}$, where $\mathcal{I}$ is a fixed deterministic interval where all our subsequent discussion takes place, and the expectation is taken over the randomness of the perturbation matrix $\{\Xi_t\}$. This is a finite, deterministic scalar by Assumption 1(a). Without loss of generality, we index the coherent signal directions so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$. We use the threshold $t \geq c_0 T$ throughout, where $c_0 > 0$ is a fixed constant chosen so that $\beta^{c_0 T}$ is negligible (Appendix B.3).

The quantitative statement is given as follows:

Assumption 2 (Signal persistence)**.** *Define the momentum-filtered signed coordinate* $\bar{\lambda}_k(t) \coloneqq (1-\beta)\sum_{\tau\geq 0} \beta^\tau \lambda_k(t-\tau)$*. There exists* $c_{\mathrm{sig}} \in (0, 1]$*, irrespective of β or $T = 1/(1-\beta)$, such that for all $t \geq c_0 T$ and $k = 1, \ldots, r$,*

$$|\bar{\lambda}_k(t)| \geq c_{\mathrm{sig}} \lambda_k.$$

This says the momentum-filtered coordinate $|\bar{\lambda}_k(t)|$ is sufficiently bounded away from zero, even if the unsmoothed $\lambda_k(t)$ could be close to zero. If the signed coordinate admits a specific form such as $\lambda_k(t) = \bar{\mu}_k + \xi_k(t)$ with $|\bar{\mu}_k| > 0$ and zero-mean sub-Gaussian $\xi_k(t)$, Assumption 2 holds with high probability for large enough window T (Appendix E.1).

3 Filtering Effect of Momentum: Spectral Gap and Subspace Reliability

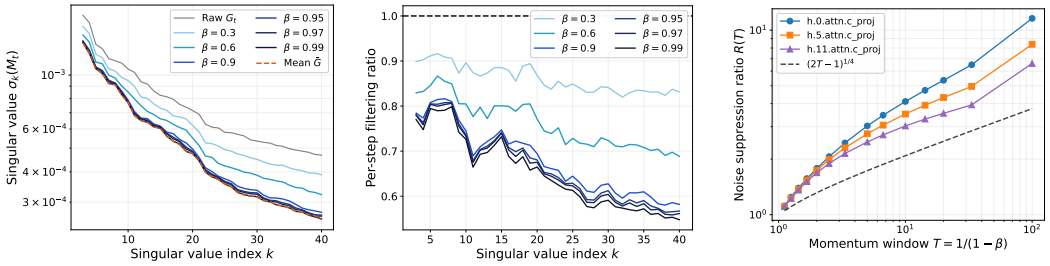
Before investigating the interaction between momentum and the polar factor, we present our first analysis focusing on the benefit of momentum in signal recovery, throughout this section. The full proofs of the statements in this section are deferred to Appendix C.

Amplitude recovery (in theory). As the effective window size $T = 1/(1-\beta)$ grows, the momentum buffer M_t averages out the perturbations while preserving the coherent signal G_t^{sig}. Theorem 1 below quantifies this: the tail singular values decay as $(2T-1)^{-1/4}$, creating a spectral gap with a larger effective window size T.

Theorem 1 (Spectral gap of the momentum buffer)**.** *Under Assumptions 1 and 2, for $m \geq n$, with probability of at least $1 - (2T-1)^{-\frac{1}{2}}$, we have*

$$\sigma_k(M_t) \geq c_{\text{sig}}\lambda_k - \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}}, \quad k = 1, \ldots, r, \quad \text{and} \quad \sigma_{r+1}(M_t) \leq \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}}. \tag{6}$$

Hence, the spectral gap widens with a larger effective window size T, or equivalently, with a heavier momentum β. The perturbation floor $O((2T-1)^{-1/4})$ in equation (6) monotonically decays. While this behavior is conceptually intuitive — momentum smoothening out noise perturbations — this spectral gap decay behavior aligns well with the practical benefit of momentum at near-1 β, in stark contrast to the standard optimization analysis whose theoretical guarantees deteriorate with large β [9, 28, 15, 47].



(a) Filtered singular value spectra. (b) Per-step filtering ratio. (c) Noise suppression ratio $R(T)$.

Figure 2: Spectral filtering visualization. (a) Filtered momentum singular value spectra (blue), the raw gradient spectrum (grey), and the mean-gradient spectrum (dashed orange) on layer `h.0`. (b) Per-step filtering ratio on `h.0`. (c) Noise-suppression ratio $R(T)$ on each layer `h.0`, `h.5`, and `h.11` ($K = 500$) versus momentum window size $T = 1/(1-\beta)$, with the dashed $(2T-1)^{1/4}$ floor.

Amplitude recovery (in simulation). We next use numerical simulations to verify that the momentum buffer indeed creates the spectral gap implied by Theorem 1, beyond the stylized gradient model. To do so, we use the *stationary probe* by collecting K mini-batch gradients on a saved checkpoint with all model weights held fixed, and applying the momentum pipelines to the gradient buffer in collection order. Holding the weights fixed removes the non-stationarity caused by parameter updates, so the probe isolates the stochastic gradient stream around a realistic training checkpoint. In this controlled setting, the coherent signal is a well-defined, time-invariant target, just as in the stylized model. At the same time, unlike the stylized model, the gradients are generated by a real learning task and therefore retain the structure of an actual neural-network training problem. Here, we use the NanoGPT stationary probe at training step 3000 (out of 5100 total steps) on the attention output projection layers. The full settings appear in Appendix F, with synthetic, CIFAR-10, and additional NanoGPT layer and checkpoint results in Appendix G.

Theorem 1 contains two predictions: signal preservation in the *head* ($k \le r$), $\sigma_k(M_t) \ge c_{\text{sig}}\lambda_k - \sqrt{\eta}\,(2T-1)^{-1/4}$, which keeps the leading singular values pinned near the signal scale $c_{\text{sig}}\lambda_k$, and perturbation attenuation in the *tail* ($k > r$), $\sigma_{r+1}(M_t) \le \sqrt{\eta}\,(2T-1)^{-1/4}$, which drives the tail to zero at rate $T^{-1/4}$. We validate these on real gradients via two empirical metrics. Let $M_K^{(\beta)}$ denote the momentum buffer equation (4) with decay β after collecting a buffer of K gradients, and G_K the last raw gradient in that buffer. The *per-step filtering ratio* $\mathrm{Filt}_k(\beta) \coloneqq \sigma_k(M_K^{(\beta)})/\sigma_k(G_K)$ is the index-wise attenuation between the raw gradient G_K and the buffer $M_K^{(\beta)}$. Theorem 1 predicts that signal preservation keeps $\mathrm{Filt}_k(\beta)$ near 1 at the head ($k \le r$), while perturbation is suppressed as β gets closer to 1.

The *noise-suppression ratio* $R(T) \coloneqq \|G_K - \bar{G}\|_2 / \|M_K^{(\beta)} - \bar{G}\|_2$, with the in-buffer mean gradient $\bar{G} \coloneqq \frac{1}{K}\sum_{t=1}^{K} G_t$ as the reference of the coherent gradient signal, is the ratio of the raw gradient's perturbation to the perturbation remaining in the momentum buffer. Under stationarity $\bar{G}$ approximates the coherent signal, so subtracting it approximately removes the signal from both numerator and denominator and isolates the perturbation. Theorem 1 predicts that this suppression grows with the momentum window size T, with $R(T)$ bounded below by the $(2T-1)^{1/4}$ floor. The full derivation of this argument is deferred to Appendix F.3.

Figures 2a and 2b support Theorem 1 via the stationary NanoGPT probe at training step 3000 on `h.0.attn.c_proj` layer. They show that as β increases, the tail singular values of $M_K^{(\beta)}$ and the tail filtering ratios are suppressed more than the head, opening a gap between head and tail in both panels. This is the spectral gap predicted by Theorem 1, and it widens with β. In figure 2a, the momentum spectrum moves toward the mean-gradient spectrum $\sigma_k(\bar{G})$ as β increases, with the head reaching $\sigma_k(\bar{G})$ ahead of the tail, consistent with Theorem 1. Figure 2c shows the noise-suppression ratio $R(T)$ on three attention output projections lies above $(2T-1)^{1/4}$ guide predicted by Theorem 1, serving as a sanity check of our theory.

Subspace recovery (in theory). Theorem 1 controls singular *values* corresponding to the coherent signal. Combined with Wedin's $\sin\Theta$ theorem [60], the spectral gap it opens also controls the singular *subspaces*: the larger the gap, the closer the top-r singular subspaces of M_t are to those of the coherent gradient signal.

Corollary 1 (Singular-subspace reliability)**.** *Under the conditions of Theorem 1, with probability of at least* $1 - (2T-1)^{-\frac{1}{2}}$*, we have*

$$\|\sin\Theta(\hat{U}_t, U)\|_2 \le \left(c_{\text{sig}}\lambda_r - \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}}\right)^{-1} \cdot \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}} = O(T^{-1/4}) \tag{7}$$

$$\|\sin\Theta(\hat{V}_t, V)\|_2 \le \left(c_{\text{sig}}\lambda_r - \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}}\right)^{-1} \cdot \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}} = O(T^{-1/4}). \tag{8}$$

where $U \coloneqq [u_1, \dots, u_r]$, $V \coloneqq [v_1, \dots, v_r]$, *and* $\hat{U}_t$, $\hat{V}_t$ *denote the top-*r *left and right singular subspaces of* M_t*, respectively. The principal angles* $0 \le \theta_1 \le \cdots \le \theta_r \le \pi/2$ *between the column spans of* $\hat{U}_t$ *and* U *are defined by* $\cos\theta_i \coloneqq \sigma_i(\hat{U}_t^\top U)$*. Stacking them into* $\Theta(\hat{U}_t, U) \coloneqq \mathrm{diag}(\theta_1, \dots, \theta_r) \in \mathbb{R}^{r\times r}$ *and applying* sin *entrywise gives* $\sin\Theta(\hat{U}_t, U) = \mathrm{diag}(\sin\theta_1, \dots, \sin\theta_r)$*, whose spectral norm* $\|\sin\Theta(\hat{U}_t, U)\|_2 = \sin\theta_r$ *is the sine of the largest principal angle.*

Each $\|\sin\Theta\|_2$ term lies in $[0, 1]$ and measures how far a top-r singular subspace of M_t is from the corresponding signal subspace: 0 when they coincide, 1 when orthogonal. Corollary 1 thus says the momentum buffer's leading singular subspaces become reliable as the momentum window size grows, with the deviation vanishing at rate $T^{-1/4}$.

Subspace recovery (in simulation). Figure 3 supports Corollary 1 at each rank $r \in \{1,5,10\}$ via the stationary NanoGPT probe at training step 3000 on `h.0.attn.c_proj` layer. The figure plots two empirical *subspace alignment errors* on real gradients: the left-subspace error $\sin\Theta_U \coloneqq \sin\theta_r(M_K^{(\beta)}; \bar{G})$ and the right-subspace error $\sin\Theta_V \coloneqq \sin\theta_r(M_K^{(\beta)\top}; \bar{G}^\top)$, where $\sin\theta_r(A;\, B) \coloneqq \|\sin\Theta(U_r(A),\, U_r(B))\|_2$. Both measure the principal-angle distance between the top-r singular subspaces of $M_K^{(\beta)}$ and those of $\bar{G}$ (full definition in Appendix F.3).

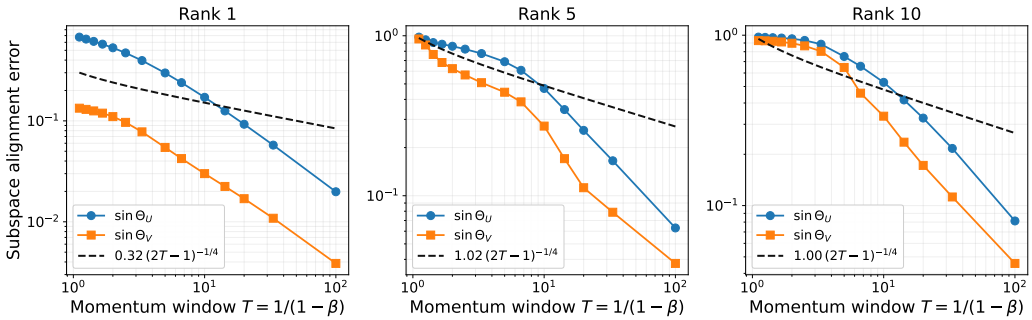


Figure 3: Stationary probe subspace alignment error $\sin\Theta_U$ and $\sin\Theta_V$ at ranks $r \in \{1, 5, 10\}$ versus momentum window size $T = 1/(1-\beta)$, with the dashed $c_r\,(2T-1)^{-1/4}$ guide (c_r fitted independently per panel).

4 Noncommutativity of Momentum and Orthogonalization

Signal-recovery separation (in theory). While Section 3 focuses on the denoising effect arising solely from momentum, this section investigates the interaction between momentum and the polar update across the variants: Pre-polar update $\mathcal{O}(M_t)$, Post-polar update $\widetilde{M}_t$, and Polar-only $\mathcal{O}(G_t)$ in equations (1) to (3). Under the same stylized gradient model introduced in Section 2, Theorem 2 shows that Pre-polar recovers the signal matrix G^{sig} correctly as momentum filters out the BVMZOS perturbation with a sufficiently large effective window. The expected scores of Polar-only and Post-polar remain separated from this optimum. Thus, our main theorem successfully demonstrates the separation between Pre- and Post-polar, explaining the practical superiority of Muon over Orthogonal-SGDM. Theorem 3 is a quantitative version of the separation result in Theorem 2, proving that the signal alignment asymptotically vanishes with the matrix size $m \to \infty$ if we do not use Pre-polar.

Before stating the main theorem in this section, we introduce the following stronger assumptions in addition to Assumption 1, which are used in the proof of Theorem 2.

Assumption 3 (Time-invariant signal component)**.** *In addition to Assumption 1(a), we assume that the signal component G_t^{sig} is time-invariant, i.e., $G_t^{\mathrm{sig}} = G^{\mathrm{sig}}$ for all t.*

Assumption 4 (Pairwise independence of perturbations)**.** *In addition to Assumption 1(b), we assume that the perturbations $\{\Xi_t\}$ are pairwise independent. As a consequence, for any measurable function f and g, and $t_1, t_2 \geq 0$ such that $t_1 \neq t_2$, $\mathbb{E}\left[f(\Xi_{t_1})g(\Xi_{t_2})\right] = \mathbb{E}\left[f(\Xi_{t_1})\right]\mathbb{E}\left[g(\Xi_{t_2})\right]$.*

Theorem 2 (Pre-polar recovery versus non-denoised baselines)**.** *Assuming Assumptions 1 to 3 and $m \geq n$, for every choice of fixed $(G^{\mathrm{sig}}, \{\Xi_t\})$ such that the law of Ξ_t is absolutely continuous with respect to the Lebesgue measure, there exists a constant $C > 0$, independent of the momentum window size T, such that the following statements hold all at once.*

(i) ***Polar-only baseline.***

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(G_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right] \leq 1 - C.$$

(ii) ***Post-polar pipeline.***

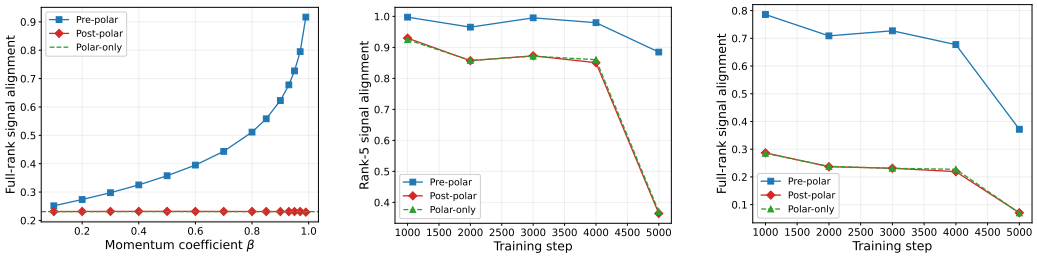
$$\mathbb{E}\left[\frac{\langle \tilde{M}_t, G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right] \leq 1 - C.$$

If we moreover assume Assumption 4, then for Post-polar pipeline,

$$\mathbb{P}\left(\frac{\langle \tilde{M}_t, G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F} \geq 1 - \frac{C}{2}\right) \leq \frac{4}{C^2(2T-1)}. \tag{9}$$

(iii) ***Pre-polar pipeline.*** *Denote $C' = \frac{2n\sqrt{\eta}}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}$. Then*

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right] \geq 1 - \frac{C'}{\sqrt{2T-1}}. \tag{10}$$



(a) Full-rank alignment vs. β. (b) Rank-5 alignment over training. (c) Full-rank alignment over training.

Figure 4: Stationary probe signal alignment for the three pipelines defined in equation (1)–equation (3) ($K = 500$). (a) β-sweep at the step-3000 checkpoint of the full-rank signal alignment. (b) rank-5 signal alignment at $\beta = 0.95$ across the five checkpoints. (c) Full-rank signal alignment at $\beta = 0.95$ across the same five checkpoints. The reference $\bar{G} = K^{-1}\sum_t G_t$ replaces G^{sig} on real gradients.

Moreover,

$$\mathbb{P}\left(\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F} \geq 1 - \frac{C'}{(2T-1)^{\frac{1}{4}}}\right) \geq 1 - \frac{1}{(2T-1)^{\frac{1}{4}}}. \tag{11}$$

(iv) ***Dominance over both non-Pre-polar baselines.*** *There exists T_0 such that for all $T \geq T_0$,*

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right] \geq \max\left\{\mathbb{E}\left[\frac{\langle \mathcal{O}(G_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right], \mathbb{E}\left[\frac{\langle \tilde{M}_t, G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right]\right\} + \frac{C}{2}. \tag{12}$$

If we moreover assume Assumption 4, then

$$\mathbb{P}\left(\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F} \geq \frac{\langle \tilde{M}_t, G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F} + \frac{C}{4}\right) \geq 1 - \frac{4}{C^2(2T-1)} - \frac{1}{(2T-1)^{\frac{1}{4}}}. \tag{13}$$

The signal alignment score in Theorem 2 uses a fixed signal-dependent normalization, different from cosine similarity. It measures the signal component carried by the update, with dependence on both direction and magnitude. For Post-polar, the gap can reflect both directional disagreement and norm contraction. Thus Pre-polar pipeline strictly dominates both Polar-only and Post-polar pipelines for sufficiently large momentum windows in expected signal alignment. With the pairwise independence assumption (Assumption 4), Pre-polar pipeline also dominates Post-polar pipeline in probability. The proof is deferred to Appendix D.

Signal-recovery separation (in simulation). Figure 4 supports Theorem 2 on the same stationary NanoGPT probe at training step 3000 used in Section 3, on the `h.0.attn.c_proj` layer. Synthetic, CIFAR-10, and other-layer NanoGPT extensions are provided in Appendix I. The figures use two empirical *signal alignment* metrics on real gradients (see Appendix F.3 for details): the rank-5 alignment $\mathrm{Align}_r(A; \bar{G}) := \|U_r(\bar{G})^\top A\, V_r(\bar{G})\|_F/\sqrt{r} \in [0,1]$, which measures how much of A lies in the top-r singular subspaces of $\bar{G}$, and the full-rank alignment $\mathrm{Align}_{\mathrm{full}}(A; \bar{G}) := \langle A, \mathcal{O}(\bar{G})\rangle_F/\min(m,n) \in [-1,1]$, the signed Frobenius inner product between A and the polar factor $\mathcal{O}(\bar{G})$, normalized by $\min(m,n) = \|\mathcal{O}(\bar{G})\|_F^2$ so that $\mathcal{O}(\bar{G})$ itself scores 1. Here A is the placeholder for any of the pipelines $\mathcal{O}(G_t)$, $\mathcal{O}(M_t)$, or $\tilde{M}_t$. Figure 4a shows the full-rank signal alignment for Pre-polar, Post-polar and Polar-only pipelines versus the momentum coefficient β. Only Pre-polar shows the monotonically improving trend in β, which is consistent with the success probability equation (11) increasing in β. Figures 4b and 4c show the same signal alignment at fixed $\beta = 0.95$ across different checkpoints, but with rank-5 and full-rank matrices, respectively. In both cases, Pre-polar is consistently higher than Post-polar and Polar-only pipelines across the different training checkpoints, which again supports Pre-polar's noise-smoothening mechanism. Note that Pre-polar outperforms especially for steps 4000 and 5000, where training switches into the NanoGPT warmdown phase, i.e., linear learning-rate decay is used over the final 1450/5100 training steps.

Quantitative signal-recovery separation. Theorem 2 delineates that Pre-polar pipeline dominates Polar-only counterpart in the signal alignment, but we do not quantitatively characterize the separation constant $C > 0$. Despite being challenging under the general coherent signal in Assumption 1, we show below that Polar-only alignment vanishes as the gradient matrix size (m, n) grows, jointly under (i) the rank-1 spiked signal model, and (ii) the low SNR regime.

Assumption 5 (Rank-1 spiked model)**.** *For each t, assume that the gradient admits the signal-perturbation decomposition $G_t = G^{\text{sig}} + \Xi_t = \lambda u v^\top + \Xi_t$, where $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ are time-invariant unit vectors constituting the signal basis $G^{\text{sig}} = \lambda u v^\top$, $\lambda > 0$ is the signal strength, and each element of $\Xi_t \in \mathbb{R}^{m\times n}$ follows the standard normal $\mathcal{N}(0, \sigma_\Xi^2)$ independently.*

Theorem 3 (Pre-polar vs. Polar-only separation)**.** *Assume $m > n$ without loss of generality. Under the rank-1 spiked model (Assumption 5), we additionally assume the low SNR regime: $\lambda < 0.25\sigma_\Xi(\sqrt{m} - \sqrt{n})$. Then, the signal alignment under Polar-only pipeline is bounded as follows:*

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(G_t), G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right] \le \frac{1}{\sqrt{m}} + \frac{4\lambda}{\sigma_\Xi(\sqrt{m} - \sqrt{n})} + 4\sqrt{n}\exp\left(-\frac{c(\sqrt{m} - \sqrt{n})^2}{4}\right),$$

where $c > 0$ is an absolute constant.

An important consequence of Theorem 3 is that the signal alignment of Polar-only pipeline becomes marginal as the matrix size $m \to \infty$ relative to the signal strength λ, which is an extreme scenario of the low SNR regime. In this case, the rank-1 signal $uv^\top$ is hidden in the bulk noise, and it will never be recovered. Strikingly, Pre-polar pipeline continues to recover the target signal thanks to Theorem 2, as long as the effective window size T is sufficiently large. The proof is deferred to Appendix D.

5 Subspace Reliability and Pipeline Ordering Under Live Training

The stationary probe of Section 3 reflects the theoretical setting of Theorems 1 and 2 by holding the model weights fixed. To closely simulate a realistic gradient stream where the weights drift step-by-step, we use the *trajectory probe*, which maintains a sliding K-step gradient buffer alongside the optimizer during end-to-end NanoGPT training and applies the momentum pipelines to the buffer in collection order. This sliding buffer creates the in-buffer mean gradient $\bar{G}$ (introduced in Section 3), which is recomputed at each training step over the latest K gradients and serves as the empirical proxy for the unobservable signal G^{sig}. Importantly, the model weights and thus the signal matrix G^{sig} can drift over time in the trajectory probe, which serves as a sensitivity analysis of the time invariance supposed in Assumption 1. We analyze the `h.0.attn.c_proj` layer every 100 training steps over 3 seeds, with buffer size $K = 50$ and momentum coefficient $\beta = 0.95$. The full settings, the all-layer NanoGPT extension, and the CIFAR-10 trajectory extension appear in Appendices F, H and I, respectively.

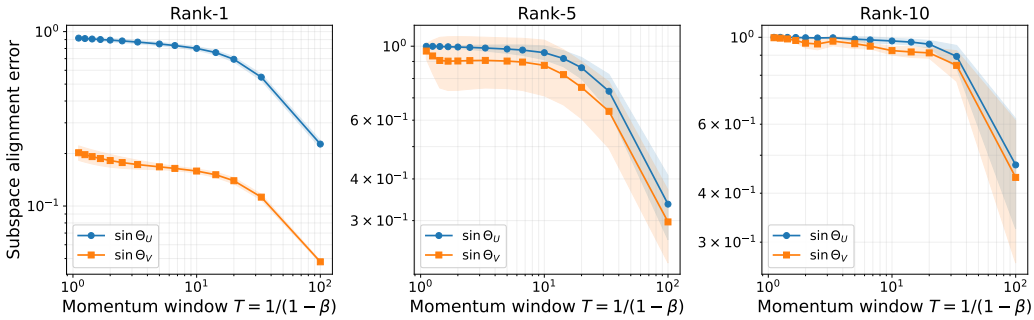
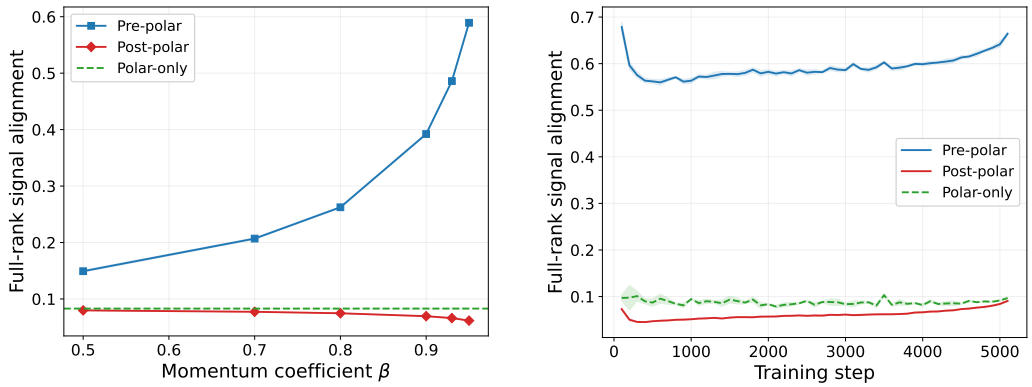


Figure 5: Trajectory probe subspace alignment errors $\sin\Theta_U, \sin\Theta_V$ at ranks $r \in \{1, 5, 10\}$ versus the momentum window size $T = 1/(1-\beta)$, with the β grid restricted to $\beta \le 0.95$ ($T \le K/2$). Curves show seed means. Shaded bands show sample standard deviation across seeds.

Subspace alignment error during training. Figure 5 reports the subspace alignment errors $\sin\Theta_U$ and $\sin\Theta_V$ of $M_K^{(\beta)}$, which is the momentum buffer with decay β after collecting a buffer of K

gradients, at ranks $r \in \{1, 5, 10\}$ versus the momentum window $T = 1/(1-\beta)$ on the layer `h.0.attn.c_proj` during NanoGPT training. The alignment errors averaged across different seeds decrease overall with T at every rank, validating that the rank-r subspace recovery (Corollary 1) is robust beyond the time-invariant signal in Assumption 1(a).

Comparison of signal alignment for three pipelines. Figure 6 reports the full-rank signal alignment (see Section 4) for the three pipelines on `h.0.attn.c_proj` during live NanoGPT training. All three pipelines are evaluated at the last gradient of the current buffer. Figure 6a sweeps β at training step 3000, and figure 6b tracks $\beta = 0.95$ across training steps.



(a) Full-rank alignment vs. β.

(b) Full-rank signal alignment over training.

Figure 6: Trajectory probe signal alignment for the three pipelines defined in equation (1)–equation (3) ($K = 50$, 3-seed mean). (a) β-sweep at step-3000 checkpoint of the full-rank signal alignment. (b) Full-rank signal alignment at $\beta = 0.95$ across training steps. Curves show seed means. Shaded bands show sample standard deviation across seeds.

In figure 6a, Pre-polar full-rank alignment rises monotonically with β, while Post-polar and Polar-only stay nearly flat at low values across the entire β range. Polar-only is β-independent by construction. The low Post-polar scores are consistent with the signal-alignment gap established in Theorem 2. In figure 6b, Pre-polar advantage at $\beta = 0.95$ persists across training steps. This trajectory result mirrors the stationary result of figure 4 supports that Pre-polar pipeline's dominance in signal alignment over both Post-polar and Polar-only survives the non-stationarity introduced by live training.

6 Conclusion

This paper explains why Muon orthogonalizes momentum rather than the raw gradient: momentum acts as a spectral filter that preserves the coherent signal while attenuating the bounded variance zero-mean orthogonal perturbation, so the polar step operates on a buffer in which the signal singular subspaces are already stabilized. Direct orthogonalization, or orthogonalizing each gradient before averaging, erases this structure and is provably worse in expected signal alignment. The same mechanism suggests a design rule that extends beyond Muon: in any matrix-aware optimizer that ends in a nonlinear spectral step, the spectral filter that separates signal from noise should come first — *denoise first, then orthogonalize*.

Limitations and future work. Several extensions are left open. First, practical Muon training often increases the momentum coefficient during early training, which enlarges the effective window size across iterations. Formalizing this schedule-dependent story requires a time-varying filter analysis. Second, Theorem 3 is proved under the rank-1 spiked Gaussian model. Extending it to general rank-r coherent signals or to non-Gaussian perturbations would require sharper concentration tools for the polar factor. Beyond these extensions, the spectral-gap mechanism may also guide the design of Muon variants with adaptive momentum updates.

Acknowledgment

XL and ZZ are supported by JST SPRING (Japan Grant Number JPMJSP2104). HB is supported by JST PRESTO (Grant Number JPMJPR24K6). A part of the experiments of this research was conducted using Wisteria/Aquarius in the Information Technology Center, the University of Tokyo.

References

[1] Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. In *Proceedings of the 14th International Conference on Learning Representations*, 2026. 16

[2] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022. 4, 16

[3] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real Algebraic Geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer-Verlag, Berlin, 1998. doi: 10.1007/978-3-662-03718-8. 22

[4] Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding orthogonalization in Muon. In *Proceedings of the 3rd Workshop on Efficient Systems for Foundation Models*, 2025. 16

[5] Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. *arXiv preprint arXiv:2601.22652*, 2026. 4, 16

[6] Dan Busbridge, Jason Ramapuram, Pierre Ablin, Tatiana Likhomanenko, Eeshan Gunesh Dhekane, Xavier Suau Cuadros, and Russell Webb. How to scale your EMA. *Advances in Neural Information Processing Systems*, 36:73122–73174, 2023. 17

[7] David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted Boltzmann machines. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 111–119. PMLR, 2015. 16

[8] David Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in Neural Information Processing Systems*, 28: 2971–2979, 2015. 16

[9] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. 2, 5, 16

[10] Yasuko Chikuse. *Statistics on Special Manifolds*, volume 174. Springer Science & Business Media, 2003. 24

[11] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2260–2268. PMLR, 2020. 1, 16

[12] Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025. 2, 16

[13] Aaron Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization. *arXiv preprint arXiv:2010.00406*, 2020. 16

[14] Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. RMNP: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026. 16

[15] 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, 2025. 2, 5, 16

[16] Everette S Gardner Jr. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4 (1):1–28, 1985. 3, 17

[17] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019. 16

[18] Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. In *Proceedings of the 14th International Conference on Learning Representations*, 2026. 4

[19] G. Goh. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. 16

[20] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018. 1, 16

[21] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. 4, 16

[22] Chuan He, Zhanwang Deng, and Zhaosong Lu. Low-rank orthogonalization for large-scale matrix optimization with applications to foundation model training. *arXiv preprint arXiv:2509.11983*, 2025. 16

[23] Nicholas J Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008. 25

[24] Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024. URL https://github.com/KellerJordan/modded-nanogpt. 52, 55

[25] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. 1, 2, 16, 55

[26] 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*, 2026. 2, 16

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. 1

[28] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025. 2, 5, 16

[29] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026. 2, 16

[30] Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. 1, 16

[31] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. 1, 16

[32] Weiyang Liu, Rongmei Lin, Zhen Liu, James M. Rehg, Liam Paull, Li Xiong, Le Song, and Adrian Weller. Orthogonal over-parameterized training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7251–7260, 2021. 16

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. 1

[34] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026. 2, 16

[35] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36: 71449–71485, 2023. 4, 16

[36] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983. 1, 16

[37] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024. 27

[38] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 49069–49104, 2025. 1, 16

[39] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. 1, 16

[40] Zeju Qiu, Simon Buchholz, Tim Z. Xiao, Maximilian Dax, Bernhard Schölkopf, and Weiyang Liu. Reparameterized LLM training via orthogonal equivalence transformation. *Advances in Neural Information Processing Systems*, 38:140775–140821, 2025. 16

[41] Zeju Qiu, Lixin Liu, Adrian Weller, Han Shi, and Weiyang Liu. POET-X: Memory-efficient LLM training by scaling orthogonal transformation. In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR, 2026. 16

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 27, 52, 55

[43] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs). In *Proceedings of the 3rd High-dimensional Learning Dynamics*, 2025. 16

[44] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017. 16

[45] Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025. 1

[46] Kexuan Shi, Hanxuan Li, Zeju Qiu, Yandong Wen, Simon Buchholz, and Weiyang Liu. Pion: A spectrum-preserving optimizer via orthogonal equivalence transformation. *arXiv preprint arXiv:2605.12492*, 2026. 16

[47] Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026. 2, 5, 16

[48] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019. 17

[49] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9058–9067. PMLR, 2020. 45

[50] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 33

[51] Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025. 2, 16

[52] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147. PMLR, 2013. 1, 3, 16, 17

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 26

[54] Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022. 2, 16

[55] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025. 16

[56] Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018. 25, 26

[57] Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. 1, 16

[58] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. 1

[59] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025. 2, 16

[60] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. 6, 20

[61] Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025. 1, 19

[62] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7654–7663. PMLR, 2019. 17

Appendix

Table of Contents

A Related Work **16**

B Setup Conventions and Variants **17**
B.1 Effective Sample Size of the Momentum Buffer 17
B.2 Polar Factor Equivalence with Heavy-ball Momentum 17
B.3 Initialization and the Post-Transient Regime 18
B.4 Practical Nesterov: Kernel Weights and Effective Sample Size 18

C Proofs in Section 3 **19**
C.1 Proof of Theorem 1 19
C.2 Proof of Corollary 1 20

D Proofs in Section 4 **20**
D.1 Proof of Theorem 2 22
D.2 Proof of Theorem 3 24

E Additional Discussions **25**
E.1 Sufficient Conditions for Signal Persistence 25
E.2 Sub-Gaussian Strengthening of Proposition 1 26

F Experimental Setups **26**
F.1 Shared Computational Settings 26
F.2 Probe Protocols 27
F.3 Measurements 29
F.4 Tasks 31
F.5 Signal-Strength Sensitivity 35

G Experimental Results of Spectral Gaps **35**

H Experimental Results of Subspace Alignment **36**

I Experimental Results of Signal Alignment Ordering **40**

J Hyperparameter Sensitivity in Signal Strength **45**

K Hyperparameter Sensitivity in Buffer Size **47**

A Related Work

Matrix-based and LMO-based optimizers. Matrix-based optimizer has moved beyond classical coordinate-wise methods. Earlier precursors of spectral descent in restricted Boltzmann machines and deep networks include Carlson et al. [7, 8]. Tuddenham et al. [54] introduce Orthogonal-SGDM, which orthogonalizes each per-step gradient before the momentum averaging. Shampoo [20] and SOAP [57] use matrix-aware preconditioning, while Scion [38] and Gluon [43] are LMO-based methods that operate in norm-constrained matrix geometry. Muon [25] applies a Newton-Schulz update to a momentum LLM training [31]. Pion [46] uses double-sided orthogonal transformation to design a matrix-based spectrum-preserving optimizer. A parallel line of work studies practical refinements of Muon's orthogonalization step, including the polar express that computes polar decomposition with optimal worst-case convergence [1], low-rank orthogonalization [22], block-wise orthogonalization [4], and row-momentum normalized preconditioning [14].

Theory of Muon and spectral updates. Recent studies have started to provide a theoretical foundation for the Muon optimizer and spectral updates. One line places Muon and related orthogonalization updates in norm-constrained geometry: Chen et al. [9] view Muon as a nuclear-norm Lion-K optimizer whose decoupled weight decay implicitly enforces spectral-norm constraints on the weights; Kovalev [28] develops a non-Euclidean trust-region interpretation of gradient orthogonalization and its momentum variant; and Shulgin et al. [47] analyze Muon as an inexact LMO method, quantifying how Newton–Schulz approximation error couples with the step size and momentum coefficient. A second line studies the implicit bias of spectral descent and Muon, proving convergence to spectral-norm max-margin solutions on multiclass separable data [15]. A third line asks when spectral orthogonalization should improve optimization, establishing preconditioning benefits in matrix factorization and in-context linear transformers [34], isotropic-curvature conditions under which spectrum homogenization or orthogonalization is optimal [51], and stable-rank/nuclear-rank conditions predicting when spectral updates outperform Euclidean gradient steps [12]. A fourth line analyzes structured tasks where per-component disparities are visible: Muon-type spectral updates balance learning across associative-memory frequencies and storage capacity [29, 26], heavy-tailed class distributions [59], imbalanced principal components [55], and anisotropic curvature in phase retrieval [5]. Additionally, a recent research line [32, 40, 41, 46] studies how spectrum preservation can serve as a guiding principle to design optimizers. Together, these works clarify spectral-update geometry, implicit bias, and task-level scaling behavior from complementary angles.

Research progress on the momentum algorithm. Momentum is among the most extensively analyzed primitives in optimization, with theoretical study organized along several complementary aspects. In classical deterministic optimization, Polyak [39] show heavy-ball achieves an accelerated linear rate on smooth strongly convex quadratics, with a geometric and spectral exposition by Goh [19]. Nesterov's accelerated gradient method attains the optimal $O(1/k^2)$ rate for smooth convex objectives [36]. Later, Sutskever et al. [52] extend these schemes to deep network training. In stochastic optimization, Cutkosky and Mehta [11] prove that momentum eliminates the large-batch requirement of normalized SGD, and Defazio [13] reformulates SGD with momentum as primal averaging to obtain sharper non-convex convergence bounds. Inspired by signal processing theory, Li et al. [30] interpret momentum in the frequency domain as a low-pass filter that amplifies low-frequency gradient components and attenuates high-frequency components, with high-frequency suppression becoming more important in late training. Together, these works examine momentum from deterministic-acceleration, stochastic-convergence, and signal-processing angles.

Gradient and noise structure in deep learning. Understanding the empirical structure of gradients and gradient noise during neural-network training has been a central concern of modern learning theory. On the gradient side, the Hessian spectrum of an overparametrized network often separates into a near-zero bulk and a few outliers, and the gradient can become largely concentrated in the corresponding top eigenspaces [44, 17]. This is consistent with the observation that gradient descent often proceeds mostly within a small subspace spanned by the top Hessian eigenvectors [21]. In high-dimensional feature-learning regimes, the first gradient update of the first-layer weights can contain a rank-1 spike aligned with the target signal [2]. Relatedly, under structured input distributions such as spiked covariance models or multiple-index teacher models, gradient-based training can exploit the hidden low-dimensional structure and drive the learned representation toward the corresponding signal subspace [35]. On the noise side, empirical studies suggest that mini-batch gradient noise

can be highly non-Gaussian and heavy-tailed [48]. Orthogonally, SGD noise has been shown to be anisotropic, with its covariance aligned with the Hessian in a way that facilitates escape from sharp minima [62]. Together, these threads establish that gradient and noise structure in modern deep learning training.

B Setup Conventions and Variants

Section 2 adopts the EMA normalization equation (4) and zero initialization $M_0 = 0$. We record here (i) the *effective sample size* identity that underlies the perturbation variance bound used in Theorems 1 and 2; (ii) the polar-factor equivalence between EMA and Polyak/heavy-ball momentum that justifies the EMA normalization choice; (iii) the buffer's behavior under arbitrary initialization, which defines the post-transient threshold $t \geq c_0 T$ characterizing when the EMA buffer has reached its asymptotic regime; and (iv) the kernel weights and effective sample size of practical Nesterov momentum.

B.1 Effective Sample Size of the Momentum Buffer

The perturbation bounds in Theorems 1 and 2 are driven by a single derived scalar quantity,

$$N_{\text{eff}} := \frac{(\sum_{s\geq 0} h_s)^2}{\sum_{s\geq 0} h_s^2} = \frac{1+\beta}{1-\beta} = 2T - 1, \tag{14}$$

the *effective sample size* of the momentum buffer. For a simple average of N uncorrelated, mean-zero random variables, the variance scales down by a factor of $1/N$. The bounded variance zero-mean orthogonal perturbation sequence $\{\Xi_t\}$ is pairwise uncorrelated in the Frobenius inner product ($\mathbb{E}\left[\langle \Xi_s, \Xi_t \rangle_F\right] = 0$ for $s \neq t$, see Assumption 1(b) and Proposition 1), so the variance of the momentum-weighted perturbation $S_t := \sum_{s\geq 0} h_s \Xi_{t-s}$ is reduced by $\sum_{s\geq 0} h_s^2 = 1/(2T-1)$ [6] (see equation (14)). Thus, the momentum buffer with decay coefficient β provides the same variance reduction as averaging over $2T-1$ independent samples. This factor determines the perturbation bound in Theorems 1 and 2.

B.2 Polar Factor Equivalence with Heavy-ball Momentum

For the decay $\beta \in [0, 1)$ and scalar $\gamma > 0$, define the general momentum recursion

$$M_t = \beta M_{t-1} + \gamma G_t, \tag{15}$$

where G_t is the incoming fresh gradient matrix. Under zero initialization $M_0 = 0$, the recursion unrolls to

$$M_t = \sum_{s=0}^{t-1} h_s \, G_{t-s}, \qquad \text{where} \qquad h_s := \gamma \beta^s.$$

The decay β sets how far back the buffer effectively remembers, while γ sets the overall scale. The two standard normalizations are

$$\text{EMA momentum [16]: } \gamma = 1 - \beta, \qquad \text{Polyak (heavy-ball) momentum [52]: } \gamma = 1.$$

Pre-polar update equation (1) applies the polar factor $\mathcal{O}(\cdot)$ to the buffer $M_t = \sum_{s=0}^{t-1} h_s \, G_{t-s}$. Because the polar factor is scale-invariant, $\mathcal{O}(\alpha X) = \mathcal{O}(X)$ for every $\alpha > 0$, we may divide M_t by the positive scalar $A := \sum_{s'\geq 0} h_{s'} = \gamma/(1-\beta)$ without changing the polar factor:

$$\mathcal{O}(M_t) = \mathcal{O}\left(\frac{M_t}{A}\right) = \mathcal{O}\left(\sum_{s=0}^{t-1} \frac{h_s}{A} \, G_{t-s}\right) = \mathcal{O}\left(\sum_{s=0}^{t-1} (1-\beta)\beta^s \, G_{t-s}\right),$$

where the last equality uses $h_s/A = \gamma\beta^s \cdot (1-\beta)/\gamma = (1-\beta)\beta^s$. The scalar γ drops out, and thus every first-order recursion with the same decay β produces the same Pre-polar update. In particular, Polyak/heavy-ball momentum ($\gamma = 1$) and EMA momentum ($\gamma = 1-\beta$) generate identical Pre-polar updates. The EMA normalization $\gamma = 1 - \beta$ adopted in equation (4) picks the unique γ for which the infinite kernel sums to one ($\sum_{s\geq 0} h_s = 1$), so the buffer can be read directly as a weighted average of past gradients.

Remark 2 (Polar factor equivalence of EMA and heavy-ball momentum)**.** *Let $M_t^{(\beta,\gamma)}$ denote the zero-initialized buffer generated by equation (15), and let $M_t^{\mathrm{ema}(\beta)} = (1-\beta)\sum_{s=0}^{t-1}\beta^s G_{t-s}$ be the EMA-normalized buffer with the same decay. Then $M_t^{(\beta,\gamma)} = \frac{\gamma}{1-\beta} M_t^{\mathrm{ema}(\beta)}$, so $\mathcal{O}(M_t^{(\beta,\gamma)}) = \mathcal{O}(M_t^{\mathrm{ema}(\beta)})$ for every $\gamma > 0$.*

B.3 Initialization and the Post-Transient Regime

Under an arbitrary initialization $M_0 = M_{\mathrm{init}}$, the recursion equation (4) unrolls to

$$M_t = \beta^t M_{\mathrm{init}} + (1-\beta)\sum_{s=0}^{t-1}\beta^s G_{t-s},$$

so the buffer carries an additional $\beta^t M_{\mathrm{init}}$ term, an exponentially decaying transient of operator-norm size at most $\beta^t \|M_{\mathrm{init}}\|_2$. We say the recursion has entered its *post-transient regime* when $t \geq c_0 T$ for a constant $c_0 > 0$ chosen so that $\beta^{c_0 T}\|M_{\mathrm{init}}\|_2$ is negligible relative to the signal scale.

The same threshold $t \geq c_0 T$ also governs the EMA's warmup under the default zero initialization. With $M_{\mathrm{init}} = 0$, the finite-time buffer $M_t = (1-\beta)\sum_{s=0}^{t-1}\beta^s G_{t-s}$ carries total kernel weight $1-\beta^t$ rather than 1, so the buffer recovers a fraction $1-\beta^t$ of a constant signal, which would violate the persistence floor $c_{\mathrm{sig}}\lambda_k$ in Assumption 2 during the warmup phase. The condition $t \geq c_0 T$ ensures this $O(\beta^t)$ error is negligible, recovering the same regime as in the arbitrary-initialization case.

B.4 Practical Nesterov: Kernel Weights and Effective Sample Size

Practical deep-learning Nesterov momentum modifies the EMA recursion equation (4) by passing a linear combination of the buffer M_t and the current gradient G_t to the polar factor, rather than M_t itself. With the general first-order recursion $M_t = \beta M_{t-1} + \gamma G_t$ and scalars ν, κ, the parameter update is formed from

$$N_t \;=\; \nu\, M_t + \kappa\, G_t.$$

Under zero initialization $M_0 = 0$, the buffer unrolls to $M_t = \sum_{s=0}^{t-1}\gamma\beta^s G_{t-s}$, so

$$N_t \;=\; (\nu\gamma+\kappa)\, G_t \;+\; \sum_{s=1}^{t-1}\nu\gamma\,\beta^s\, G_{t-s} \;=\; \sum_{s=0}^{t-1} h_s\, G_{t-s}, \qquad h_0 = \nu\gamma+\kappa, \quad h_s = \nu\gamma\,\beta^s\ (s\geq 1).$$

Non-equivalence with plain momentum. Appendix B.2 shows that $\mathcal{O}(M_t)$ depends only on the decay β, since multiplying all geometric weights $\gamma\beta^s$ by the same positive constant leaves the polar factor unchanged. For N_t, however, changing γ while keeping κ fixed changes the relative weight of the current gradient versus the geometric tail. Thus the kernel shape is not determined by β alone. In the nondegenerate case $\nu\gamma \neq 0$ and $0 < \beta < 1$, matching plain-momentum geometric weights $\tilde{h}_s = \tilde{\gamma}\tilde{\beta}^s$ to h_s for all $s \geq 1$ forces $\tilde{\beta} = \beta$ and $\tilde{\gamma} = \nu\gamma$. Matching the current-gradient weight then requires $h_0 = \nu\gamma + \kappa = \tilde{\gamma}$, hence $\kappa = 0$. Therefore practical Nesterov reduces to plain momentum exactly when $\kappa = 0$.

EMA-normalized Muon Nesterov. The common implementation uses $\gamma = 1-\beta$, $\nu = \beta$, $\kappa = 1-\beta$, giving

$$h_0 \;=\; 1-\beta^2, \qquad h_s \;=\; (1-\beta)\beta^{s+1} \quad (s\geq 1).$$

In the stationary/infinite-tail approximation, the kernel mass is one,

$$\sum_{s\geq 0} h_s \;=\; (1-\beta^2) \;+\; (1-\beta)\beta^2\sum_{s\geq 0}\beta^s \;=\; (1-\beta^2)+\beta^2 \;=\; 1,$$

so N_t remains a weighted average of past gradients. The squared kernel mass, which controls the variance proxy of Proposition 1 through equation (14), is

$$\sum_{s\geq 0} h_s^2 \;=\; (1-\beta^2)^2 + \frac{(1-\beta)^2\beta^4}{1-\beta^2} \;=\; \frac{(1-\beta)\,(1+2\beta-2\beta^3)}{1+\beta},$$

so the effective sample size becomes

$$N_{\text{eff}}^{\text{nest}} = \frac{1}{\sum_{s\geq 0} h_s^2} = \frac{1+\beta}{(1-\beta)(1+2\beta-2\beta^3)} < \frac{1+\beta}{1-\beta} = 2T-1,$$

with strict inequality because $1+2\beta-2\beta^3 = 1+2\beta(1-\beta^2) > 1$ for $\beta \in (0,1)$.

Compared with plain momentum at the same β, practical Nesterov has a strictly smaller effective sample size and therefore attenuates the momentum-filtered perturbation in Proposition 1 less aggressively, while placing more weight on the current gradient ($h_0 = 1-\beta^2 > 1-\beta$). This trade-off explains why enabling Nesterov can help at fixed hyperparameters, while retuned plain momentum can produce very similar training curves in practice [61].

C Proofs in Section 3

To control the perturbation terms in Theorem 1, we first establish the following concentration estimate.

Proposition 1 (Concentration of the momentum-filtered perturbation)**.** *Let $\{\Xi_t\}_{t\in\mathbb{Z}}$ be an $m\times n$ matrix-valued perturbation sequence that satisfies Assumption 1(b) for $t\geq 0$, and $\Xi_t = 0$ for $t<0$. Define*

$$S_t := (1-\beta)\sum_{s\geq 0}\beta^s \Xi_{t-s}, \qquad T := \frac{1}{1-\beta}.$$

Then for every $u>0$, we have

$$\mathbb{P}\left(\|S_t\|_F \geq \sqrt{\frac{\eta}{2T-1}}u\right) \leq \frac{1}{u^2}.$$

As a corollary,

$$\mathbb{P}\left(\|S_t\|_2 \geq \sqrt{\frac{\eta}{2T-1}}u\right) \leq \frac{1}{u^2}.$$

Proof of Proposition 1. By Assumption 1(b) the perturbation $\{\Xi_t\}$ is pairwise uncorrelated since orthogonal in the Frobenius inner product, $\mathbb{E}\left[\langle \Xi_{t_1}, \Xi_{t_2}\rangle_F\right] = 0$ for $t_1 \neq t_2$.

Then we have

$$\mathrm{Var}_F\left[S_t\right] = \mathbb{E}\left[\|S_t\|_F^2\right] = \sum_{s=0}^{\infty} h_s^2\, \mathbb{E}\left[\|\Xi_{t-s}\|_F^2\right] \leq \frac{\eta}{2T-1}$$

by equation (14).

By Chebyshev's bound we have

$$\mathbb{P}\left(\|S_t\|_F \geq \sqrt{\frac{\eta}{2T-1}}u\right) \leq \frac{1}{u^2}.$$

The corollary is a direct consequence of the fact that the Frobenius norm is an upper bound of the operator norm. □

C.1 Proof of Theorem 1

Proof. By the decomposition equation (5), defining $G_t = 0$ for $t<0$, we have

$$\begin{aligned} M_t &= (1-\beta)\sum_{s\geq 0}\beta^s G_{t-s} \\ &= (1-\beta)\sum_{s\geq 0}\beta^s G_{t-s}^{\text{sig}} + (1-\beta)\sum_{s\geq 0}\beta^s \Xi_{t-s} \\ &= \sum_{i=1}^{r}\bar{\lambda}_i(t) u_i v_i^\top + S_t. \end{aligned}$$

Let M_t^{sig} denote $\sum_{i=1}^r \bar{\lambda}_i(t) u_i v_i^\top$.

Signal preservation. For $k = 1, \dots, r$, by the min-max theorem of singular values, the k-th singular value of M_t is exactly

$$\sigma_k(M_t) = \max_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=k}} \min_{\substack{x \in S \\ \|x\|=1}} \|M_t x\|.$$

We may take $S = \operatorname{span}\{v_1, \dots, v_k\}$. Then, for every unit vector $x \in S$, the triangle inequality and the definition of the operator norm imply

$$\|M_t x\| \geq \|M_t^{\text{sig}} x\| - \|S_t x\| \geq \min_{1 \leq i \leq k} |\bar{\lambda}_i(t)| - \|S_t\|_2 \geq c_{\text{sig}} \lambda_k - \|S_t\|_2 .$$

Here the first inequality uses the triangle inequality. The second uses the orthonormality of $\{u_i\}$ and $\{v_i\}$ together with $\|S_t x\| \leq \|S_t\|_2$ for $\|x\| = 1$. The last uses Assumption 2 and $\lambda_1 \geq \cdots \geq \lambda_r$. Eventually, Proposition 1 leads to the following bound:

$$\mathbb{P}\left(\sigma_k(M_t) \leq c_{\text{sig}} \lambda_k - \sqrt{\frac{\eta}{2T-1}} u\right) \leq \mathbb{P}\left(\|S_t\|_2 > \sqrt{\frac{\eta}{2T-1}} u\right) \leq \frac{1}{u^2}.$$

Perturbation attenuation. Similarly, for $k = r+1, \dots, n$, by the min-max theorem,

$$\sigma_k(M_t) = \min_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=n-k+1}} \max_{\substack{x \in S \\ \|x\|=1}} \|M_t x\|.$$

We may take S be one arbitrary subspace of $\operatorname{span}\{v_1, \dots, v_{k-1}\}^\perp$ so that M_t restricted to S is zero, then for every unit $x \in S$,

$$\|M_t x\| \leq \|M_t^{\text{sig}} x\| + \|S_t x\| \leq 0 + \|S_t\|_2 .$$

Then

$$\mathbb{P}\left(\sigma_k(M_t) \geq \sqrt{\frac{\eta}{2T-1}} u\right) \leq \mathbb{P}\left(\|S_t\|_2 \geq \sqrt{\frac{\eta}{2T-1}} u\right) \leq \frac{1}{u^2}.$$

Taking $u = (2T-1)^{\frac{1}{4}}$ concludes the proof. □

C.2 Proof of Corollary 1

Proof. Apply Wedin's $\sin\Theta$ theorem for singular subspaces [60] to the decomposition

$$M_t = M_t^{\text{sig}} + R_t.$$

The clean matrix M_t^{sig} has left and right signal subspaces exactly $\operatorname{span}(U)$ and $\operatorname{span}(V)$, and we have

$$\sigma_r(M_t^{\text{sig}}) - \sigma_{r+1}(M_t) \geq c_{\text{sig}} \lambda_r - \frac{\sqrt{\eta}}{(2T-1)^{1/4}}$$

with probability at least $1 - (2T-1)^{-1/2}$ by Theorem 1. Wedin's $\sin\Theta$ theorem therefore gives,

$$\max\left\{\|\sin\Theta(\hat{U}_t, U)\|_2,\ \|\sin\Theta(\hat{V}_t, V)\|_2\right\} \leq \frac{\|R_t\|_2}{\sigma_r(M_t^{\text{sig}}) - \sigma_{r+1}(M_t)} \leq \frac{\sqrt{\eta}/(2T-1)^{1/4}}{c_{\text{sig}}\lambda_r - \sqrt{\eta}/(2T-1)^{1/4}},$$

where the last inequality uses $\|R_t\|_2 = \|S_t\|_2 \leq \sqrt{\eta}/(2T-1)^{1/4}$ from Proposition 1 on the same event. Both individual bounds in Corollary 1 follow. □

D Proofs in Section 4

Recall the notation in Assumption 1 and Assumption 2. For the ordering theorem, we additionally assume Assumption 3.

The following key lemma describes how orthogonalization over noise creates a bias that prevents recovery of the signal direction.

Lemma 1 (Quantitative gap for the polar factor)**.** *Let $G \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $\operatorname{rank}(G) = r$. G has thin SVD*

$$G = U_r \Sigma_r V_r^\top, \qquad \Sigma_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0), \qquad \text{and } \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

where $U_r \in \mathbb{R}^{m\times n}$, $V_r \in \mathbb{R}^{n\times n}$, and $\Sigma_r \in \mathbb{R}^{n\times n}$.

Let $\Xi \in \mathbb{R}^{m\times n}$ be any random matrix. Then we have

$$\mathbb{E}\left[\langle \mathcal{O}(G+\Xi), G\rangle_F\right] \leq \langle \mathcal{O}(G), G\rangle_F - \frac{\sigma_r}{2}\,\mathbb{E}\left[\left\|\big(\mathcal{O}(G+\Xi) - \mathcal{O}(G)\big)V_r\right\|_F^2\right].$$

Proof. Fix a deterministic matrix $X \in \mathbb{R}^{m\times n}$, and write

$$Q \coloneqq \mathcal{O}(X).$$

Since the singular values of $\mathcal{O}(X)$ are all either 0 or 1, one has $\|Q\|_2 \leq 1$. Moreover,

$$\mathcal{O}(G) = U_r V_r^\top.$$

Define

$$B \coloneqq U_r^\top Q V_r \in \mathbb{R}^{r\times r}.$$

Then $\|B\|_2 \leq 1$, since U_r and V_r have orthonormal columns.

Now

$$\langle Q, G\rangle_F = \operatorname{Tr}(Q^\top U_r \Sigma_r V_r^\top) = \operatorname{Tr}(B^\top \Sigma_r),$$

hence

$$\langle \mathcal{O}(G), G\rangle_F - \langle Q, G\rangle_F = \operatorname{Tr}(\Sigma_r) - \operatorname{Tr}(B^\top \Sigma_r) = \sum_{i=1}^{r} \sigma_i (1 - B_{ii}).$$

Since $\sigma_i \geq \sigma_r$ for all i,

$$\langle \mathcal{O}(G), G\rangle_F - \langle Q, G\rangle_F \geq \sigma_r \sum_{i=1}^{r} (1 - B_{ii}) = \sigma_r\big(r - \operatorname{Tr}(B)\big).$$

On the other hand,

$$\|(Q - \mathcal{O}(G))V_r\|_F^2 = \|QV_r - U_r\|_F^2 = \|QV_r\|_F^2 + \|U_r\|_F^2 - 2\operatorname{Tr}(U_r^\top Q V_r).$$

Because $\|Q\|_2 \leq 1$ and V_r has orthonormal columns,

$$\|QV_r\|_F^2 \leq r, \qquad \|U_r\|_F^2 = r, \qquad \operatorname{Tr}(U_r^\top Q V_r) = \operatorname{Tr}(B).$$

Therefore

$$\|(Q - \mathcal{O}(G))V_r\|_F^2 \leq 2r - 2\operatorname{Tr}(B),$$

or equivalently,

$$r - \operatorname{Tr}(B) \geq \frac{1}{2}\,\|(Q - \mathcal{O}(G))V_r\|_F^2.$$

Combining the last two estimates yields the pointwise bound

$$\langle \mathcal{O}(G), G\rangle_F - \langle \mathcal{O}(X), G\rangle_F \geq \frac{\sigma_r}{2}\,\|(\mathcal{O}(X) - \mathcal{O}(G))V_r\|_F^2.$$

Applying this with $X = G + \Xi$ and taking expectations proves the claim. □

Remark 3. *Let*

$$\mathcal{N}_G \coloneqq \{X \in \mathbb{R}^{m\times n} : \mathcal{O}(X) = \mathcal{O}(G)\}.$$

We have

$$\mathcal{O}(X) \coloneqq X(X^\top X)^{\dagger/2},$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. Thus,

$$\begin{aligned} &\mathcal{O}(X) = \mathcal{O}(G) \\ \iff &X(X^\top X)^{\dagger/2} = U_r V_r^\top \\ \iff &X = U_r V_r^\top (X^\top X)^{1/2}. \end{aligned}$$

A necessary condition for the last equation to hold is that

$$XX^\top = U_r V_r^\top (X^\top X) V_r U_r^\top . \tag{16}$$

Thus, $\mathcal{N}_G$ is a subset of the solution set to equation (16). The latter is a collection of non-trivial algebraic equations (i.e. non-zero polynomial equations in the entries of X) whose solution set has Lebesgue measure zero [3, §2.8]. In particular, if Ξ is a continuous random matrix, then $\mathbb{P}(\mathcal{O}(G+\Xi) = \mathcal{O}(G)) = 0$. Thus the gap in Lemma 1 is positive with probability one under continuous perturbations.

In conclusion, as long as the law of Ξ is absolutely continuous with respect to the Lebesgue measure, we have a strictly positive gap in Lemma 1.

The following lemma is used to control the bias for Pre-polar pipeline.

Lemma 2. *Let $G \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $\operatorname{rank}(G) = r$. G has thin SVD*

$$G = U_r \Sigma_r V_r^\top, \qquad \Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0), \qquad \text{and } \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

where $U_r \in \mathbb{R}^{m\times n}$, $V_r \in \mathbb{R}^{n\times n}$, and $\Sigma_r \in \mathbb{R}^{n\times n}$. For any $R \in \mathbb{R}^{m\times n}$, we have

$$\frac{\langle \mathcal{O}(G+R), G\rangle_F}{\langle \mathcal{O}(G), G\rangle_F} \geq 1 - \frac{2n\,\|R\|_2}{\langle \mathcal{O}(G), G\rangle_F}.$$

Proof. In this proof, we write $\|X\|_* = \sum_i \sigma_i(X)$ for the nuclear norm, the sum of all singular values. Since $\mathcal{O}(G) = U_r V_r^\top$,

$$\langle \mathcal{O}(G), G\rangle_F = \langle U_r V_r^\top, U_r \Sigma_r V_r^\top\rangle_F = \operatorname{Tr}(\Sigma_r) = \|G\|_* .$$

Now let $B = G + R$.

$$\langle \mathcal{O}(B), G\rangle_F = \langle \mathcal{O}(B), B - R\rangle_F = \|B\|_* - \langle \mathcal{O}(B), R\rangle_F .$$

By the triangle inequality,

$$\|B\|_* \geq \|G\|_* - \|R\|_* .$$

Since R has rank at most n, we have $\|R\|_* \leq n\,\|R\|_2$, hence

$$\|B\|_* \geq \|G\|_* - n\,\|R\|_2 .$$

Also, $\mathcal{O}(B)$ has at most n singular values equal to 1, so

$$\|\mathcal{O}(B)\|_* \leq n.$$

Hence, by duality of nuclear and operator norms,

$$|\langle \mathcal{O}(B), R\rangle_F| \leq \|\mathcal{O}(B)\|_*\,\|R\|_2 \leq n\,\|R\|_2 .$$

Therefore

$$\langle \mathcal{O}(B), G\rangle_F \geq \|G\|_* - 2n\,\|R\|_2 .$$

Dividing by $\langle \mathcal{O}(G), G\rangle_F = \|G\|_*$ gives

$$\frac{\langle \mathcal{O}(G+R), G\rangle_F}{\langle \mathcal{O}(G), G\rangle_F} \geq 1 - \frac{2n\,\|R\|_2}{\|G\|_*}.$$

□

D.1 Proof of Theorem 2

Proof. **Part (i).** By Lemma 1 and Remark 3, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(G_t), G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right] \leq 1 - C.$$

Part (ii). By the definition of $\widetilde{M}_t$, Lemma 1, and Remark 3, for the same constant $C > 0$ we have

$$\mathbb{E}\left[\frac{\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right] = (1-\beta)\sum_{s=0}^{t} \beta^s\, \mathbb{E}\left[\frac{\langle \mathcal{O}(G_{t-s}), G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right] \leq (1-\beta)\sum_{s\geq 0} \beta^s (1-C) = 1-C.$$

Now consider the Assumption 4. Since for any X, $\left|\langle \mathcal{O}(X), G^{\text{sig}}\rangle_F\right| \le \langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F$, the variance

$$\text{Var}\left[\langle \mathcal{O}(G_t), G^{\text{sig}}\rangle_F\right] \le \langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F^2$$

for every $t \ge 0$. So we have

$$\begin{aligned}
\text{Var}\left[\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F\right] &= \text{Var}\left[(1-\beta)\sum_{s=0}^{t}\beta^s \langle \mathcal{O}(G_{t-s}), G^{\text{sig}}\rangle_F\right] \\
&= (1-\beta)^2 \sum_{s=0}^{t}\beta^{2s}\,\text{Var}\left[\langle \mathcal{O}(G_{t-s}), G^{\text{sig}}\rangle_F\right] \\
&\le (1-\beta)^2 \sum_{s=0}^{t}\beta^{2s}\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F^2 \\
&\le (1-\beta)^2 \cdot \frac{1}{1-\beta^2}\cdot \langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F^2 \\
&= \frac{1-\beta}{1+\beta}\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F^2 \\
&= \frac{1}{2T-1}\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F^2.
\end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F} \ge 1-\frac{C}{2}\right) \\
&= \mathbb{P}\left(\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F \ge \left(1-\frac{C}{2}\right)\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F\right) \\
&= \mathbb{P}\left(\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F - \mathbb{E}\left[\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F\right] \ge \left(1-\frac{C}{2}\right)\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F - \mathbb{E}\left[\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F\right]\right) \\
&\le \mathbb{P}\left(\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F - \mathbb{E}\left[\langle \widetilde{M}_t, G^{\text{sig}}\rangle_F\right] \ge \frac{C\sqrt{2T-1}}{2}\frac{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}{\sqrt{2T-1}}\right) \\
&\le \frac{4}{C^2(2T-1)}.
\end{aligned}$$

Part (iii). By Lemma 2, we have

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(M_t), G^{\text{sig}}\rangle_F}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right] \ge 1 - \mathbb{E}\left[\frac{2n\|M_t - G^{\text{sig}}\|_2}{\langle \mathcal{O}(G^{\text{sig}}), G^{\text{sig}}\rangle_F}\right].$$

Similar to the proof of Proposition 1, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|M_t - G^{\text{sig}}\right\|_F^2\right] &= \mathbb{E}\left[\left\|(1-\beta)\sum_{s=0}^{t}\beta^s \Xi_{t-s}\right\|_F^2\right] \\
&= (1-\beta)^2\sum_{s=0}^{t}\beta^{2s}\,\mathbb{E}\left[\|\Xi_{t-s}\|_F^2\right] \\
&\le \frac{1}{2T-1}\,\eta,
\end{aligned}$$

where the second equality uses pairwise uncorrelatedness and the inequality uses $\mathbb{E}\,\|\Xi_{t-s}\|_F^2 \le \eta$ from Assumption 1(b).

By Lyapunov's inequality,

$$\mathbb{E}\left[\|M_t - G^{\text{sig}}\|_2\right] \le \mathbb{E}\left[\|M_t - G^{\text{sig}}\|_2^2\right]^{\frac{1}{2}} \le \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{2}}}.$$

Therefore,

$$\mathbb{E}\left[\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}\right] \geq 1 - \frac{2n\sqrt{\eta}}{(2T-1)^{\frac{1}{2}}\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}.$$

By taking $C' = \frac{2n\sqrt{\eta}}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F}$, we obtain equation (10).

By Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\langle \mathcal{O}(M_t), G^{\mathrm{sig}}\rangle_F}{\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F} \leq 1 - \frac{C'}{(2T-1)^{\frac{1}{4}}}\right) \\
&\leq \mathbb{P}\left(\|M_t - G^{\mathrm{sig}}\|_2 \geq \frac{\sqrt{\eta}}{(2T-1)^{\frac{1}{4}}}\right) \\
&\leq \mathbb{E}\left[\|M_t - G^{\mathrm{sig}}\|_2\right] \cdot (2T-1)^{\frac{1}{4}} \\
&\leq \frac{1}{(2T-1)^{\frac{1}{4}}}.
\end{aligned}$$

Part (iv). Taking T_0 such that $\frac{C'}{\sqrt{2T_0-1}} \leq \frac{C}{2}$ proves the first claim. The second claim is a union bound of the probability bounds for Pre-polar and Post-polar pipelines. □

D.2 Proof of Theorem 3

Proof. The rank-1 spiked model (Assumption 5) yields $\langle \mathcal{O}(G^{\mathrm{sig}}), G^{\mathrm{sig}}\rangle_F = \lambda \|uv^\top\|_F^2 = \lambda$, which means that the signal alignment under Polar-only pipeline is $\mathbb{E}[\langle \mathcal{O}(G_t), uv^\top\rangle_F]$. We derive the upper bound on this quantity, whose integrand is bounded as follows via the Cauchy–Schwarz inequality:

$$\begin{aligned}
\langle \mathcal{O}(G_t), uv^\top\rangle_F &= \langle \mathcal{O}(\Xi_t), uv^\top\rangle_F + \langle \mathcal{O}(G_t) - \mathcal{O}(\Xi_t), uv^\top\rangle_F \\
&\leq \underbrace{\langle \mathcal{O}(\Xi_t), uv^\top\rangle_F}_{=:(\clubsuit)} + \underbrace{\|\mathcal{O}(G_t) - \mathcal{O}(\Xi_t)\|_F}_{=:(\diamondsuit)}.
\end{aligned}$$

To address the term (♣), note that the polar factor of an element-wise Gaussian matrix $\mathcal{O}(\Xi_t)$ is uniformly distributed over the Stiefel manifold $\mathrm{St}(m,n) \coloneqq \{U \in \mathbb{R}^{m\times n} : U^\top U = I_n\}$ [10, Theorem 2.2.1]. This implies that $\mathcal{O}(\Xi_t)$ and $V_1\,\mathcal{O}(\Xi_t)V_2$ for the pair of orthogonal actions $(V_1, V_2) \in O(m)\times O(n)$ follow the same law. Thus, we can take $u = e_m^{(1)} \in \mathbb{R}^m$ and $v = e_n^{(1)} \in \mathbb{R}^n$ (the standard basis vector with the first coordinate being one) without loss of generality because the underlying law of $\langle \mathcal{O}(\Xi_t), uv^\top\rangle_F$ and $\langle \mathcal{O}(\Xi_t), e_m^{(1)}e_n^{(1)\top}\rangle_F$ remains the same with any pair of orthogonal actions (V_1, V_2) satisfying $u = V_1^\top e_m^{(1)}$ and $v = V_2 e_n^{(1)}$. Now, we can focus on $\langle \mathcal{O}(\Xi_t), e_m^{(1)}e_n^{(1)\top}\rangle_F = [\mathcal{O}(\Xi_t)]_{11}$ instead. Since $\mathcal{O}(\Xi_t) \in \mathrm{St}(m,n)$, we immediately have $\sum_{i=1}^m [\mathcal{O}(\Xi_t)]_{i1}^2 = 1$. By noting that the law of $\mathcal{O}(\Xi_t)$ is invariant with the orthogonally actioned $V\,\mathcal{O}(\Xi_t)$ with $V \in O(m)$, all of $[\mathcal{O}(\Xi_t)]_{11}, \ldots, [\mathcal{O}(\Xi_t)]_{m1}$ follow the same law. This indicates that we have $\mathbb{E}[\sum_{i=1}^m [\mathcal{O}(\Xi_t)]_{i1}^2] = m\mathbb{E}[[\mathcal{O}(\Xi_t)]_{11}^2] = 1$. Therefore,

$$\mathbb{E}[\langle \mathcal{O}(\Xi_t), uv^\top\rangle_F] = \mathbb{E}[[\mathcal{O}(\Xi_t)]_{11}] \leq \sqrt{\mathbb{E}[[\mathcal{O}(\Xi_t)]_{11}^2]} = \frac{1}{\sqrt{m}},$$

where we used Jensen's inequality.

To address the term (♢), let $\mathcal{E} \coloneqq \{\sigma_{\min}(\Xi_t) \geq s_*/2\}$ for $s_* \coloneqq \sigma_\Xi(\sqrt{m} - \sqrt{n})$ denotes the "good" event, where $\sigma_{\min}(\Xi_t) > \lambda$ holds. We bound (♢) by viewing G_t as a perturbation of Ξ_t toward the rank-1 spike $\lambda uv^\top$. Write $G_{t,\tau} \coloneqq \Xi_t + \tau\lambda uv^\top$ $(0 < \tau < 1)$ as an intermediate perturbation. Before proceeding, let us ensure the differentiability of the polar factor $\mathcal{O}(\cdot)$ at $G_{t,\tau}$, for which $\sigma_{\min}(G_{t,\tau}) > 0$ is necessary by the definition of the polar factor — in fact, to ensure the invertibility of $G_{t,\tau}^\top G_{t,\tau}$. By Weyl's inequality for singular values, we have

$$\sigma_{\min}(G_{t,\tau}) \geq \sigma_{\min}(\Xi_t) - \tau\lambda \left\|uv^\top\right\|_2 = \sigma_{\min}(\Xi_t) - \tau\lambda \geq \sigma_{\min}(\Xi_t) - \lambda. \tag{17}$$

Thus, $\sigma_{\min}(G_{t,\tau}) > 0$ is ensured on the "good" event $\mathcal{E}$. In this case, the fundamental theorem of calculus gives

$$\mathcal{O}(G_t) - \mathcal{O}(\Xi_t) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathcal{O}(G_{t,\tau})\mathrm{d}\tau = \int_0^1 D\,\mathcal{O}(G_{t,\tau})[\lambda uv^\top]\mathrm{d}\tau,$$

where $D\,\mathcal{O}(\cdot)[H]$ is the Gâteaux derivative of the polar factor in the direction of $H \in \mathbb{R}^{m\times n}$. Note that the polar factor is locally Lipschitz [23], that is, $\|D\,\mathcal{O}(X)[H]\|_F \le \|H\|_F/\sigma_{\min}(X)$. Using this, we have

$$\|\,\mathcal{O}(G_t) - \mathcal{O}(\Xi_t)\|_F \le \int_0^1 \left\|D\,\mathcal{O}(G_{t,\tau})[\lambda uv^\top]\right\|_F \mathrm{d}\tau \le \frac{\lambda}{\sigma_{\min}(G_{t,\tau})} \le \frac{\lambda}{\sigma_{\min}(\Xi_t) - \lambda},$$

where Weyl's inequality equation (17) is used at the last inequality. By combining the conditioned event $\mathcal{E}$ and the low SNR assumption $\lambda < s_*/4$, we finally have $\|\,\mathcal{O}(G_t) - \mathcal{O}(\Xi_t)\|_F \le 4\lambda/s_*$. In contrast, conditioning on the "bad" event $\mathcal{E}^\complement$, we use a more elementary bound $\|\,\mathcal{O}(G_t) - \mathcal{O}(\Xi_t)\|_F \le \|\,\mathcal{O}(G_t)\|_F + \|\,\mathcal{O}(\Xi_t)\|_F = 2\sqrt{n}$. To properly control, let us evaluate $\Pr(\mathcal{E}^\complement)$. By Vershynin [56, Exercise 7.3.4], we have the two-sided bound on the (least) singular value of a Gaussian random matrix: for any $u \ge 0$, there exists an absolute constant $c > 0$ such that

$$\sigma_{\min}(\Xi_t) \ge \sigma_\Xi(\sqrt{m} - \sqrt{n} - u) \quad \text{with probability at least } 1 - 2\exp(-cu^2).$$

By choosing $u = (\sqrt{m} - \sqrt{n})/2$, we have the following probability bound:

$$\Pr(\mathcal{E}^\complement) = \Pr\left(\sigma_{\min}(\Xi_t) < \frac{\sigma_\Xi(\sqrt{m} - \sqrt{n})}{2}\right) \le 2\exp\left(-\frac{c(\sqrt{m} - \sqrt{n})^2}{4}\right).$$

After all, we can evaluate the expectation of $(\diamondsuit)$ by decomposing the event into $\mathcal{E}$ and $\mathcal{E}^\complement$ as follows:

$$\begin{aligned}\mathbb{E}\|\,\mathcal{O}(G_t) - \mathcal{O}(\Xi_t)\|_F &\le \frac{4\lambda}{\sigma_\Xi(\sqrt{m} - \sqrt{n})} \cdot \Pr(\mathcal{E}) + 2\sqrt{n} \cdot \Pr(\mathcal{E}^\complement) \\ &\le \frac{4\lambda}{\sigma_\Xi(\sqrt{m} - \sqrt{n})} + 4\sqrt{n}\exp\left(-\frac{c(\sqrt{m} - \sqrt{n})^2}{4}\right).\end{aligned}$$

Combining the above all of $(\clubsuit)$ and $(\diamondsuit)$, we have the desired inequality. □

E Additional Discussions

E.1 Sufficient Conditions for Signal Persistence

Assumption 2 bounds the momentum-filtered signed coordinate $\bar{\lambda}_k(t)$ from below by $c_{\mathrm{sig}}\lambda_k$ in absolute value. In this subsection, we demonstrate that Assumption 2 is indeed satisfied by a specific signed coordinate $\lambda_k(t)$.

Suppose $\lambda_k(t) = \bar{\mu}_k + \xi_k(t)$ with $\bar{\mu}_k \in \mathbb{R}$ and $\{\xi_k(t)\}_t$ independent zero-mean sub-Gaussian with parameter σ_λ^2. The momentum-filtered coordinate $\bar{\lambda}_k(t) = (1-\beta)\sum_{\tau=0}^t \beta^\tau \lambda_k(t-\tau)$ has mean $(1-\beta^{t+1})\bar{\mu}_k$, which equals $\bar{\mu}_k$ up to an $O(\beta^t)$ term once $t \ge c_0 T$, and sub-Gaussian parameter at most $\sigma_\lambda^2(1-\beta)^2\sum_{\tau\ge 0}\beta^{2\tau} = \sigma_\lambda^2/(2T-1)$ by the weight identity equation (14). Choosing c_0 so that $\beta^{c_0 T} \le 1/4$, the mean has magnitude at least $\frac{3}{4}|\bar{\mu}_k|$, so for every fixed $t \ge c_0 T$ a single sub-Gaussian tail bound on the fluctuation gives

$$\mathbb{P}\big(|\bar{\lambda}_k(t)| < |\bar{\mu}_k|/2\big) \le 2\exp\left(-\frac{\bar{\mu}_k^2(2T-1)}{32\sigma_\lambda^2}\right).$$

Taking a union bound over $k = 1, \dots, r$, we have

$$\mathbb{P}\big(\exists k \le r : |\bar{\lambda}_k(t)| < |\bar{\mu}_k|/2\big) \le 2r\exp\left(-\frac{\min_{k\le r}\bar{\mu}_k^2\,(2T-1)}{32\sigma_\lambda^2}\right).$$

Therefore, under the scenario $\min_k |\bar{\mu}_k| \gg \sigma_\lambda$, Assumption 2 holds simultaneously for all $k \le r$ with

$$c_{\mathrm{sig}} = \min_{k\le r}\frac{|\bar{\mu}_k|}{2\lambda_k}$$

on an event of overwhelming probability at every fixed $t \ge c_0 T$.

E.2 Sub-Gaussian Strengthening of Proposition 1

The Chebyshev-based proof of Proposition 1 as stated (under bounded Frobenius-norm second moment) appears in Appendix C. The result below is a *separate strengthening* of Proposition 1 under a stronger hypothesis: it assumes that each bilinear projection $x^\top \Xi_t y$ is centered sub-Gaussian, and concludes with an exponential tail rather than the Chebyshev $1/u^2$ tail of Proposition 1. We show this strengthening here for completeness. It is not used in the main theorem chain, which goes through the Chebyshev term only. Note that under the variance bound $\mathbb{E}\|\Xi_t\|_F^2 \le \eta$ of Assumption 1(b), the entrywise sub-Gaussian parameter v below corresponds to $\eta = mn\,v^2$.

Proposition 2 (Momentum-weighted bilinear sub-Gaussian concentration)**.** *Let $\{\Xi_t\}_{t\ge 0}$ be a sequence of independent $m \times n$ random matrices. Suppose that for every fixed unit vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, the bilinear projection $x^\top \Xi_t y$ is sub-Gaussian with parameter $v > 0$. Define*

$$S_t := (1-\beta)\sum_{s\ge 0} \beta^s \Xi_{t-s} \qquad \textit{and} \qquad T := \frac{1}{1-\beta}.$$

Then for every fixed unit vectors x, y and every $\varepsilon > 0$,

$$\mathbb{P}\big(|x^\top S_t y| > \varepsilon\big) \le 2\exp\left(-\frac{(2T-1)\varepsilon^2}{2v^2}\right).$$

An ε-net and union bound (e.g. [56, Theorem 4.4.5]) upgrade this to operator-norm concentration with an additional $m+n$ factor in the exponent.

Proof. Let $h_s := (1-\beta)\beta^s$, so that $S_t = \sum_{s\ge 0} h_s \Xi_{t-s}$. For fixed unit vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, set

$$z_s := x^\top \Xi_{t-s} y \qquad \text{and} \qquad W_L := \sum_{s=0}^{L} h_s z_s.$$

Since the z_s are independent and each is sub-Gaussian with parameter v, the moment generating function (MGF) of W_L factorizes over s. For all $\tau \in \mathbb{R}$,

$$\mathbb{E}[e^{\tau W_L}] = \prod_{s=0}^{L} \mathbb{E}[e^{\tau h_s z_s}] \le \prod_{s=0}^{L} \exp\left(\frac{\tau^2 h_s^2 v^2}{2}\right) = \exp\left(\frac{\tau^2 v^2}{2}\sum_{s=0}^{L} h_s^2\right) \overset{\text{equation (14)}}{\le} \exp\left(\frac{\tau^2 v^2}{2(2T-1)}\right).$$

A standard one-sided Chernoff bound followed by symmetrization (apply the same MGF bound to $-W_L$) gives the two-sided tail

$$\mathbb{P}\big(|x^\top S_t y| > \varepsilon\big) \le 2\exp\left(-\frac{(2T-1)\varepsilon^2}{2v^2}\right).$$

An ε-net and union bound then upgrade this to operator-norm concentration.

□

F Experimental Setups

F.1 Shared Computational Settings

The experiments in this paper run on two hardware environments. The end-to-end NanoGPT and LLaMA 350M [53] training comparison reported in figure 1 (Pre-polar, Post-polar, and Polar-only) runs on NVIDIA L40S GPUs. All remaining experiments, the synthetic simulations, the CIFAR-10 stationary and trajectory probes, and the NanoGPT stationary and trajectory probes, run on NVIDIA A100 40GB GPUs on the Wisteria Aquarius [1] system at the University of Tokyo.

The L40S experiments split into two end-to-end runs: a NanoGPT three-pipeline comparison on a single-node, 4-GPU PyTorch DDP setup with `bfloat16` autocast (Appendix F.4.4), and a LLaMA 350M three-pipeline comparison on a single-node, 8-GPU PyTorch DDP setup with `bfloat16`

[1] https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/

training (Appendix F.4.5). For the NanoGPT experiment, we use a GPT-2-style 12-layer / width-768 / 6-head modded-NanoGPT[2] speedrun model trained on FineWeb-10B [37]. For the LLaMA experiment, we used a reduced-scale language model that preserves the corresponding design choices: a LLaMA-style 24-layer / width-1024 / 16-head causal model with a 32k LLaMA-family tokenizer trained on C4 [42]. The shared L40S settings for the two studies are recorded in Appendices F.4.4 and F.4.5, and the best hyperparameter selections for each Muon pipeline are reported in the same appendices.

Table 1: Experimental settings for the NanoGPT three-pipeline end-to-end training comparison on NVIDIA L40S GPUs.

Setting	Value
Compute setup	4 NVIDIA L40S GPUs, single-node PyTorch DDP
Training precision	`bfloat16` autocast
Model family	GPT-2-style modded-NanoGPT speedrun causal language model
Model scale	12 layers, width 768, 6 heads, MLP width 3072
Core design choices	RMSNorm, RoPE, GPT-2 BPE (vocab 50304)
Data pipeline	FineWeb-10B pre-tokenized binary shards
Sequence length	1024
Batch shape	64 sequences/GPU, gradient accumulation 2, global batch 512
Training steps	5100
Learning-rate schedule	plateau–warmdown, 1450 warmdown steps
Regularization / clipping	weight decay 0, gradient clipping disabled

Table 2: Experimental settings for the LLaMA 350M three-pipeline end-to-end training comparison on NVIDIA L40S GPUs.

Setting	Value
Compute setup	8 NVIDIA L40S GPUs, single-node PyTorch DDP
Training precision	`bfloat16`
Model family	Reduced-scale LLaMA-style causal language model
Model scale	24 layers, width 1024, 16 heads, MLP width 2736
Core design choices	RMSNorm, 32k LLaMA-family tokenizer
Data pipeline	C4 dataset with document-level truncation and right-padding
Sequence length	1024
Batch shape	128 sequences/GPU, gradient accumulation 1, global batch 1024
Training steps	3000
Learning-rate schedule	cosine decay, 1000 warmup steps, minimum LR ratio 0.1
Regularization / clipping	weight decay 0, gradient clipping disabled

The Wisteria runs share a common compute environment. The synthetic simulation uses a single A100 GPU for its spiked-model simulations. The CIFAR-10 stationary and trajectory probes use a single A100. The NanoGPT probe jobs use four A100s under PyTorch DDP with `bfloat16` mixed precision. Spectrum summaries, polar factors, signal and subspace alignments are computed with exact `float32` SVDs inside the **Analysis procedure**, while training runs in `bfloat16`. Table 3 records the shared Wisteria settings.

F.2 Probe Protocols

The CIFAR-10 and NanoGPT experiments use the same two probe protocols on a designated target weight matrix $W \in \mathbb{R}^{m\times n}$. The two protocols differ only in how the per-step mini-batch gradients are collected. The downstream **Analysis procedure** is identical.

Let $G_t \in \mathbb{R}^{m\times n}$ denote the mini-batch gradient of W at step t. The probe records a gradient buffer $\{G_t\}_{t=1}^{K}$ of K size. Gradients are collected immediately after backpropagation and before any optimizer state update is applied. We used two distinct momentum coefficients throughout: β_{train} is the training-time Muon momentum coefficient (used by the optimizer that generates $\{G_t\}$), and β

[2] https://github.com/KellerJordan/modded-nanogpt

Table 3: Shared experimental settings for the Wisteria Aquarius runs (the synthetic simulations, the CIFAR-10 probes, and the NanoGPT probes).

Setting	Value
Cluster	Wisteria Aquarius
Compute GPUs	NVIDIA A100 40GB
Compiler / CUDA / Python	`gcc/8.3.1`, `cuda/11.8`, `python/3.11.7`
Training precision	`bfloat16` for NanoGPT, `float32` for CIFAR-10 and synthetic
Probe analysis precision	exact `float32` SVD for all spectrum and polar computations
Random seeds	1337, 1338, 1339

is the probe-side momentum coefficient that the probe analysis below applies to the gradient buffer. The two are decoupled by construction: the **Analysis procedure** operates solely on the collected raw gradient buffer $\{G_t\}$ and never accesses the training optimizer's internal state; at every probe step the three pipelines are recomputed from the same gradient buffer $\{G_t\}$. Pre-polar, Post-polar, and Polar-only comparisons are therefore fair comparisons on a common gradient stream, and training-time optimizer settings affect the results only through the trajectory that generates $\{G_t\}$. Next, we introduce the stationary probe and the trajectory probe.

Stationary probe.

1. Load the model from a saved checkpoint and hold every weight fixed, including the target weight W. No weight is updated during gradient collection.
2. For $t = 1, \ldots, K$, draw one mini-batch from the dataloader in its natural sequential order (the dataloader's default ordering without shuffling), run one step of forward and backward propagation, and record the gradient G_t of the target weight W. We refer to this as the *sequential* collection order, which is the default setting used throughout the paper. For the NanoGPT stationary probe at the step-3000 checkpoint, we additionally reran the same protocol under a *shuffled* collection order as a robustness check, where each mini-batch starts at a random position in the current data shard. The shuffled results are reported only in Appendix G.
3. Save the gradient sequence $\{G_t\}_{t=1}^{K}$ to disk in the order the mini-batches were drawn. The collection order is preserved because the downstream momentum buffers are order-dependent.

Since the model weights do not change during gradient collection, every G_t is drawn from the same gradient distribution. This protocol therefore synthetically simulates the stationary special case of the BVMZOS perturbation model of Assumption 1. The three pipelines are applied to the gradient buffer in its collection order, so Pre-polar and Post-polar momentum buffers see the gradient stream $\{G_t\}_{t=1}^{K}$ exactly as it was recorded. For the stationary probe at the step-3000 checkpoint we report both the sequential and shuffled collection orders. All other stationary checkpoints use the sequential order only.

Trajectory probe. During end-to-end training:

1. Maintain a fixed-capacity First-In, First-Out (FIFO) buffer of the most recent K target weight's gradients alongside the regular optimizer step. Each training step appends G_t to the buffer and pops the oldest entry once the buffer is full.
2. At every I training step (we used $I = 100$ in all CIFAR-10 and NanoGPT trajectory runs), take a checkpoint of the current buffer $\{G_{t-K+1}, \ldots, G_t\}$, run the **Analysis procedure** below, and save the resulting summary.
3. Continue training without modification. The probe does not alter weight updates, optimizer state, random seeds, or data ordering.

The trajectory buffer therefore represents a sliding buffer over the live training trajectory, and the corresponding analysis quantities inherit any non-stationarity in the gradient stream.

Buffer-size selection. The buffer size K obeys different constraints in the two protocols. With zero initialization, the momentum buffer (equation (4)) truncated to K steps carries kernel mass $1 - \beta^K$

(see Appendix B.3), so $K \geq c_0 T$ keeps the momentum-filtered buffers (e.g., $M_K^{(\beta)}$ in equation (18)) close to their asymptotic values and allows the in-buffer mean gradient $\bar{G}$ to serve as the reference of the coherent gradient signal.

In the stationary probe, the weights are held fixed during gradient collection, so the gradient distribution of G_t does not drift across the K steps. The only constraint on K is the lower bound $K \geq c_0 T$ at the largest β in the per-task grid (Appendix F.4). A larger K is therefore preferable.

In the trajectory probe, the weights are updated by the optimizer across the K steps, so the gradient distribution of G_t drifts with t. The lower bound $K \geq c_0 T$ still applies, and K must also remain small enough to limit this drift across the steps contributing to the in-buffer mean gradient $\bar{G}$ (defined in the **Analysis procedure** below). The choice of K therefore balances these two requirements.

Analysis procedure. Given a buffer $\{G_t\}_{t=1}^K$, the analysis performs the following steps:

1. *Signal reference.* Compute the mean gradient $\bar{G} := K^{-1} \sum_{t=1}^K G_t$ and its exact SVD $\bar{G} = U \Sigma V^\top$. The top-$r$ left and right singular vectors (U_r, V_r) define the signal subspace used as the alignment target on the CIFAR-10 and NanoGPT experiments. On the synthetic simulation the spiked model of Appendix F.4.1 plants a known rank-$r_\star$ signal G_t^{sig} with singular bases $U_{\mathrm{true}}, V_{\mathrm{true}}$, so the alignment target is the planted top-r subspace $(U_{\mathrm{true}}, V_{\mathrm{true}})$ directly rather than (U_r, V_r) from $\bar{G}$ (Appendix F.3).
2. *Three pipelines.* For each β in the per-task grid (Appendix F.4), starting from $M_0 = \widetilde{M}_0 = 0$, Pre-polar and Post-polar pipelines maintain two separate momentum buffers:

$$\text{Pre-polar:} \quad \mathcal{O}\big(M_K^{(\beta)}\big), \quad \text{where} \quad M_K^{(\beta)} := (1-\beta) \sum_{s=0}^{K-1} \beta^s \, G_{K-s}, \tag{18}$$

$$\text{Post-polar:} \quad \widetilde{M}_K^{(\beta)} := (1-\beta) \sum_{s=0}^{K-1} \beta^s \, \mathcal{O}(G_{K-s}), \tag{19}$$

$$\text{Polar-only:} \quad \mathcal{O}(G_K), \tag{20}$$

   where $\mathcal{O}(\cdot)$ is the polar factor introduced in equation (1). Pre-polar buffer $M_K^{(\beta)}$ averages the raw gradients and is orthogonalized once at the end, whereas Post-polar buffer $\widetilde{M}_K^{(\beta)}$ orthogonalizes each per-step gradient first and then averages the resulting polar factors. Polar-only baseline skips the momentum entirely and orthogonalizes only the final raw gradient G_K.
3. *Spectral summaries.* Record (i) the singular-value sequences of $\bar{G}$ and of Pre-polar momentum buffer $M_K^{(\beta)}$, (ii) the per-step filtering ratio $\sigma_k(M_K^{(\beta)})/\sigma_k(G_K)$ at the final collection index, and (iii) the noise-suppression ratio $R(T)$ that compares the operator norm of the raw-gradient residual with that of the momentum residual at momentum window size $T = 1/(1-\beta)$ (introduced in Appendix F.3). The two ratios serve different purposes and use different denominators. The per-step filtering ratio measures the per-singular-value attenuation between $M_K^{(\beta)}$ and G_K. $R(T)$ uses a signal reference in both numerator and denominator and measures the operator-norm attenuation of the momentum-filtered perturbation. The signal reference is $\bar{G}$ on the CIFAR-10 and NanoGPT experiments and the planted signal G_t^{sig} on the synthetic simulation. The synthetic case additionally applies the bias correction $(1-\beta^K)\, G_t^{\mathrm{sig}}$ in the denominator to account for the zero-initialized momentum buffer (Appendix F.3).
4. *Signal alignment and subspace alignment error metrics.* Signal alignment is reported with the rank-r and full-rank signal alignment metrics Align_r and $\mathrm{Align}_{\mathrm{full}}$ of Appendix F.3. Subspace alignment error panels report the $\sin\Theta$ principal-angle distance at fixed ranks $r \in \{1,5,10\}$.

All SVDs inside the **Analysis procedure** are exact `float32` decompositions. Newton–Schulz iteration is used only inside the training-time Muon optimizer and never inside the probe.

F.3 Measurements

Every probe run produces the same four measurements from the gradient buffer $\{G_t\}_{t=1}^K$ and the three pipelines of Appendix F.2. Each measurement is tied to a specific theoretical claim:

1. *Per-step filtering ratio* — the ratio of the k-th singular value of Pre-polar momentum buffer $M_K^{(\beta)}$ (equation (18)) to that of the latest collected raw gradient G_K,

$$\mathrm{Filt}_k(\beta) \;\coloneqq\; \frac{\sigma_k\big(M_K^{(\beta)}\big)}{\sigma_k(G_K)}.$$

   Both spectra are computed from the same gradient buffer at the final collection step: G_K is the last raw mini-batch gradient (equivalently, the momentum buffer at $\beta = 0$), and $M_K^{(\beta)}$ is the momentum over the same window. By construction $\mathrm{Filt}_k(0) = 1$. Theorem 1 predicts that when β is close enough to 1, the tail filtering ratios ($k > r$) are suppressed more than the head ($k \le r$), opening a spectral gap between head and tail that widens with β. The per-step filtering ratio supports Theorem 1 qualitatively.

2. *Noise-suppression ratio* — residual operator-norm ratio $R(T)$ of raw gradient vs. Pre-polar momentum buffer with probe-side momentum coefficient β (associated with the effective sample size $2T-1$). Explicitly,

$$R(T) \;\coloneqq\; \frac{\big\|G_K - \bar{G}\big\|_2}{\big\|M_K^{(\beta)} - \bar{G}\big\|_2}, \qquad \bar{G} \;\coloneqq\; \frac{1}{K}\sum_{t=1}^{K} G_t,$$

   where $\bar{G}$ is the in-buffer approximation of G_t^{sig}. Subtracting $\bar{G}$ from both numerator and denominator approximately removes the coherent signal component G_t^{sig} from $G_t = G_t^{\mathrm{sig}} + \Xi_t$, so the numerator $\|G_K - \bar{G}\|_2 \approx \|\Xi_K\|_2$ approximates the single-step perturbation operator norm, while the denominator $\|M_K^{(\beta)} - \bar{G}\|_2 \approx \|S_K\|_2$ approximates the operator norm of the momentum-filtered perturbation S_K from Proposition 1. Their ratio $R(T)$ therefore measures how much the momentum reduces the perturbation operator norm relative to a single step (see Remark 4 for the derivation). The figure conventions plot $R(T)$ against $T = 1/(1-\beta)$, with the theoretical prediction $(2T-1)^{1/4}$ (derived in Remark 4) shown as the dashed guide. The choice of $\bar{G}$ as the signal reference is what distinguishes $R(T)$ from the per-step filtering ratio: $\mathrm{Filt}_k(\beta)$ uses the latest single-batch gradient G_K as the denominator and reports per-index attenuation, while $R(T)$ uses the $\bar{G}$ as the reference and reports the perturbation operator-norm reduction relative to a single step, compared against the $(2T-1)^{1/4}$ guide.[3] The noise-suppression ratio is therefore an empirical check that $R(T)$ grows with the momentum window size T at a rate no slower than $(2T-1)^{1/4}$ with high probability, as predicted by Proposition 1 and Theorem 1.

3. *Subspace alignment error* — the $\sin\theta$ subspace distance from the top-r singular subspaces of the reference to those of Pre-polar momentum buffer. Explicitly,

$$\sin\theta_r(A;\,B) \;\coloneqq\; \big\|\sin\Theta\big(U_r(A),\,U_r(B)\big)\big\|_2,$$

   with the right-subspace version $\sin\theta_r(A^\top;\,B^\top) = \|\sin\Theta(V_r(A),\,V_r(B))\|_2$ reported separately. In this paper, we also define $\sin\Theta_U \coloneqq \sin\theta_r(M_K^{(\beta)};\,\bar{G})$ and $\sin\Theta_V \coloneqq \sin\theta_r(M_K^{(\beta)\top};\,\bar{G}^\top)$ for convenience. The gradient signal reference is $\bar{G}$ on the CIFAR-10 and NanoGPT experiments at rank $r \in \{1{,}5{,}10\}$.[4] The subspace alignment error supports Corollary 1, which predicts that $\sin\Theta_U$ and $\sin\Theta_V$ decrease as the momentum window size T grows at a rate no slower than $(2T-1)^{1/4}$ with high probability.

4. *Signal alignment* — the signal alignment comparison applied to Pre-polar $= \mathcal{O}(M_K^{(\beta)})$, Post-polar $= \widetilde{M}_K^{(\beta)}$, and Polar-only $= \mathcal{O}(G_K)$, reported through the following two metrics:
   Rank-r signal alignment.

$$\mathrm{Align}_r(A;\,B) \;\coloneqq\; \frac{\big\|U_r(B)^\top A\, V_r(B)\big\|_F}{\sqrt{r}} \;\in\; [0,1].$$

[3] On the synthetic simulation the planted signal G_t^{sig} is known exactly, so we replace $\bar{G}$ with G_t^{sig} and apply a zero-init bias correction to the momentum buffer: the synthetic $R(T)$ is $\|G_K - G_t^{\mathrm{sig}}\|_2 / \|M_K^{(\beta)} - (1-\beta^K)\,G_t^{\mathrm{sig}}\|_2$. The bias factor $(1-\beta^K)$ accounts for the momentum buffer starting from zero, which makes $\mathbb{E}[M_K^{(\beta)}] = (1-\beta^K)\,G_t^{\mathrm{sig}}$ under $G_t = G_t^{\mathrm{sig}} + \Xi_t$. Without this correction the synthetic denominator would be inflated by an $O(\beta^K)$ transient at large β.

[4] On the synthetic simulation the reference is the planted top-r singular subspace $(U_{\mathrm{true}}, V_{\mathrm{true}})$ at $r \in \{1,2,3\}$. Restricting to $r \le r_\star$ is forced by the model: the planted bases have only $r_\star$ orthonormal columns, so a query at $r > r_\star$ is either ill-defined or pads with arbitrary directions from the noise null space.

Larger values indicate stronger signal alignment. Theorem 2 predicts that Pre-polar achieves higher Align_r than Post-polar and Polar-only for sufficiently large momentum window size T. On the CIFAR-10 and NanoGPT experiments the reference is $B = \bar{G}$ and the rank ladder is $r \in \{1,5,10\}$.[5]

Full-rank signal alignment.

$$\mathrm{Align}_{\mathrm{full}}(A;\, B) \;\coloneqq\; \frac{\langle A, \mathcal{O}(B)\rangle_F}{\min(m,n)} \;\in\; [-1,1].$$

We report $\mathrm{Align}_{\mathrm{full}}$ as a full-rank cross-check on the CIFAR-10 and NanoGPT experiments only.[6] Larger values indicate stronger full-rank alignment. Theorem 2 predicts that Pre-polar achieves higher $\mathrm{Align}_{\mathrm{full}}$ than Post-polar and Polar-only for sufficiently large momentum window size T.

The probe-side momentum coefficient grids are measurement- and task-specific and are listed with the corresponding task definitions in Appendix F.4. The default probe-side momentum coefficient $\beta = 0.95$ is shared across all experiments.

Remark 4 (Predicted $(2T-1)^{1/4}$ guide on $R(T)$)**.** *The dashed $(2T-1)^{1/4}$ guide in $R(T)$ panels comes from inverting the operator-norm corollary of Proposition 1.*

Bridge. *Substituting $G_t = G_t^{\mathrm{sig}} + \Xi_t$ and writing $\bar{G}^{\mathrm{sig}} \coloneqq K^{-1}\sum_t G_t^{\mathrm{sig}}$, $\bar{\Xi} \coloneqq K^{-1}\sum_t \Xi_t$,*

$$G_K - \bar{G} \;=\; (G_K^{\mathrm{sig}} - \bar{G}^{\mathrm{sig}}) + (\Xi_K - \bar{\Xi}), \quad M_K^{(\beta)} - \bar{G} \;=\; (M_K^{\mathrm{sig}} - \bar{G}^{\mathrm{sig}}) + (S_K - \bar{\Xi}).$$

The BVMZOS structure gives $\mathbb{E}\left\|\bar{\Xi}\right\|_F^2 \le \eta/K$, while Proposition 1 gives $\mathbb{E}\,\|S_K\|_F^2 \le \eta/(2T-1)$. Since $\|\cdot\|_2 \le \|\cdot\|_F$, the typical operator norms of $\bar{\Xi}$ and S_K scale as $\sqrt{\eta/K}$ and $\sqrt{\eta/(2T-1)}$, so $\left\|\bar{\Xi}\right\|_2$ is negligible against $\|S_K\|_2$ once $K \gg 2T-1$. The signal residuals $G_K^{\mathrm{sig}} - \bar{G}^{\mathrm{sig}}$ and $M_K^{\mathrm{sig}} - \bar{G}^{\mathrm{sig}}$ vanish exactly when $\lambda_k(t)$ is time-invariant, and are dominated by their perturbation counterparts in our experiments otherwise. Hence

$$\left\|M_K^{(\beta)} - \bar{G}\right\|_2 \approx \|S_K\|_2\,, \qquad \left\|G_K - \bar{G}\right\|_2 \approx \|\Xi_K\|_2\,.$$

Lower bound. *Setting $u = (2T-1)^{1/4}$ in Proposition 1 gives*

$$\|S_K\|_2 \;\le\; \sqrt{\eta}\,(2T-1)^{-1/4} \qquad \textit{with probability at least} \quad 1-(2T-1)^{-1/2}.$$

The numerator $\|\Xi_K\|_2$ is T-independent under Assumption 1(b). Dividing,

$$R(T) \;\gtrsim\; (2T-1)^{1/4}$$

on the same event, up to a noise-distribution-dependent constant.

Sharper rate under sub-Gaussian projections. *Under the bilinear sub-Gaussian hypothesis of Proposition 2 (Appendix E.2), the same bridge yields the sharper $R(T) \gtrsim \sqrt{2T-1}$, attained by the synthetic Gaussian curves of figure 9a. The $(2T-1)^{1/4}$ floor uses only the second-moment bound of Assumption 1(b) and therefore remains valid under heavy-tailed gradient noise with finite variance.*

F.4 Tasks

We describe the five experimental tasks below, organized into probe experiments (Appendices F.4.1 to F.4.3), where the **Analysis procedure** is applied to gradients collected from a fixed or live model, and end-to-end training runs (Appendices F.4.4 and F.4.5), where Pre-polar, Post-polar, and Polar-only are compared as full training optimizers.

F.4.1 Synthetic Simulation

Signal Generator. The synthetic simulation uses a rank-3 spiked model: a time-invariant low-rank signal plus a perturbation satisfying Assumption 1(b).

[5] On the synthetic simulation $B = U_{\mathrm{true}}V_{\mathrm{true}}^\top$ and the rank ladder is $r \in \{1,2,3\}$.

[6] The synthetic simulation drops it because the relevant signal reference is the planted rank-$r_\star$ partial polar factor $U_{\mathrm{true}}V_{\mathrm{true}}^\top$, which has Frobenius norm $\sqrt{r_\star}$ and caps the corresponding normalized full-rank signal alignment at $r_\star/\min(m,n) = 0.03$ on our 100×100 synthetic simulation.

Table 4: Synthetic task summary.

Setting	Value
Matrix size	100×100
Signal model	low-rank spike + BVMZOS perturbation

Perturbation generator. All synthetic simulations use the perturbation $\Xi_t = b\,\epsilon\,B_t + Z_t$, where $\epsilon \in \{-1, +1\}$ is a single Rademacher sign drawn once per trajectory, $\{B_t\}$ are deterministic dense matrices with $\|B_t\|_F = 1$ and $\langle B_s, B_t\rangle_F = 0$ for $s \neq t$, Z_t are i.i.d. isotropic noise with entrywise variance $\sigma_n^2/2$ (Gaussian, or rescaled finite-variance Student-t with four degrees of freedom for the heavy-tailed panel of figure 7), and $b = \sigma_n\sqrt{mn/2}$. This satisfies Assumption 1(b): mean zero $\mathbb{E}[\Xi_t] = 0$, bounded variance $\mathbb{E}\,\|\Xi_t\|_F^2 = \sigma_n^2 mn$, and Frobenius orthogonality $\mathbb{E}\langle \Xi_s, \Xi_t\rangle_F = b^2\langle B_s, B_t\rangle_F = 0$ for $s \neq t$.

Measurement scope. We report the same four measurements as the CIFAR-10 and NanoGPT probes (Appendix F.3). The synthetic stationary probe uses $K = 1000$ collected gradients per trial over 10 random-seed trials, and the signal-strength sweep (Appendix F.5) uses $K = 500$.

Probe-side β grids. For each of the following probes, we used different β grids. We use a six-point grid {0.3,0.6,0.9,0.95,0.97,0.99} to report the filtered singular value spectrum (Figure 7). We use a fourteen-point grid {0.05,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95,0.97,0.99,0.995} for the noise-suppression ratio (Figure 9a) and the subspace alignment error (Figure 13). The Pre-polar / Post-polar / Polar-only β-sweep (Figure 4a) uses the same fourteen-point grid. We use the eight-point grid {0.5,0.7,0.8,0.9,0.93,0.95,0.97,0.99} for the all-layer 3×4 stationary grid (Figure 21) and the per-layer / per-checkpoint numerical summaries (Tables 7 to 9). For simulations that do not sweep β, we use $\beta = 0.95$ as the default probe-side momentum coefficient.

F.4.2 CIFAR-10 Probe

The target weight matrix to be probed is the convolution layer `layer2.0.conv1` of a ResNet-18, whose weight tensor is matricized to shape 128×576. The stationary and trajectory probes follow the general protocol of Appendix F.2.

Stationary probe. The CIFAR-10 stationary probe collects mini-batch gradients after warmup on `layer2.0.conv1` from a ResNet-18 trained on CIFAR-10. In the language of Assumption 1, the signal proxy is the gradient mean over the collection buffer after warmup and the perturbation proxy is the residual around that mean. K is chosen large enough to fully warm the momentum buffer at the largest probe-side $\beta = 0.99$ ($T = 100$). At $K = 2000$, the kernel mass $1 - \beta^K$ is within 2×10^{-9} of one per Appendix B.3.

Trajectory probe. The CIFAR-10 trajectory probe maintains a sliding buffer of recent target-layer gradients during continued training after the warmup phase. The analysis interval, stored ordering metric, and direction-cut ranks follow the cross-experiment conventions of Appendices F.2 and F.3. We use $K = 100$ (The sensitivity analysis with respect to K is provided in Appendix K).

Probe-side β grids. The four CIFAR-10 measurements use the following grids. We use a six-point grid {0.3,0.6,0.9,0.95,0.97,0.99} for the stationary per-step filtering ratio (Figure 8). We use a fourteen-point grid {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.85,0.9,0.93,0.95,0.97,0.99} for the stationary noise-suppression ratio (Figure 9b) and the stationary subspace alignment error (Figure 14). We use an eight-point grid {0.5,0.7,0.8,0.9,0.93,0.95,0.97,0.99} for the stationary signal alignment ordering (Figure 20). On the trajectory probe at $K = 100$, we use the fourteen-point grid {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.85,0.9,0.93,0.95,0.97,0.99} for the noise-suppression ratio and the subspace alignment error (Figure 16), and the eight-point grid {0.5,0.7,0.8,0.9,0.93,0.95,0.97,0.99} for the ordering panel (Figure 22). For experiments that do not sweep β, we use $\beta = 0.95$ as the default probe-side momentum coefficient.

F.4.3 NanoGPT Probe

The NanoGPT training code is adapted from modded-NanoGPT. All NanoGPT probe runs, the stationary probe across three representative attention layers, and the trajectory probe across every

attention and MLP output projection, share a single training configuration, described below, except for the L40S end-to-end run configuration shown in Appendix F.4.4 separately.

Training. The shared NanoGPT training configuration is a 12-layer, 768-width GPT-2 style model with 6 heads, rotary position embeddings (RoPE) [50], RMSNorm, scaled the squared-ReLU activations, FineWeb-10B training data, global batch size 512 (device batch size 64, sequence length 1024), zero warmup, 1450 warmdown iterations, and a 5100-step budget. The Muon optimizer runs with momentum coefficient $\beta_{\text{train}} = 0.8$, Nesterov momentum enabled, learning rate 0.05, and five Newton–Schulz steps per update.

The trajectory probe is applied to all 24 output projections of the model: the 12 attention output projections `h.{0,...,11}.attn.c_proj` (each 768×768) and the 12 MLP output projections `h.{0,...,11}.mlp.c_proj` (each 768×3072), all with buffer size $K = 50$ and analysis every 100 training steps. The stationary probe additionally covers three representative depths `h.0.attn.c_proj`, `h.5.attn.c_proj`, and `h.11.attn.c_proj` at the saved checkpoints {1000, 2000, 3000, 4000, 5000}.

Stationary probe. The NanoGPT stationary probe reuses stored checkpoints at training steps 1000, 2000, 3000, 4000, and 5000 and applies the protocol of Appendix F.2 to three representative attention output projections: `h.0.attn.c_proj`, `h.5.attn.c_proj`, and `h.11.attn.c_proj`. The default protocol uses the sequential collection order of Appendix F.2, with collection buffer of $K = 500$ and $K = 1000$ gradients. The stationary probe holds the weights fixed during collection, so K is chosen large enough to warm the momentum buffer at the largest probe-side $\beta = 0.99$ ($T = 100$). At the step-3000 checkpoint we additionally rerun the same two windows under the shuffled collection order as the robustness cross-check of Appendix G. Figure 4 uses `h.0` as the representative slice, while the appendix keeps `h.5` and `h.11` as cross-layer checks.

Trajectory probe. The NanoGPT trajectory probe is reported at two scopes: representative-layer evidence on a single attention output projection (Section 5), and all-layer evidence spanning every attention and MLP output projection (Appendix I). Both scopes share the same 3-seed run group, buffer size $K = 50$, analysis interval, and probe-side β grids, differing only in the target layer. The representative-layer slice uses `h.0.attn.c_proj`, matching the stationary-probe representative layer (Figure 3). Appendix I aggregates over all 24 projection targets. The K-sweep in Appendix K verifies that Pre-polar dominance over Post-polar and Polar-only, and the Pre-polar to Polar-only alignment ratio, hold consistently for K ranging from $2.5T$ to $10T$.

The probe stores per-step filtered singular-value spectra, $\sin\Theta$ subspace alignment errors at ranks $r \in \{1,5,10\}$, and Pre-polar / Post-polar / Polar-only full-rank alignment values. Three training seeds {1337, 1338, 1339} are run for both instances. We report 3-seed mean and sample standard deviation.

Probe-side β grids. The NanoGPT panels use the following grids. We use a six-point grid {0.3,0.6,0.9,0.95,0.97,0.99} for the stationary per-step filtering ratio (Figure 2b). We use a fourteen-point grid {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.85,0.9,0.93,0.95,0.97,0.99} for the stationary noise-suppression ratio (Figure 2c) and the stationary subspace alignment error (Figure 3). We use an eight-point grid {0.5,0.7,0.8,0.9,0.93,0.95,0.97,0.99} for the stationary Pre-polar / Post-polar / Polar-only signal-alignment comparison (Figure 4a). For the trajectory subspace alignment error (Figure 5) we use a twelve-point grid {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.85,0.9,0.93,0.95}, and for the trajectory Pre-polar / Post-polar / Polar-only signal-alignment comparison (Figure 6a) the six-point grid {0.5,0.7,0.8,0.9,0.93,0.95}. We display the full fourteen-point grid for the all-layer trajectory subspace alignment error (Figures 17 and 18). For experiments that do not sweep β, we use $\beta = 0.95$ as the default probe-side momentum coefficient.

F.4.4 NanoGPT End-to-End Training

The NanoGPT experiment is the first of the two end-to-end comparisons in figure 1. The shared L40S training recipe (architecture, data, batch shape, schedule, precision, DDP layout, training steps) is given in table 1. Hidden two-dimensional matrix parameters in the transformer blocks are updated by the selected Muon pipeline (with five Newton–Schulz iterations per update), while token embeddings, the language modeling head, and scalar tensors are updated by an auxiliary fused AdamW optimizer with learning rate 0.008, betas (0.8, 0.95), and $\epsilon = 10^{-10}$. Pre-polar and Post-polar additionally use the Muon momentum warmup that linearly interpolates from a fixed initial momentum 0.85 to the pipeline's best β over the first 300 steps and enable Nesterov

momentum. Polar-only carries no momentum buffer, skips the momentum warmup, and turns off Nesterov momentum. The pipeline-specific best Muon learning rate and β are selected by the sweep below.

Hyperparameter sweep and best configurations. Before the seed-averaging stage, we ran a single-seed (`seed=0`) sweep at the full 5100-step training budget to pick the best Muon learning rate and momentum coefficient per pipeline. The main sweep covers three Muon learning rates $\{0.025, 0.05, 0.075\}$ crossed with three momentum coefficients $\{0.8,\ 0.9, 0.95\}$ for Pre-polar and Post-polar pipelines (nine (lr, β) cells per pipeline), and six learning rates $\{0.01, 0.015, 0.02, 0.025, 0.05, 0.075\}$ for Polar-only pipeline. The selected configurations reported in table 5 are those that minimize the FineWeb-10B validation loss at step 5100 within the swept grid. The selected configurations were then rerun with three independent seeds (1-3) under identical settings, and the per-step three-seed mean and sample standard deviation feed the NanoGPT panel of figure 1.

Table 5: NanoGPT end-to-end best configurations selected from the single-seed (`seed=0`) sweep and used for the three-seed final comparison.

Variant	Pipeline	Muon LR	β	Seeds
Pre-polar	momentum $\rightarrow$ NS	0.025	0.8	3
Post-polar	NS $\rightarrow$ momentum	0.05	0.95	3
Polar-only	NS only	0.02	–	3

F.4.5 LLaMA 350M End-to-End Training

The LLaMA 350M experiment is the second of the two end-to-end comparisons in figure 1. The shared L40S training recipe (architecture, data, batch shape, schedule, precision, DDP layout, training steps) is given in table 2. Hidden two-dimensional matrix parameters in the transformer blocks are updated by the selected Muon pipeline (with five Newton–Schulz iterations per update), while embeddings, the output head, normalization parameters, and other non-Muon tensors are updated by an auxiliary AdamW optimizer with learning rate 3×10^{-4}, betas (0.9,0.95), and $\epsilon = 10^{-10}$. Pre-polar and Post-polar hold the Muon momentum coefficient constant at the pipeline's best β throughout training and enable Nesterov momentum. No momentum warmup is applied here, in contrast to the NanoGPT recipe of Appendix F.4.4. Polar-only carries no momentum buffer and turns off Nesterov momentum. Validation cross-entropy is computed every 500 steps over $\sim$10M C4 tokens, and the reported final loss / perplexity in table 6 is the value at step 3000. The pipeline-specific best Muon learning rate and β are selected by the sweep below.

Hyperparameter sweep and best configurations. Before the seed-averaging stage, we ran a single-seed (`seed=0`) sweep at the full 3000-step training budget to pick the best Muon learning rate and momentum coefficient per pipeline. The main sweep covers six Muon learning rates $\{0.005, 0.01, 0.02, 0.025, 0.03, 0.035\}$ crossed with four momentum coefficients $\{0.8, 0.9, 0.95, 0.975\}$ for Pre-polar and Post-polar pipelines, and the same six learning rates for Polar-only pipeline. The selected configurations reported in table 6 are those that minimize the C4 validation loss at step 3000 within the swept grid. The selected configurations were then rerun with three independent seeds (1-3). The per-step three-seed mean and sample standard deviation feed the LLaMA 350M panel of figure 1.

Table 6: LLaMA 350M end-to-end best configurations selected from the single-seed (`seed=0`) sweep and seed-averaged final validation loss / perplexity at step 3000. Final loss and perplexity are reported as mean $\pm$ sample standard deviation over completed runs.

Variant	Pipeline	Muon LR	β	Seeds	Final loss / PPL
Pre-polar	momentum $\rightarrow$ NS	0.020	0.95	3	2.9696 ± 0.0002 / 19.48
Post-polar	NS $\rightarrow$ momentum	0.020	0.95	3	3.1568 ± 0.0058 / 23.50
Polar-only	NS only	0.035	–	3	3.2402 ± 0.0076 / 25.54

F.5 Signal-Strength Sensitivity

The signal-strength sweep of Appendix J reuses the synthetic, CIFAR-10, and NanoGPT probe protocols defined above with three task-specific configurations. All three sweeps share the probe-side momentum coefficient $\beta = 0.95$ and the single-pass momentum over the stored collection order described in Appendix F.2. The metric definitions are those of Appendix F.3.

Synthetic λ sweep (Figure 26). The synthetic generator is the rank-3 spiked model shared with figure 19: $m = n = 100$, planted singular values $\lambda \cdot [1, 8/12, 5/12]$ (so that λ acts directly as the leading signal singular value), $\sigma_n = 1$, BVMZOS perturbation, $K = 500$ collection steps per trial, and 10 random-seed trials. The signal subspace bases (U, V) are random orthogonal matrices fixed across λ for reproducibility (seed 123). The reported metric is rank-3 subspace alignment against the planted (U, V) subspaces, plotted with $\pm 1\sigma$ trial-variance bands. The swept grid is $\lambda \in \{0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, 10.0, 12.0, 15.0, 18.0, 20.0\}$.

CIFAR-10 batch sweep (Figure 27). The probe target weight matrix is layer `layer2.0.conv1` of the same ResNet-18 used in the default CIFAR-10 stationary probe. The network is warmed up for 500 steps, then frozen. Six independent probes are run with mini-batch sizes $b \in \{16, 32, 64, 128, 256, 512\}$, each collecting $K = 200$ stationary mini-batch gradients on the same frozen network. Per-batch the K gradients are processed through the three pipelines at $\beta = 0.95$ in collection order. The four reported panels are rank-1, rank-5, rank-10 subspace alignment against the empirical mean-gradient subspace, and full-rank signal alignment.

NanoGPT batch sweep (Figure 28). The probe target weight matrix is layer `h.0.attn.c_proj` (matrix shape 768×768) of the NanoGPT step-3000 checkpoint. Six independent probes are run with mini-batch sizes $b \in \{16, 32, 64, 128, 256, 512\}$ at sequence length 1024, each collecting $K = 500$ stationary mini-batch gradients with single-step accumulation. Per-batch the K gradients are processed through the three pipelines at $\beta = 0.95$ in collection order. The four reported panels are rank-1 subspace alignment, rank-5 and rank-10 subspace alignment, and full-rank signal alignment.

G Experimental Results of Spectral Gaps

This section extends the validation of Theorem 1 from the NanoGPT probe of Section 3 to synthetic and CIFAR-10 stationary probes and to the remaining NanoGPT attention output projections and saved checkpoints. The full settings are shown in Appendices F.4.1 to F.4.3. All panels follow the curve, marker, and reference conventions of figure 2. Across all panels, the same qualitative pattern repeats: the head of the momentum buffer's singular value spectrum tracks the signal singular values, the tail collapses toward the perturbation floor $(2T-1)^{-1/4}$ of Theorem 1 as β grows, and the noise-suppression ratio $R(T)$ sits above the $(2T-1)^{1/4}$ rate predicted by Theorem 1.

Synthetic and CIFAR-10 Stationary Probes. Figure 7 shows the filtered singular value spectrum on the rank-3 spiked model under Gaussian-noise (left panel) and heavy-tailed Student-t-noise (right panel) BVMZOS perturbations, with planted singular values $\sigma_1 = 12$, $\sigma_2 = 8$, and $\sigma_3 = 5$ (black diamonds at $k = 1, 2, 3$). As β grows, the top-3 singular values of the momentum buffer concentrate near the true signal, while singular values at indices $k \geq 4$ are more suppressed, supporting Theorem 1. Figure 8 reproduces this pattern on the CIFAR-10 stationary probe at `layer2.0.conv1` of a ResNet-18 (matricized to 128×576), warmup step 500, $K = 2000$, with the filtered singular value spectra in figure 8a and the per-step filtering ratio $\mathrm{Filt}_k(\beta)$ (Appendix F.3) in figure 8b. As β grows, the per-step filtering ratios decrease, with the head indices suppressed less than the tail. Figure 9 shows the noise-suppression ratio $R(T)$ on synthetic and CIFAR-10 stationary probe experiments. On synthetic (Figure 9a), where the planted G_t^{sig} replaces $\bar{G}$ in the denominator with a zero-init bias correction (Appendix F.3), $R(T)$ grows above the $(2T-1)^{1/4}$ rate under the BVMZOS perturbation. On CIFAR-10 (Figure 9b), the empirical curve grows at the similar rate above the same floor across the full β sweep, supporting Theorem 1.

NanoGPT Stationary Probes Across Layers and Checkpoints. Figures 10 and 11 extend figures 2a and 2b, respectively, to a 3×5 grid over three attention output projections (`h.0`, `h.5`, `h.11`) as rows and five saved training checkpoints (steps 1000, 2000, 3000, 4000, and 5000) as columns, $K = 500$ per cell. At every cell, as β grows, the buffer spectrum moves toward the mean-gradient

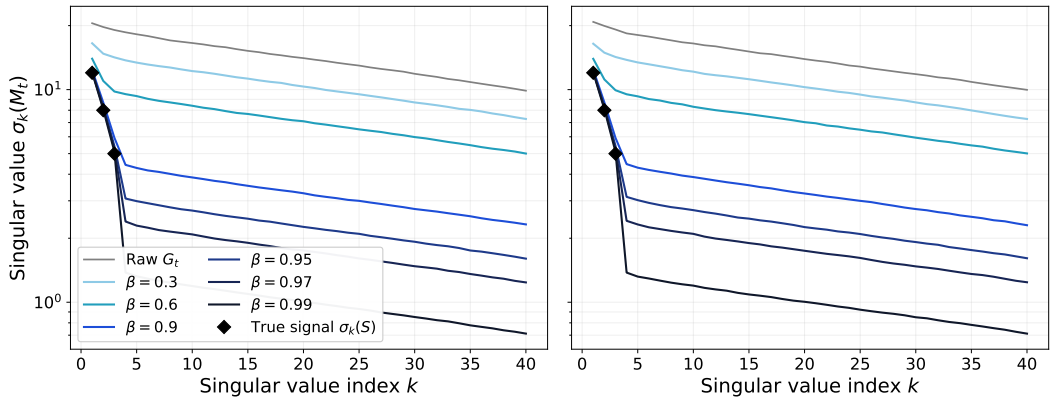
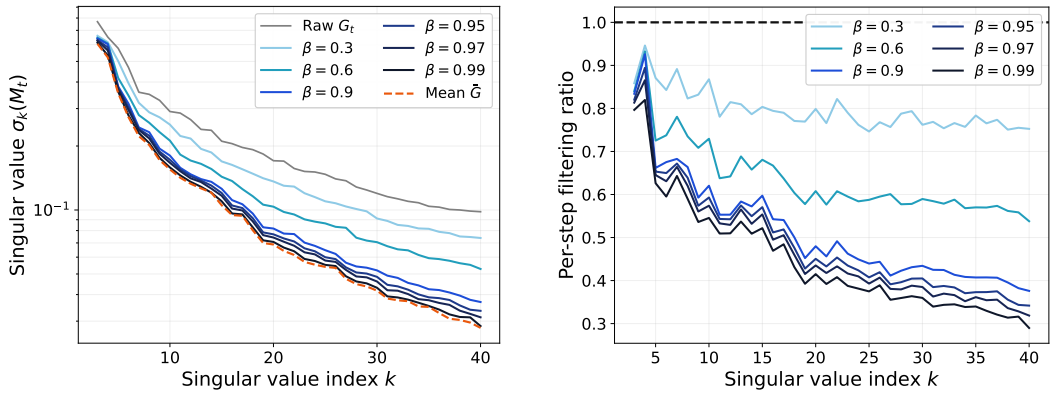


Figure 7: Synthetic stationary filtered singular value spectra under the rank-3 spiked model ($m = n = 100$, $\sigma_n = 1$, $K = 1000$, 10-trial mean) with a BVMZOS perturbation. (Left). The rank-3 spiked model with the Gaussian noise. (Right) The rank-3 spiked model with the heavy-tailed Student-t noise. The black diamonds at $k = 1, 2, 3$ mark the planted signal singular values $\sigma_k \in \{12, 8, 5\}$. The experimental details are described in Appendix F.4.1.



(a) Filtered singular value spectra.

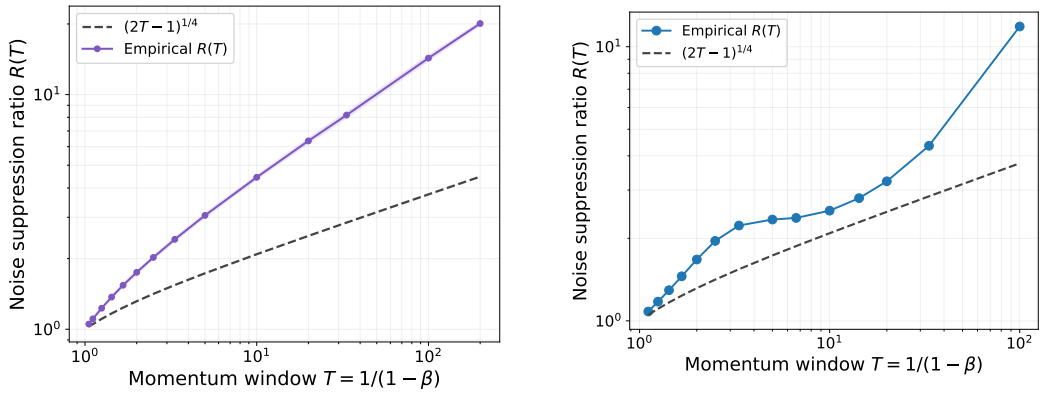
(b) Per-step filtering ratio.

Figure 8: CIFAR-10 stationary probe at `layer2.0.conv1` (128×576), warmup step 500, $K = 2000$. Index range $k \in \{3, \ldots, 40\}$. Curve and reference conventions follow figures 2a and 2b.

spectrum $\sigma_k(\bar{G})$ with the head reaching $\sigma_k(\bar{G})$ ahead of the tail, and the per-step filtering ratios decrease with the head suppressed less than the tail. This pattern, where the head filtering ratios are suppressed less than the tail, is consistent with the spectral gap predicted by Theorem 1. Figure 12 extends figure 2c to the early- and late-training checkpoints (steps 1000 and 5000) on the same three attention output projections. At both checkpoints, $R(T)$ grows at the rate above the dashed $(2T-1)^{1/4}$ floor. The depth ordering of figure 2c (`h.0` furthest above the floor, `h.11` closest to it) is preserved across training.

H Experimental Results of Subspace Alignment

In this section, we extend the subspace alignment experiments (theoretically suggested by Corollary 1) provided in figures 3 and 5 for the representative NanoGPT layer, by additionally running the same experiments on synthetic and CIFAR-10 stationary probes, the remaining stationary NanoGPT layers, and the trajectory probes on CIFAR-10 and NanoGPT. All panels follow the curve, marker, and reference conventions of figure 3 for stationary probe and figure 5 for trajectory probe. Each dashed reference is the rate of $c_r\,(2T-1)^{-1/4}$ guide with a fitted c_r, which corresponds to the high-probability Corollary 1 bound. The signal reference is the planted top-r singular subspace $(U_{\text{true}}, V_{\text{true}})$ on the synthetic simulation (where ground truth is available) and the empirical mean



(a) Synthetic noise-suppression ratio.

(b) CIFAR-10 stationary noise-suppression ratio.

Figure 9: Noise-suppression ratio $R(T)$ (Appendix F.3) on (a) synthetic rank-3 spiked gradients ($m = n = 100$, $\sigma_n = 1$, 1000 steps, 10 trials) under a BVMZOS perturbation and (b) CIFAR-10 stationary gradients at `layer2.0.conv1`. The noise-suppression ratio $R(T)$ in the synthetic simulation uses the planted signal G_t^{sig} in place of $\bar{G}$ with a zero-init bias correction (Appendix F.3). Dashed line: $(2T-1)^{1/4}$ floor.

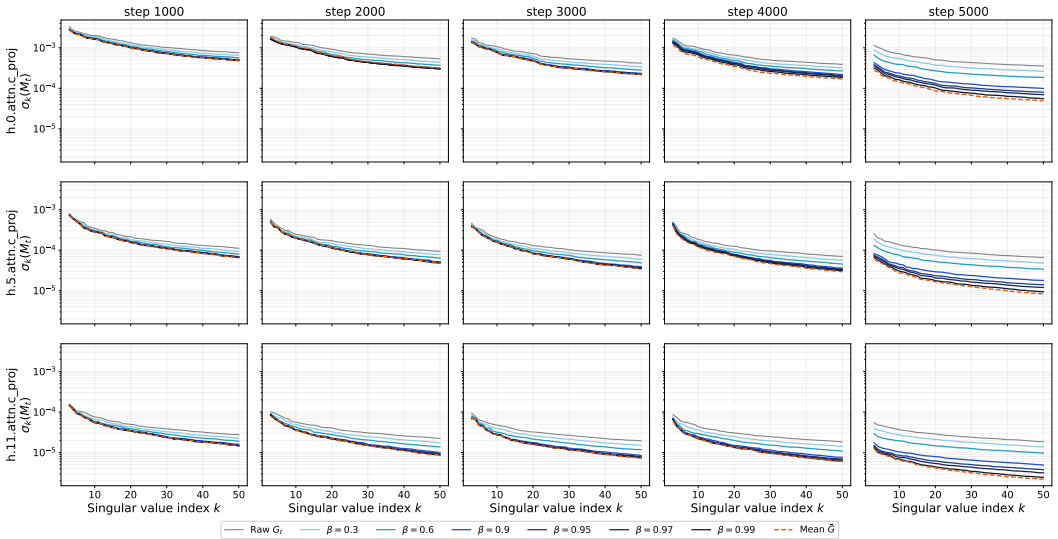


Figure 10: Stationary NanoGPT filtered singular value spectra over attention output projections `h.0`, `h.5`, `h.11` (rows) and training checkpoints 1000, 2000, 3000, 4000, and 5000 (columns), $K = 500$. Mean-gradient spectrum $\sigma_k(\bar{G})$ shown in dashed orange. Axes are shared across all fifteen cells.

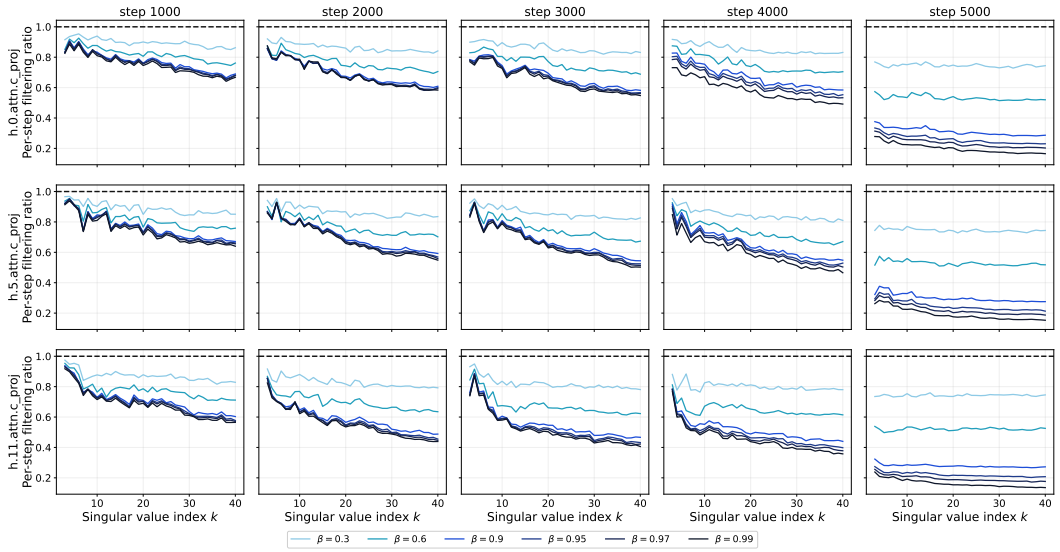


Figure 11: Stationary NanoGPT per-step filtering ratio $\mathrm{Filt}_k(\beta) = \sigma_k(M_K^{(\beta)})/\sigma_k(G_K)$ over the same (layer, step) grid as figure 10. Dashed reference at $y = 1$. Axes are shared across all fifteen cells.

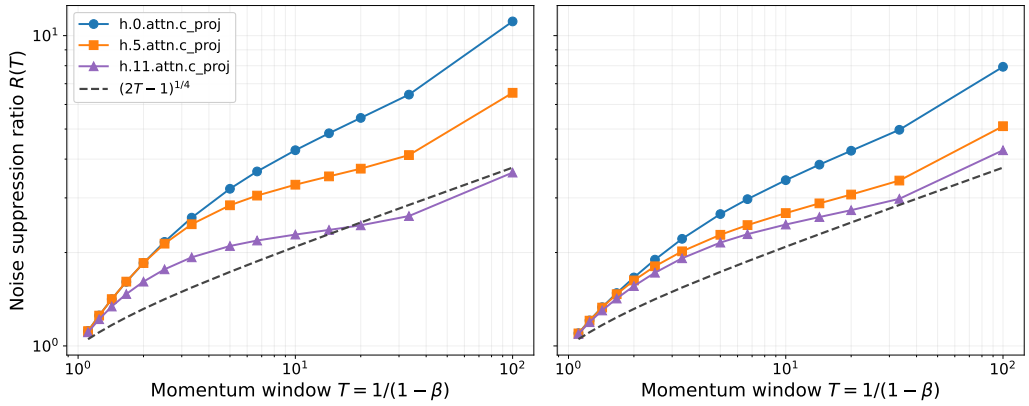


Figure 12: Stationary NanoGPT noise-suppression ratio $R(T)$ (Appendix F.3) at training checkpoints (a) step 1000 and (b) step 5000. Three attention output projections per panel. Dashed line: $(2T-1)^{1/4}$ floor.

gradient $\bar{G}$ on CIFAR-10 and NanoGPT. Across all panels, $\sin\Theta_U$ and $\sin\Theta_V$ decrease as the momentum window size T grows.

Synthetic and CIFAR-10 Stationary Probes. Figure 13 shows the subspace alignment errors $\sin\Theta_U$ and $\sin\Theta_V$ on the rank-3 spiked model shared with figure 7 at ranks $r \in \{1, 2, 3\}$, measured against the planted top-r singular subspace $(U_{\text{true}}, V_{\text{true}})$. Both errors decrease as β grows at all ranks, supporting Corollary 1. Figure 14 shows the subspace alignment errors $\sin\Theta_U$ and $\sin\Theta_V$ on the CIFAR-10 stationary probe at `layer2.0.conv1` of a ResNet-18 (matricized to 128×576), warmup step 500, $K = 2000$ at ranks $r \in \{1,5,10\}$. Both errors decrease as β grows, supporting Corollary 1.

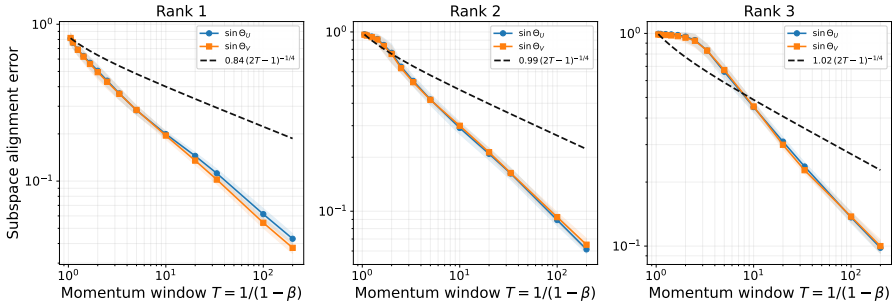


Figure 13: Subspace alignment error on the synthetic rank-3 spiked model under a BVMZOS perturbation. Panels report ranks $r \in \{1, 2, 3\}$. $\sin\Theta_U$ (blue) and $\sin\Theta_V$ (orange) are computed against the planted top-r singular subspace $(U_{\text{true}}, V_{\text{true}})$. Dashed line: fitted $c_r\,(2T-1)^{-1/4}$ guide. Shaded bands: ± 1 trial standard deviation across 10 random-seed trials.

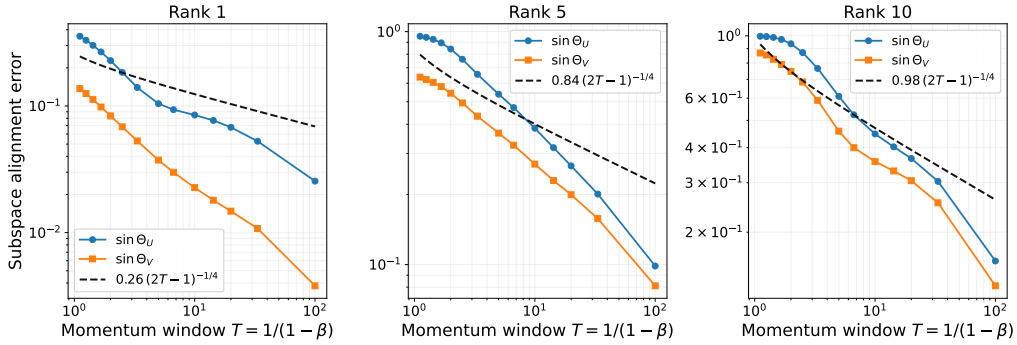


Figure 14: CIFAR-10 stationary subspace alignment error on `layer2.0.conv1` (128×576), warmup step 500, $K = 2000$, at ranks $r \in \{1,5,10\}$. Curve and reference conventions follow figure 3.

NanoGPT Stationary Probes Across Layers. Figure 15 extends figure 3 to the two remaining attention output projections `h.5.attn.c_proj` and `h.11.attn.c_proj` at step 3000, $K = 500$ at ranks $r \in \{1,5,10\}$.

CIFAR-10 Trajectory Probes. Figure 16 shows the subspace alignment errors $\sin\Theta_U$ and $\sin\Theta_V$ on the CIFAR-10 trajectory probe at `layer2.0.conv1`, $K = 100$, $I = 100$, final analysis step $t = 1500$, six-seed mean (seeds 42–47). Both errors decrease as β grows at all ranks, confirming that the rank-r subspace reliability bound survives the local-in-time relaxation of Assumption 1(a) and supporting Corollary 1.

NanoGPT Trajectory Probes Across Layers. Figures 17 and 18 extend figure 5 to the two remaining attention output projections and the three MLP output projections at training step 3000, $K = 50$, $I = 100$, three-seed mean with sample standard deviation bands across seeds.

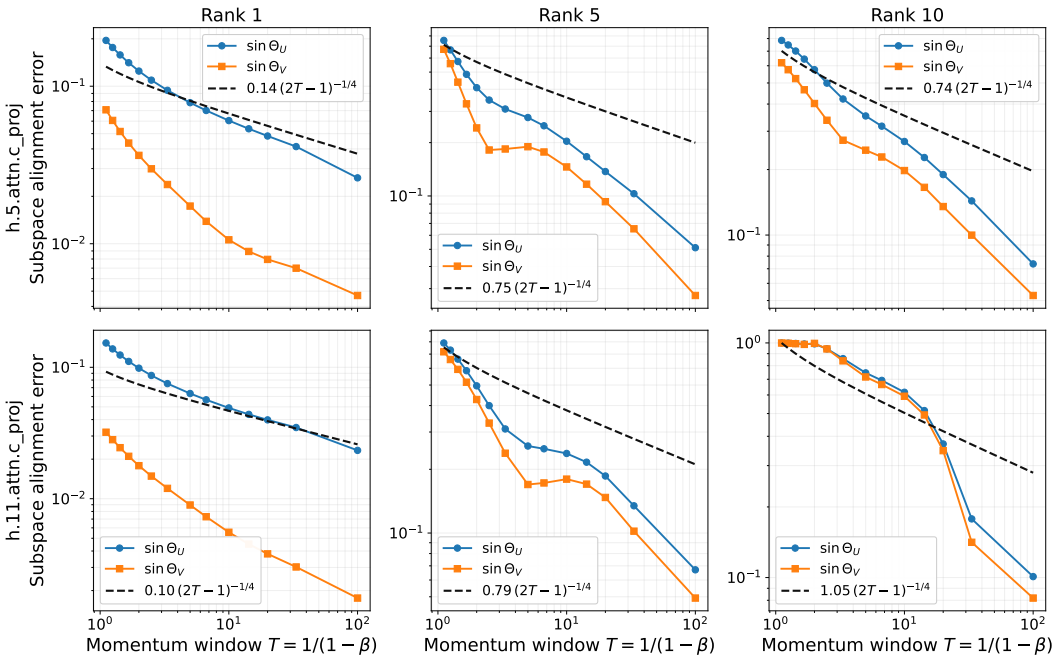


Figure 15: Stationary NanoGPT subspace alignment error on attention output projections `h.5.attn.c_proj` (top) and `h.11.attn.c_proj` (bottom) at checkpoint step 3000, $K = 500$, at ranks $r \in \{1,5,10\}$ (columns). The representative `h.0.attn.c_proj` panel is in the main text as figure 3.

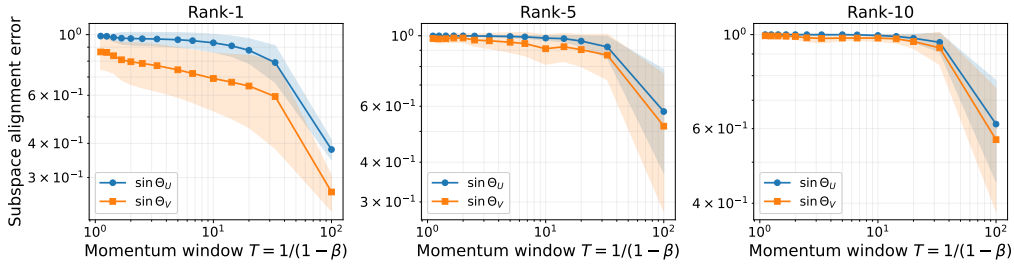


Figure 16: CIFAR-10 trajectory subspace alignment error at training step 1500 on `layer2.0.conv1`, six-seed mean (seeds 42–47, $K = 100$), at ranks $r \in \{1,5,10\}$. Solid lines: six-seed mean. Shaded bands: sample standard deviation across seeds.

I Experimental Results of Signal Alignment Ordering

In this section, we extend the signal alignment experiments (theoretically suggested by Theorem 2) provided in figures 4 and 6 for the representative NanoGPT layer, by additionally running the same experiments on synthetic and CIFAR-10 stationary probes, the remaining stationary NanoGPT layers and checkpoints, and the trajectory probes on CIFAR-10 and NanoGPT. All panels follow the curve, marker, and reference conventions of figure 4 for stationary probe and figure 6 for trajectory probe. The signal reference is the planted top-r singular subspace $(U_{\text{true}}, V_{\text{true}})$ on the synthetic simulation (where ground truth is available) and the empirical mean gradient $\bar{G}$ on CIFAR-10 and NanoGPT. Theorem 2 predicts only Pre-polar dominance over both non-denoised pipelines. The relative position of Post-polar and Polar-only is layer- and rank-dependent on real gradients and is not predicted by the theorem. Across all panels, the Pre-polar signal alignment rises monotonically with β and dominates the Post-polar and Polar-only signal alignments, while the Post-polar and Polar-only signal alignments stay nearly flat at low values.

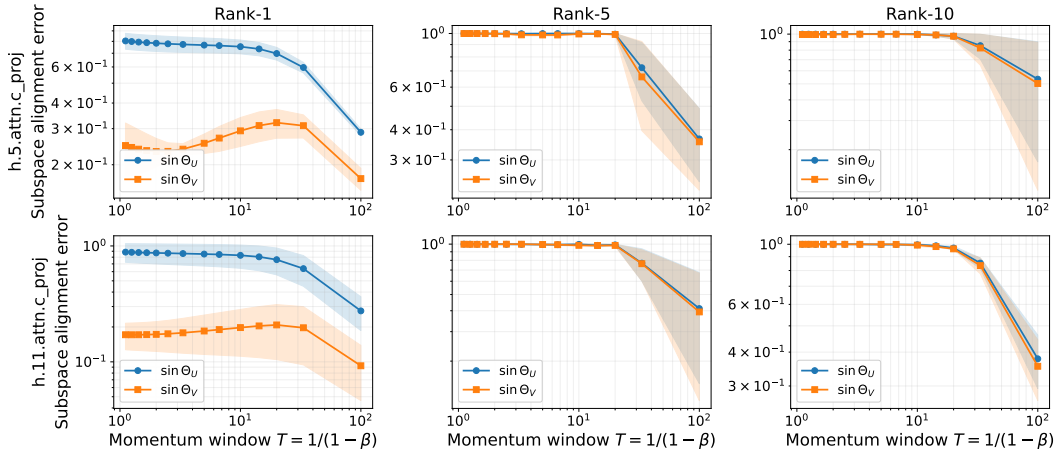


Figure 17: NanoGPT trajectory subspace alignment error on attention output projections `h.5.attn.c_proj` (top) and `h.11.attn.c_proj` (bottom) at training step 3000, trajectory buffer $K = 50$, at ranks $r \in \{1,5,10\}$ (columns). 3-seed mean with ± 1 sample-standard-deviation bands.

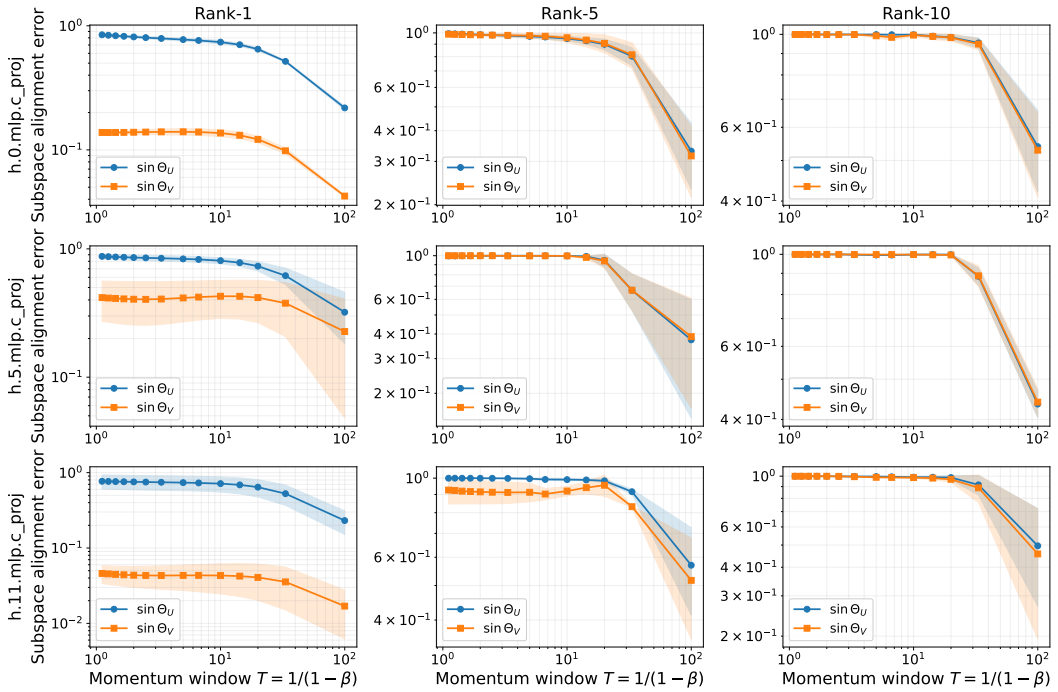


Figure 18: NanoGPT trajectory subspace alignment error on MLP output projections `h.0.mlp.c_proj` (top), `h.5.mlp.c_proj` (middle), `h.11.mlp.c_proj` (rows) at training step 3000, trajectory buffer $K = 50$, at ranks $r \in \{1,5,10\}$ (columns). Same plotting conventions as figure 17.

Synthetic and CIFAR-10 Stationary Probes. Figure 19 shows the signal alignment Align_r on the rank-3 spiked model shared with figure 7 at ranks $r \in \{1, 2, 3\}$. The Pre-polar signal alignment rises monotonically with β at every rank, while the Post-polar signal alignment declines or is nearly flat in β and the Polar-only signal alignment is β-independent by construction, supporting Theorem 2. Figure 20 supports Theorem 2 on the CIFAR-10 stationary probe at `layer2.0.conv1` of a ResNet-18 (matricized to 128×576), warmup step 500, $K = 2000$ at ranks $r \in \{1,5,10\}$ and the full-rank signal alignment. The Pre-polar signal alignment dominates both the signal alignments of non-denoised pipelines at every panel. The relative position of the Post-polar signal alignment and the Polar-only signal alignment varies across ranks on this layer.

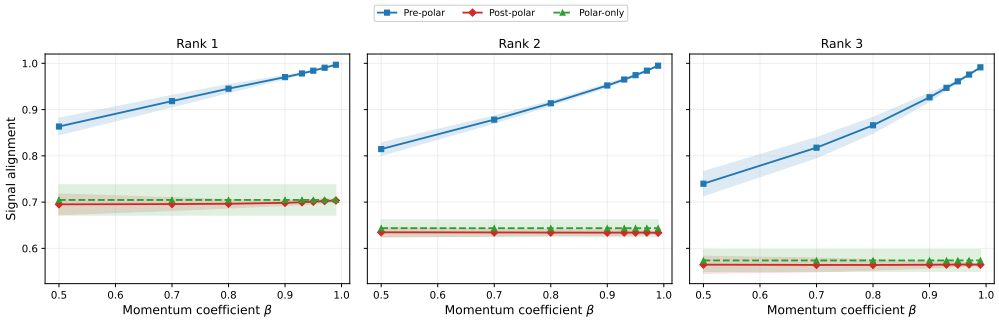


Figure 19: Synthetic signal alignment versus momentum coefficient β on the rank-3 spiked model ($m = n = 100$, $\sigma_n = 1$, $K = 1000$, 10 random-seed trials) under a BVMZOS perturbation. Panels report ranks $r \in \{1, 2, 3\}$ against the planted top-r singular subspace $(U_{\mathrm{true}}, V_{\mathrm{true}})$. Curve and reference conventions follow figure 4. Shaded bands show trial standard deviation across the 10 trials.

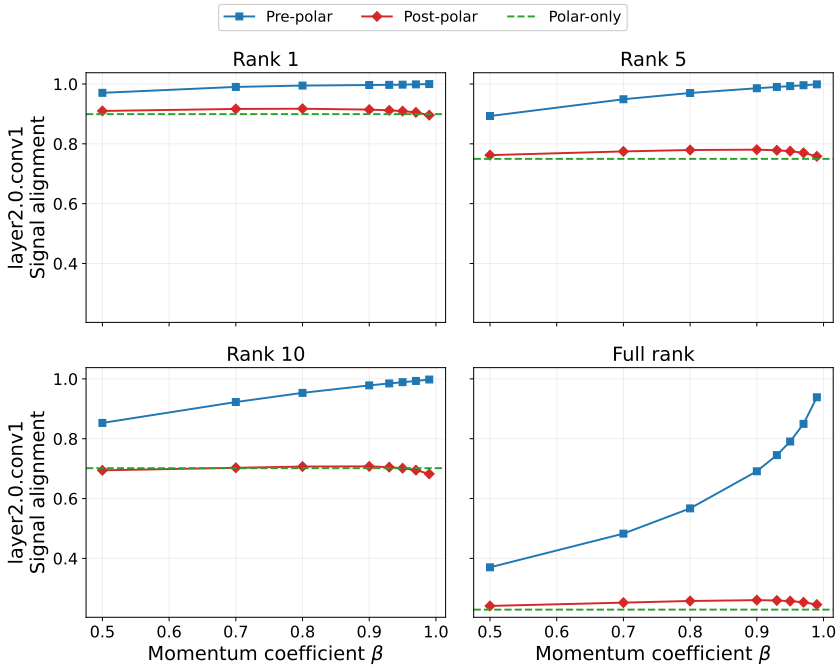


Figure 20: CIFAR-10 stationary signal alignment on `layer2.0.conv1` (128×576), warmup step 500, $K = 2000$, at ranks $r \in \{1,5,10\}$ and the full-rank signal alignment. Curve and reference conventions follow figure 4.

NanoGPT Stationary Probes Across Layers and Checkpoints. Figure 21 extends figure 4 to a 3×4 grid over three attention output projections (`h.0`, `h.5`, `h.11`) as rows and four ranks (rank-1, rank-5, rank-10, full-rank) as columns at step 3000, $K = 500$. Pre-polar dominates both non-denoised

pipelines at every panel. Rank-5 and rank-10 are the stable subspace ranks. Rank-1 is unstable on layers with a small σ_1/σ_2 gap. Tables 7 to 9 report the corresponding numerical summaries at $\beta = 0.95$ across the three attention output projections, the four (collection order, K) settings, and the five checkpoints. The Pre-polar vs. Post-polar gap is positive at every collection-order, layer, and checkpoint, and grows from step 1000 to step 5000.

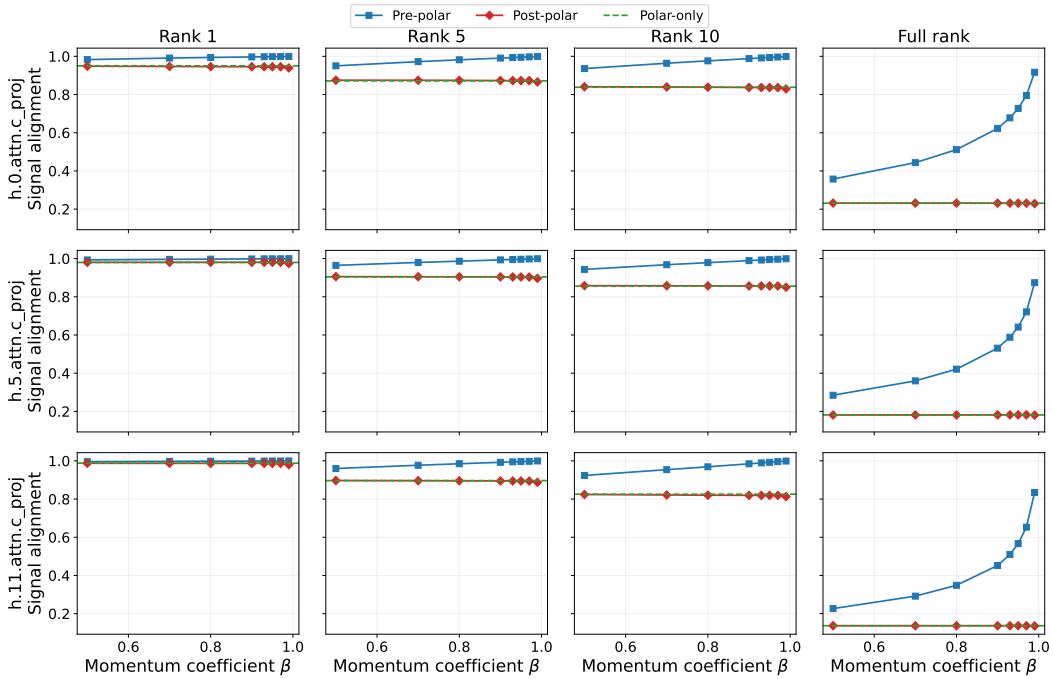


Figure 21: Stationary NanoGPT signal alignment over attention output projections `h.0` (top), `h.5` (middle), `h.11` (rows) and ranks $r \in \{1,5,10\}$ plus full rank signal alignment (columns) at checkpoint step 3000, $K = 500$. Curve and reference conventions follow figure 4. All twelve cells share the same 8-point β grid $\{0.5, 0.7, 0.8, 0.9, 0.93, 0.95, 0.97, 0.99\}$.

Table 7: Fixed-rank stationary probe averages at step 3000 across the three attention output projections, at $\beta = 0.95$. Gap columns report Pre-polar vs. Post-polar signal alignment at the indicated rank.

Collection order	Gap r1	Gap r5	Gap r10
sequential, 500 gradients	0.0277	0.1057	0.1562
sequential, 1000 gradients	0.0228	0.0963	0.1432
shuffled, 500 gradients	0.0296	0.1102	0.1619
shuffled, 1000 gradients	0.0291	0.1091	0.1595

CIFAR-10 Trajectory Probes. Figure 22 supports Theorem 2 on the CIFAR-10 trajectory probe at `layer2.0.conv1` at the final analysis checkpoint (training step 1500, $K = 100$). Panels report rank-5 signal alignment and full-rank signal alignment, both against $\mathcal{O}(\bar{G})$ of the same buffer. The Pre-polar signal alignment grows monotonically with β on both ranks, while the Post-polar signal alignment declines or is nearly flat in β and the Polar-only signal alignment is β-independent by construction. Figure 23 extends the trajectory ordering across all 15 analysis checkpoints at $\beta = 0.95$ ($T = 20$, $I = 100$). The Pre-polar vs. Post-polar full-rank alignment gap stays at 0.480 ± 0.011 across the trajectory and Pre-polar vs. Polar-only gap at 0.474 ± 0.012. Pre-polar advantage at $\beta = 0.95$ persists across training steps, mirroring the NanoGPT trajectory result of figure 6b and validating that Pre-polar signal alignment's dominance over both the signal alignments of Post-polar and Polar-only baselines survives the non-stationarity introduced by live training.

Table 8: Stationary rank-5 Pre-polar and Post-polar signal alignments and their gap per attention output projection at step 3000, $\beta = 0.95$, one row per (collection order, layer) pair under collection orders {sequential, shuffled} at $K \in \{500, 1000\}$ gradients.

Collection order	Layer	Pre r5	Post r5	Gap r5
sequential, 500 gradients	h.0	0.9953	0.8728	0.1224
sequential, 500 gradients	h.5	0.9966	0.9038	0.0928
sequential, 500 gradients	h.11	0.9963	0.8945	0.1018
sequential, 1000 gradients	h.0	0.9845	0.8771	0.1075
sequential, 1000 gradients	h.5	0.9905	0.9056	0.0849
sequential, 1000 gradients	h.11	0.9946	0.8981	0.0965
shuffled, 500 gradients	h.0	0.9903	0.8631	0.1272
shuffled, 500 gradients	h.5	0.9936	0.8962	0.0974
shuffled, 500 gradients	h.11	0.9952	0.8893	0.1059
shuffled, 1000 gradients	h.0	0.9918	0.8641	0.1277
shuffled, 1000 gradients	h.5	0.9949	0.8990	0.0959
shuffled, 1000 gradients	h.11	0.9953	0.8916	0.1036

Table 9: Stationary Pre-polar vs. Post-polar signal alignment gap averaged across the three attention output projections at $\beta = 0.95$, sequential collection order, $K = 500$, at each of the five checkpoints.

Checkpoint step	Gap r1	Gap r5	Gap r10
1000	0.0150	0.0613	0.0897
2000	0.0245	0.0959	0.1378
3000	0.0277	0.1057	0.1562
4000	0.0360	0.1197	0.1771
5000	0.3546	0.4921	0.5291

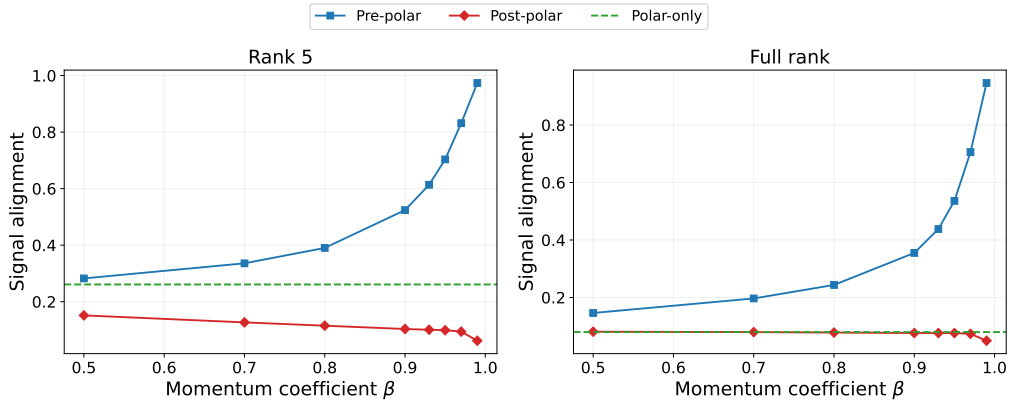


Figure 22: CIFAR-10 trajectory signal alignment versus β at training step 1500, $K = 100$, on `layer2.0.conv1`, at the rank-5 (left) and full-rank alignment (right) panels. Curve and reference conventions follow figure 4.

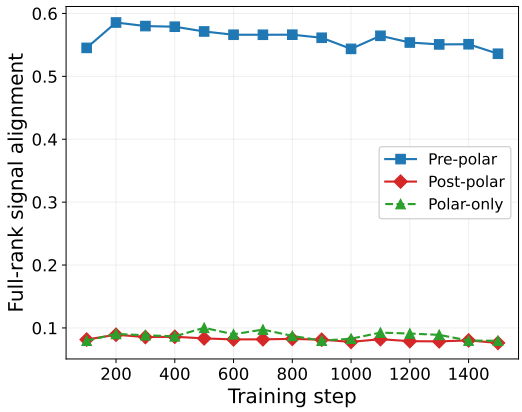


Figure 23: CIFAR-10 trajectory ordering history on `layer2.0.conv1` at $\beta = 0.95$ (trajectory buffer $K = 100$, analysis interval $I = 100$, 15 checkpoints over training step 100–1500). Curve and reference conventions follow figure 4.

NanoGPT Trajectory Probes Across Layers and Checkpoints. Figure 24 extends figure 6a to all 24 projection targets of the GPT-2 backbone (12 attention + 12 MLP output projections) at the step-3000 checkpoint, three trajectory seeds at $K = 50$, $\beta = 0.95$. Every one of the $24 \times 3 = 72$ (layer, seed) triples satisfies Pre-polar $>$ Post-polar and Pre-polar $>$ Polar-only, with MLP output projections reaching a slightly larger absolute gap than attention output projections. Table 10 reports the corresponding mean trajectory metrics aggregated over all 24 projection targets and three seeds at four common checkpoints. The Pre-polar vs. Post-polar full-rank signal alignment gap is essentially flat across the tracked checkpoints. Figure 25 tracks the trajectory ordering across training on `h.0`, `h.5`, `h.11` at $K = 50$, $\beta = 0.95$, three-seed mean. Pre-polar $>$ Post-polar and Pre-polar $>$ Polar-only at every checkpoint and every layer. The mean trajectory gap is 0.528 on `h.0`, 0.563 on `h.5`, and 0.566 on `h.11`.

Table 10: Mean full-rank Pre-polar / Post-polar / Polar-only signal alignments against $\mathcal{O}(\bar{G})$ and their pairwise gaps, averaged over all 24 projection targets and three seeds (1337, 1338, 1339) at $\beta = 0.95$, $K = 50$, at four trajectory checkpoints.

Step	Pre-polar align	Post-polar align	Polar-only align	Pre vs. Post	Pre vs. Polar-only
100	0.780	0.105	0.125	0.675	0.655
1000	0.681	0.074	0.103	0.607	0.578
2000	0.683	0.078	0.094	0.606	0.589
3000	0.682	0.080	0.100	0.602	0.582

J Hyperparameter Sensitivity in Signal Strength

This appendix sweeps the signal-to-noise ratio across three data sources at fixed momentum coefficient $\beta = 0.95$ (full settings in Appendix F.5). Synthetic λ acts directly as the leading signal singular value. On CIFAR-10 and NanoGPT, we use the mini-batch size as a proxy for SNR: at a fixed checkpoint, the stochastic mini-batch gradient is an unbiased estimator of the full-batch gradient, and its sampling noise decreases as the mini-batch size increases. Thus, increasing the mini-batch size lowers the per-step gradient noise and raises the effective gradient SNR [49]. All panels follow the curve, marker, and reference conventions of figure 4. Across all panels, the Pre-polar signal alignment dominates the signal alignments of both non-denoised pipelines: as SNR decreases (smaller λ or smaller batch), the Post-polar and Polar-only signal alignments drop sharply while Pre-polar retains high signal alignment.

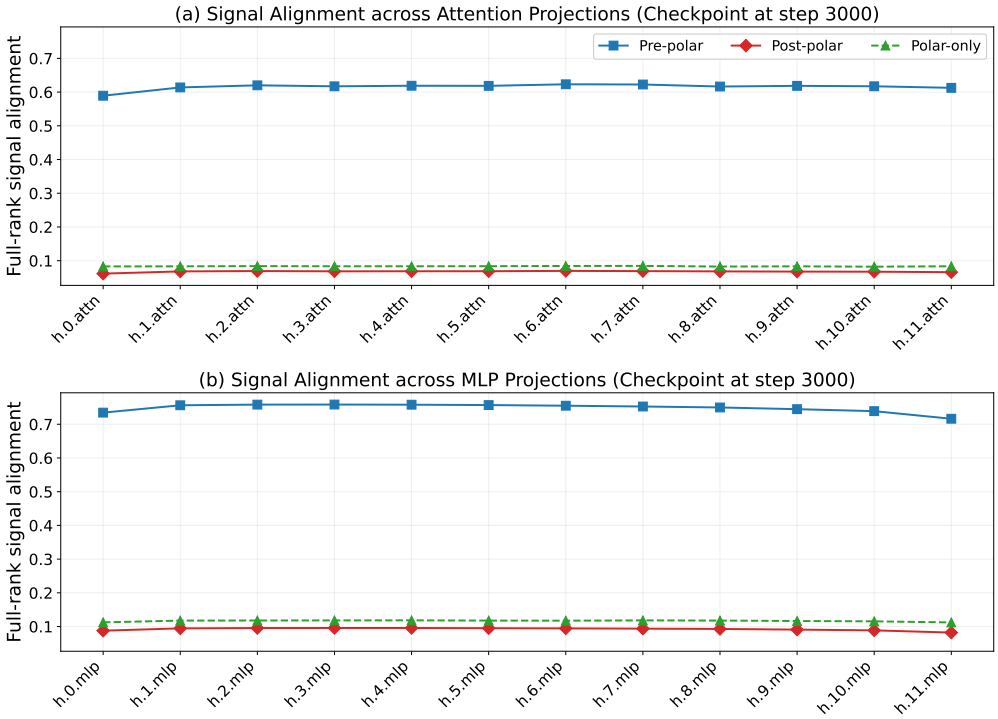


Figure 24: All-layer NanoGPT trajectory full-rank signal alignment across every attention and MLP output projection at training step 3000, $K = 50$, $\beta = 0.95$, aggregated over three seeds (1337, 1338, 1339). Curve and reference conventions follow figure 4.

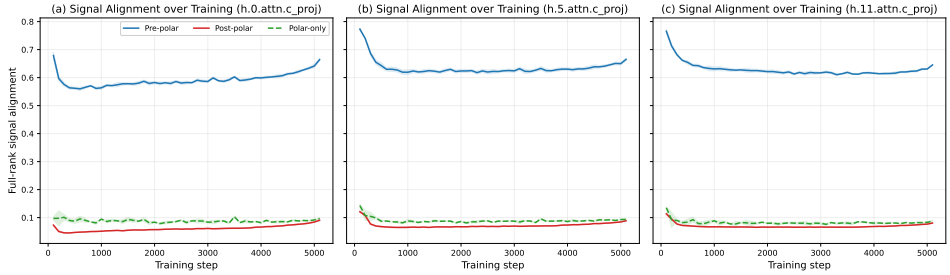


Figure 25: NanoGPT trajectory signal alignment over training on attention output projections `h.0`, `h.5`, and `h.11` at $K = 50$, $\beta = 0.95$, three-seed mean with sample standard deviation bands across seeds. Curve and reference conventions follow figure 4.

Synthetic λ Sweep Figure 26 extends Theorem 2 on the rank-3 spiked model shared with figure 19 to a sweep over the leading signal singular value λ. The Pre-polar signal alignment dominates the signal alignments of both non-denoised pipelines at every λ.

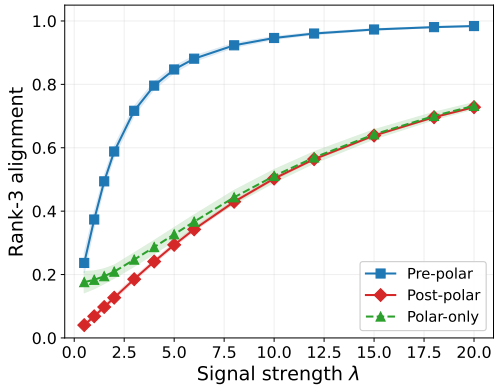


Figure 26: Synthetic rank-3 subspace alignment $\|U_3^\top A V_3\|_F/\sqrt{3}$ against the planted $(U_{\text{true}}, V_{\text{true}})$ versus signal strength λ, on the rank-3 spiked model shared with figure 19. Pre-polar (blue squares, $\mathcal{O}(M_K^{(\beta)})$), Post-polar (red diamonds, $\widetilde{M}_K^{(\beta)}$), Polar-only (green dashed, $\mathcal{O}(G_K)$). Bands are the standard deviation across 10 random-seed trials at $\beta = 0.95$.

CIFAR-10 Batch Sweep Figure 27 extends Theorem 2 to six independent CIFAR-10 stationary probes at `layer2.0.conv1` of a ResNet-18 warmed to step 500, sweeping the mini-batch size with $K = 200$ collected gradients per probe. As the mini-batch size shrinks and the per-step gradient noise grows, Post-polar and Polar-only alignment drop sharply while Pre-polar retains a high alignment, because the momentum buffer averages out the per-step noise before the polar factor is applied. The Pre-polar signal alignment therefore dominates the signal alignments of both non-denoised pipelines in every panel and at every mini-batch size.

NanoGPT Batch Sweep Figure 28 extends Theorem 2 to six independent NanoGPT stationary probes at `h.0.attn.c_proj` (768×768) of the step-3000 checkpoint, sweeping the mini-batch size with $K = 500$ collected gradients per probe experiment. The same phenomenon appears: as the mini-batch size shrinks and the per-step gradient noise grows, Post-polar and Polar-only alignment drop while Pre-polar retains a higher alignment because of the momentum filtering. The Pre-polar signal alignment dominates the signal alignments of both non-denoised pipelines in every panel and at every mini-batch size.

K Hyperparameter Sensitivity in Buffer Size

This appendix tests robustness of Pre-polar advantage to the trajectory buffer size. The two sweeps share $\beta = 0.95$ ($T = 20$) and vary K from $2.5T$ to $10T$. The Pre-polar signal alignment dominance over the signal alignments of Post-polar and Polar-only holds at every K, and Pre-polar to Polar-only alignment ratio stays approximately constant at $6\times$ on both CIFAR-10 and NanoGPT. Absolute alignment values decrease with K as $\bar{G}$ smooths over more training steps. The trajectory headline K choices in CIFAR-10 and NanoGPT are robust within this range.

CIFAR-10 K-Sweep Table 11 reports four buffer sizes $K \in \{50,100,150,200\}$ at $\beta = 0.95$ ($T = 20$) on the CIFAR-10 trajectory probe of `layer2.0.conv1` at training step 1500. As K grows from 50 to 200, the Pre-polar – Post-polar full-rank signal alignment gap falls from 0.65 to 0.33, with positive Pre-polar advantage at every row. Pre-polar to Polar-only alignment ratio stays approximately constant at $6\times$ across the sweep. The CIFAR-10 trajectory probe (Figure 23) uses $K = 100$.

NanoGPT K-Sweep table 12 reports four buffer sizes $K \in \{50,100,150,200\}$ at $\beta = 0.95$ ($T = 20$) on the NanoGPT trajectory probe of `h.0.attn.c_proj`, single seed 1337. As K grows

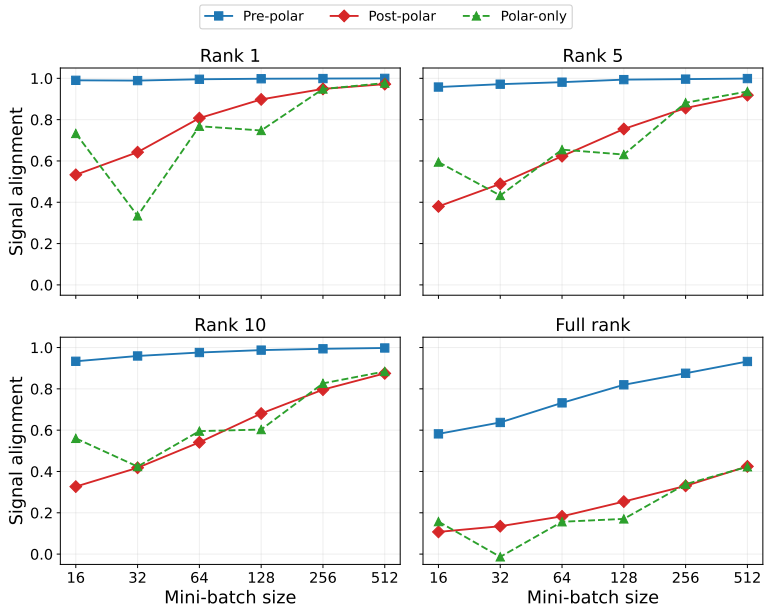


Figure 27: CIFAR-10 stationary signal alignment versus mini-batch size on `layer2.0.conv1` (128 $\times$ 576) of a ResNet-18, warmup step 500, $K = 200$ per probe, $\beta = 0.95$. Pre-polar (blue squares, $\mathcal{O}(M_K^{(\beta)})$), Post-polar (red diamonds, $\widetilde{M}_K^{(\beta)}$), Polar-only (green dashed, $\mathcal{O}(G_K)$). Panels are rank-1, rank-5, rank-10, and full-rank alignment against $\bar{G}$.

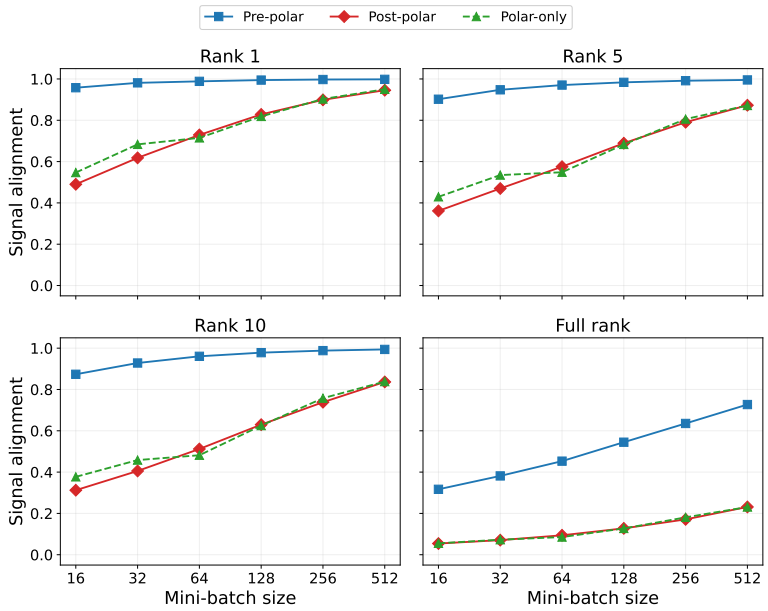


Figure 28: NanoGPT stationary signal alignment versus mini-batch size on `h.0.attn.c_proj` (768 $\times$ 768) at the step-3000 checkpoint, $K = 500$ per probe, $\beta = 0.95$. Pre-polar (blue squares, $\mathcal{O}(M_K^{(\beta)})$), Post-polar (red diamonds, $\widetilde{M}_K^{(\beta)}$), Polar-only (green dashed, $\mathcal{O}(G_K)$). Panels are rank-1, rank-5, rank-10, and full-rank alignment against $\bar{G}$.

Table 11: CIFAR-10 trajectory K-sweep on `layer2.0.conv1` at training step 1500, $\beta = 0.95$. Rank-5 columns are Align_5 and full-align columns are $\mathrm{Align}_{\mathrm{full}}$, both against $\bar{G}$ (Appendix F.3). The Pre/Polar ratio is Pre-polar full-rank alignment divided by Polar-only full-rank alignment.

K	Pre r5	Post r5	Gap r5	Pre full-align	Post full-align	Polar-only full-align	Pre/Polar ratio
50	0.8817	0.1354	0.7463	0.7479	0.1007	0.1192	6.27
100	0.6703	0.0947	0.5756	0.5420	0.0770	0.0827	6.55
150	0.5781	0.0855	0.4926	0.4452	0.0659	0.0676	6.59
200	0.5222	0.0732	0.4490	0.3837	0.0566	0.0629	6.10

from 50 to 200, the Pre-polar – Post-polar full-rank signal alignment gap falls monotonically from 0.57 to 0.28, with positive Pre-polar advantage at every row. Pre-polar to Polar-only alignment ratio stays approximately constant at 6× across the sweep. The NanoGPT trajectory probe uses $K = 50$, matching the all-layer view of Appendix I.

Table 12: NanoGPT trajectory K-sweep on `h.0.attn.c_proj` at the final training step, single seed 1337, $\beta = 0.95$. Settings beyond K and seed match Appendix F. Column conventions follow table 11.

K	Pre r5	Post r5	Gap r5	Pre full-align	Post full-align	Polar-only full-align	Pre/Polar ratio
50	0.8586	0.1603	0.6983	0.6621	0.0911	0.1001	6.61
100	0.7037	0.1278	0.5758	0.4673	0.0705	0.0710	6.58
150	0.6382	0.1103	0.5279	0.3784	0.0591	0.0602	6.29
200	0.5901	0.1018	0.4883	0.3289	0.0524	0.0544	6.05

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 abstract and the contribution list in Section 1 state three claims, all of which are formally established and empirically validated in the paper: (i) momentum opens a spectral gap and stabilizes the singular subspaces of the momentum buffer (Theorem 1 and Corollary 1 in Section 3); (ii) the Pre-polar pipeline provably dominates the Post-polar and Polar-only pipelines in expected signal alignment, with a quantitative separation under a rank-1 spiked Gaussian model in the low-SNR regime (Theorems 2 and 3 in Section 4); (iii) empirical validation across synthetic, CIFAR-10, NanoGPT (stationary and trajectory probes), and end-to-end NanoGPT and LLaMA 350M training (Sections 3, 4, 5 and Appendices G–K).

   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: A dedicated "Limitations and future work" paragraph in Section 6 flags two extensions left open: (i) practical Muon training schedules often ramp the momentum coefficient during early training, which enlarges the effective window size across iterations and would require a time-varying filter analysis to formalize; (ii) the quantitative separation in Theorem 3 is proved only under the rank-1 spiked Gaussian model, and extending it to general rank-r coherent signals or non-Gaussian perturbations would require sharper concentration tools for the polar factor.

   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: All theoretical results are numbered and cross-referenced. The assumptions are explicitly stated: Assumption 1 (structured gradient decomposition with coherent signal and MDS perturbation), Assumption 2 (signal persistence), and Assumption 5 (rank-1 spiked model). Theorem 1, Corollary 1, Theorem 2, and Theorem 3 reference the relevant assumptions in their statements. Complete proofs appear in Appendices C and D, with the supporting Proposition 1 (concentration of the momentum-filtered perturbation), Lemma 1 (quantitative gap for the polar factor), and Lemma 2 (perturbation bound for the inner product) used and properly referenced. A sub-Gaussian strengthening (Proposition 2) and sufficient conditions for signal persistence are recorded in Appendices E.2 and E.1 for completeness.

   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: Appendix F (Experimental Setups) documents every experimental detail. The hardware, precision, model architecture, data pipeline, batch shape, learning-rate schedule, regularization, and DDP layout are reported in Tables 1, 2, and 3. The probe protocols (Appendix F.2), measurement definitions (Appendix F.3), and metrics (Appendix F.3) are described step-by-step, including the analysis pipeline that recomputes Pre-polar, Post-polar, and Polar-only buffers from a shared gradient stream. Per-tier best configurations selected from the hyperparameter sweeps are listed in Tables 5 and 6. Per-task probe-side β_{probe} grids, buffer sizes K, analysis intervals, and seeds are listed in Appendices F.5–F.4.5. Together these details fully characterize the protocols required to reproduce all main claims.

   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: All datasets used are publicly available: CIFAR-10, FineWeb-10B (used through the modded-NanoGPT pipeline), and C4 [42]. The NanoGPT training code is adapted from the publicly released modded-NanoGPT speedrun [24]. An anonymized code archive containing the analysis pipeline and the end-to-end training scripts is provided alongside the paper as supplementary material. Detailed instructions on data preparation, environment configuration (compiler, CUDA, Python versions in Table 3).

   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: Appendix F provides the comprehensive specification: model family, scale, design choices, data pipeline, sequence length, batch shape, training steps, learning-rate schedule, regularization, and DDP layout for both the NanoGPT (Table 1) and LLaMA 350M (Table 2) end-to-end runs, and the shared Wisteria configuration (Table 3) for the synthetic, CIFAR-10, and NanoGPT probes. The hyperparameter selection procedure (single-seed sweep on Muon learning rate $\times$ momentum coefficient grid, then 3-seed final runs at the selected configurations) is documented in Appendices F.4.4 and F.4.5. Per-task probe-side β_{probe} grids and buffer sizes K are listed in Appendix F.5.

   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: [Yes]

   Justification: Experimental variability is reported across all main empirical claims. The end-to-end NanoGPT and LLaMA 350M comparisons (Figure 1, Tables 5 and 6) report 3-seed mean and sample standard deviation across independent training runs (seeds 1–3). The NanoGPT trajectory probe (Figures 5, 6, and Figures 17 and 18) uses three seeds (1337, 1338, 1339) and reports per-step seed mean with sample standard deviation bands. The CIFAR-10 trajectory probe (Figure 16) uses six seeds (42–47), and the synthetic experiments (Appendices G, H, I) use 10-trial means with ± 1 trial-standard-deviation bands. The variability factor is the random seed (training initialization, data ordering, and synthetic perturbation draws), and shaded bands are sample standard deviations of the seed/trial-mean estimator. Tables 7–10, 11, 12 report numerical mean values, with mean $\pm$ sample standard deviation in Table 6.

   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: Appendix F.1 documents the compute environments. The end-to-end NanoGPT three-pipeline comparison runs on 4 NVIDIA L40S GPUs in a single-node PyTorch DDP setup with `bfloat16` autocast, and the LLaMA 350M three-pipeline comparison runs on 8 NVIDIA L40S GPUs in a single-node PyTorch DDP setup with `bfloat16` (Tables 1 and 2). The synthetic, CIFAR-10, and NanoGPT probe experiments run on NVIDIA A100 40GB GPUs on the Wisteria Aquarius cluster at the University of Tokyo, with the compiler / CUDA / Python versions and analysis precision (exact `float32` SVDs) recorded in Table 3. The hyperparameter sweep size (3 LRs $\times$ 3 betas for NanoGPT, 6 LRs $\times$ 4 betas for LLaMA), the 5100/3000 training-step budgets, and the 3-seed final stage establish the per-experiment compute footprint. The probe runs are inference-only single-node A100 jobs with negligible additional cost relative to the training 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: The paper is a theoretical analysis of an existing optimizer (Muon) with experimental validation on standard public datasets (CIFAR-10, FineWeb-10B, C4) and reduced-scale, non-deployed academic models. No human subjects, personal data, scraped data, or sensitive datasets are used. No new pre-trained models or datasets with misuse potential are released. The work has been conducted in accordance with the NeurIPS Code of Ethics, and anonymity is preserved in the submission.

   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 foundational research on the theoretical mechanism by which momentum filters gradient streams inside matrix-based optimizers, with the aim of explaining an existing algorithm (Muon) rather than proposing a deployable system or application. The contribution is generic optimizer theory analogous to classical momentum analyses; it does not enable novel applications or new data sources, and there is no direct path from the theoretical understanding to a specific positive or negative societal impact beyond the diffuse efficiency benefits shared by all general-purpose neural-network optimizers. Following the NeurIPS guideline, we mark this question as not applicable.

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 paper does not release pre-trained generative models, image generators, scraped datasets, or any other asset with a high risk for misuse. The contribution is theoretical, and the experiments use existing public benchmarks (CIFAR-10, FineWeb-10B, C4) together with reduced-scale academic models trained for analysis purposes. The accompanying code release covers analysis pipelines and training scripts, not deployable model weights.

    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: All existing assets used in the paper are properly credited via citation. The Muon optimizer is cited to Jordan et al. [25]; the modded-NanoGPT speedrun training code is cited to Jordan et al. [24]; the C4 dataset is cited to Raffel et al. [42]; CIFAR-10 and FineWeb-10B are used as standard public benchmarks. We use the standard public versions of all assets and follow their respective publicly available terms of use. No re-packaged or derived datasets are released.

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 release any new dataset, pre-trained model, or benchmark. The contribution is a theoretical analysis of an existing optimizer (Muon), with experiments on existing public benchmarks. The accompanying analysis code is documented within the codebase via the protocols, measurements, and metrics in Appendices F.2, F.3, and F.3, but is supporting infrastructure rather than a primary research asset.

    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 paper does not involve crowdsourcing or research with human subjects. All experiments are computational and use existing public benchmarks.

    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 research with human subjects, so IRB approval (or equivalent) is 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 (NanoGPT and a reduced-scale LLaMA-style 350M model) appear in the paper as experimental subjects on which the analyzed optimization pipelines are evaluated, not as components of the core methodology. The core contributions (Theorem 1, Corollary 1, Theorem 2, Theorem 3) are theoretical results about momentum, the polar factor, and signal recovery in matrix-based optimization, derived without reference to LLMs as a methodological component. LLMs were not used in any important, original, or non-standard role in developing the methodology.

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.