MUON AS A KRONECKER-FACTORED CURVATURE APPROXIMATION WITHOUT ACCUMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Kronecker-factored methods such as K-FAC, Shampoo, and SOAP approximate layer curvature with input and output factors that they accumulate across steps, whereas Muon orthogonalizes the current momentum and stores no curvature matrix. The algebraic link between Muon and Shampoo without accumulation does not show whether Muon's momentum statistics follow the curvature of trained networks. We study Muon as an accumulation-free Kronecker-factored curvature approximation with two substitutions: the input factor is read from the Gram of the current momentum, and the output factor is replaced by the identity. A signal and noise decomposition explains when the momentum Gram follows the leading curvature subspace and separates subspace recovery from spectral accuracy. Measured against a sequence-level generalized Gauss-Newton reference in GPT-2, the momentum Gram captures most of the leading input curvature energy, whereas the output factor is nearly diagonal pairwise but has unequal diagonal scales and collective structure. Controlled quadratic and nonlinear experiments show that estimation noise, unequal output scales, and output coupling cause distinct failures. These results support a conditional curvature interpretation of Muon and specify where it falls short.

1 INTRODUCTION

Curvature determines how an update to one parameter interacts with updates to others. Diagonal methods such as Adam and AdamW adapt coordinate scales, while matrix methods also represent interactions within a weight matrix (Kingma & Ba, 2014; Loshchilov & Hutter, 2017; Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2025). Kronecker-factored methods make this tractable by approximating a layer's curvature with an input factor R and an output factor L, and they accumulate both factors across steps. Muon follows a different route: it orthogonalizes the momentum M of each weight matrix and stores no curvature matrix, yet it trains large language models competitively (Jordan et al., 2024; Liu et al., 2025a).

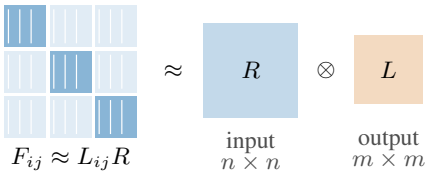
The link between Muon and matrix preconditioning is known: removing accumulation from Shampoo yields a polar update (Jordan et al., 2024; Bernstein & Newhouse, 2024b). This algebraic identity does not show that momentum estimates the curvature met during training, nor which part of a curvature correction Muon retains. We therefore ask how well the factors implied by Muon match a measured curvature reference, and which failures of this match affect optimization.

We study Muon as an accumulation-free Kronecker-factored curvature approximation (fig. 1). The identity $\operatorname{polar}(M) = M(M^\top M)^{\dagger/2}$, where $A^{\dagger/2}$ is the positive-semidefinite square root of the pseudoinverse of A, equals Kronecker whitening with the momentum Gram $M^\top M$ as the input factor and the identity as the output factor. Muon thus behaves as a curvature method to the extent that these two substitutions hold.

The input substitution can hold even when the momentum is dominated by its mean: its expected Gram splits into a gradient second moment, which approximates curvature when the empirical Fisher does, and a signal Gram of the form RAR with a residual-dependent $A \succeq 0$, which also emphasizes leading curvature directions. The output substitution makes two approximations, negligible coupling and equal scales, whose failures require different corrections.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(a) Kronecker model of a layer



Kronecker whitening: $L^{-1/2}\, G\, R^{-1/2}$
Muon, $\mathrm{polar}(M)$: $I\, M\, (M^\top M)^{-1/2}$

(b) Factor estimates

Method	Input $\widehat{R}$	Output $\widehat{L}$	Stored
K-FAC / EKFAC	$\mathbb{E}[xx^\top]$	$\mathbb{E}[\delta\delta^\top]$	factors
Shampoo / SOAP	$\mathbb{E}[G^\top G]$	$\mathbb{E}[GG^\top]$	Grams
Running averages over steps and batches			
Muon	$M^\top M$	I	none
Current momentum only; no curvature accumulator			

Figure 1: **Muon Through Kronecker Whitening.** (a) Every block F_{ij} between output neurons i and j shares one input factor R and is scaled by L_{ij} (eq. (3)). Muon's update equals Kronecker whitening with L replaced by I and R by the momentum Gram. (b) Stored-factor methods accumulate both factors, with x the layer input, δ the backpropagated output gradient, and G the gradient; Muon reads the input factor from the current momentum M. The methods also differ in curvature power and in eigenbasis corrections.

We test these conditions against a sequence-level generalized Gauss-Newton (GGN) reference in a modified GPT-2. Muon's leading 16 input directions capture 77% to 94% of the best achievable leading energy, compared with 94% to 99% for SOAP's stored factor. When the curvature window is set separately from the update, the window with the most accurate spectrum also yields the lowest loss wherever differences exceed seed variation. The output factor is nearly diagonal pairwise, but its diagonal scales are strongly unequal in query and key projections. Our contributions are diagnostic rather than a new optimizer:

- We read Muon as an accumulation-free Kronecker-factored curvature approximation with two substitutions, the momentum Gram for the input factor and the identity for the output factor (section 4.1).
- We derive when the momentum Gram tracks leading input curvature, separating realization noise from signal bias, and split the identity output factor into two approximations with different corrections (sections 4.2 and 4.3).
- We measure the Kronecker premise and both substitutions module by module against a GGN reference and relate curvature windows to loss (section 5; ViT-Tiny and Gated DeltaNet in section E).
- We break one approximation at a time in quadratic and nonlinear problems and reproduce the predicted failures (section 6).

2 Related Work

Interpretations of Muon. Muon is commonly motivated as steepest descent under the spectral norm or as modular dualization (Jordan et al., 2024; Bernstein & Newhouse, 2024b;a), with extensions to trust-region, linear-minimization-oracle, and constrained formulations and to convergence analyses (Kovalev, 2025; Pethick et al., 2025; Riabinin et al., 2025; Chen et al., 2025; Sfyraki & Wang, 2025; Shen et al., 2025; Davis & Drusvyatskiy, 2025). A second line connects Muon to curvature. Without accumulation, Shampoo's two-sided update reduces to a polar update (Jordan et al., 2024; Bernstein & Newhouse, 2024b), as do one-sided ASGO and KL-Shampoo without momentum (An et al., 2025; Lin et al., 2025); structured-Fisher and symmetry-based estimates yield Muon-like updates under further approximations (Gong et al., 2025; Artemev et al., 2026), and Shampoo decomposes into an adapted Muon update (Eschenhagen et al., 2026). Other analyses assume a Kronecker or isotropic curvature model (Nguyen et al., 2026; Su, 2025), measure the curvature penalty along Muon updates (Wang et al., 2026), or read Muon as omitting an input-side preconditioner (Du & Su, 2026). We instead assign the momentum Gram to the input side and compare the implied factors with a GGN reference in trained Transformers.

Kronecker-factored curvature. K-FAC approximates layer Fisher blocks by Kronecker products of activation and backpropagation factors, EKFAC re-estimates diagonal scales in their eigenbasis, and Shampoo and SOAP precondition with accumulated gradient Grams (Martens & Grosse, 2015; George et al., 2018; Gupta et al., 2018; Anil et al., 2020; Vyas et al., 2025; Morwani et al., 2024). All of them accumulate both factors. Approximate Kronecker and near-block-diagonal Hessian structure supports this premise (Wu et al., 2020; Zhang et al., 2024; Dong et al., 2025), but self-attention Hessians are heterogeneous (Ormaniec et al., 2024) and block approximations can fail in large models (Granziol & Juarev, 2026); accurate layerwise Gauss-Newton steps also outperform SOAP and Muon (Abreu et al., 2025), although K-FAC's gains need not reflect accurate second-order updates (Benzing, 2022). We test the shared-input-factor premise per module.

Gradient noise and curvature estimates. Mean-plus-covariance decompositions underlie analyses of variance adaptation and batch size (Balles & Hennig, 2018; McCandlish et al., 2018; Zhang et al., 2019). Gradient covariance, Fisher, and Hessian are distinct objects (Thomas et al., 2020; Zhang et al., 2026c), and the observed-label empirical Fisher generally differs from the model Fisher (Kunstner et al., 2019). Muon-specific analyses cover momentum as a spectral filter, large-batch square-root preconditioning, and batch sizes (Li et al., 2026; Paquette et al., 2026; Sato et al., 2025; Naganuma et al., 2026), and sharing samples between gradient and preconditioner adds bias (Nayak et al., 2026). We add the RAR structure of the signal Gram, which gives conditions under which the momentum Gram tracks leading curvature.

Muon variants. Several variants reintroduce what the two substitutions discard. Neuron-wise and elementwise second-moment scaling after orthogonalization reintroduce output scales (Li et al., 2025; Si et al., 2025), and per-head or grouped-head orthogonalization lets attention heads update at different scales (GLM-5-Team, 2026; Kimi Team, 2026; Zhang et al., 2026a). Stored or curvature-aware factors (Zhang et al., 2026b; Xu et al., 2026; Che, 2026; Liu et al., 2025b; Du & Su, 2026) and variance reduction (Liu et al., 2025c) target the input side and its noise, while equilibration (Chang et al., 2026) and QK-Clip (Kimi Team, 2025) address conditioning and logit growth. Our measurements indicate which correction each module needs.

3 Preliminaries

3.1 Kronecker-Factored Curvature

Consider a weight matrix $W \in \mathbb{R}^{m\times n}$ in a linear map $y_t = Wx_t$ applied at positions t of an input sequence. With ℓ the sequence loss and $\delta_t = \partial\ell/\partial y_t$ the backpropagated output gradient, sequence s contributes the gradient $G_s = \sum_t \delta_{s,t} x_{s,t}^\top$, which sums all uses of the shared weight (Eschenhagen et al., 2023). With column-major vectorization, the model Fisher (Amari, 1998) is

$$F = \mathbb{E}_s \mathbb{E}_{\widetilde{y}\sim p_\theta}[\mathrm{vec}(G_s)\,\mathrm{vec}(G_s)^\top], \tag{1}$$

where labels $\widetilde{y}$ are sampled from the model's predictive distribution p_θ on the fixed input sequence. For softmax cross-entropy on logits, F equals the GGN matrix (Schraudolph, 2002; Martens, 2020), which differs in general from both the exact Hessian and the observed-label empirical Fisher (Kunstner et al., 2019).

A Kronecker model approximates this $mn \times mn$ block by two small factors,

$$F \approx R \otimes L, \qquad (R \otimes L)^{-p}\,\mathrm{vec}(G) = \mathrm{vec}(L^{-p} G R^{-p}), \tag{2}$$

where $R \in \mathbb{R}^{n\times n}$ acts on inputs, $L \in \mathbb{R}^{m\times m}$ on outputs, G is any gradient matrix, and $p > 0$ is the curvature power; the identity holds for positive-definite factors and on their support otherwise. Reordering coordinates by output neuron, the $n \times n$ block that couples output neurons i and j is

$$F_{ij} \approx L_{ij} R. \tag{3}$$

Every block thus shares one input factor, and L sets their relative scales and couplings. Since the factors are defined only up to a reciprocal scalar, we fix $R = \mathbb{E}[G_s^\top G_s]$ and $L = \mathbb{E}[G_s G_s^\top]/\,\mathrm{tr}(F)$ with sampled labels, so that $\mathrm{tr}(L) = 1$ and $R \otimes L = F$ whenever F is exactly Kronecker. These partial traces differ from the activation and backpropagation factors of K-FAC.

Stored-factor methods accumulate factor estimates across steps. K-FAC applies an approximate inverse Fisher ($p = 1$), Shampoo applies inverse fourth roots of two Gram accumulators, and EKFAC and SOAP further adapt scales in the factor eigenbasis (Martens & Grosse, 2015; Gupta et al., 2018; George et al., 2018; Vyas et al., 2025). Our Right-EMA comparator stores only the input Gram and adapts factored diagonal scales in its eigenbasis (section A.1).

3.2 Muon

Muon keeps one momentum buffer per matrix. In the form used in our experiments,

$$M_t = \beta M_{t-1} + (1-\beta)G_t, \qquad W_{t+1} = (1-\eta_t\gamma)W_t - \eta_t c_{m,n}\,\mathrm{polar}(M_t), \tag{4}$$

where G_t is the minibatch gradient, β the momentum coefficient, η_t the learning rate, γ the weight decay, and $c_{m,n} = \sqrt{\max(1, m/n)}$. For a rank-$r$ compact singular value decomposition $M = U_r S_r V_r^\top$, $\mathrm{polar}(M) = U_r V_r^\top$. The implementation approximates this polar factor with five Newton-Schulz polynomial steps (Jordan et al., 2024; Amsel et al., 2025). Our GPT-2 runs use bias-corrected momentum without Nesterov, unlike the original implementation.

4 Muon as Accumulation-Free Kronecker Whitening

4.1 Two Factor Substitutions

The polar factor satisfies

$$\mathrm{vec}(\mathrm{polar}(M)) = \left((M^\top M) \otimes I_m\right)^{\dagger/2} \mathrm{vec}(M). \tag{5}$$

Comparing with eq. (2) at $p = 1/2$, Muon is Kronecker whitening with $\widehat{R} = M^\top M$ and $\widehat{L} = I_m$. The first substitution reads input curvature from the current momentum; the second discards output coupling and relative output scales. Factors are compared up to a positive scalar, which the learning rate absorbs; $L \approx I$ below means $L \approx I/m$ under our trace convention.

Agreement of leading eigenvectors is weaker than agreement of updates. On the subspace spanned by M, $\mathrm{polar}(M)$ matches $MR^{-1/2}$ up to scale if $M^\top M \propto R$. Shared eigenvectors alone do not suffice, because the relative eigenvalues set the whitening strengths. We therefore measure subspaces, captured energy, and spectra separately.

The identity $\mathrm{polar}(M) = (MM^\top)^{\dagger/2}M$ equally admits a left-sided reading. We adopt the right-sided reading to test shared input geometry and output-neuron heterogeneity. This is a testable modeling choice, not a consequence of the algebra. It also makes Muon's economy explicit: it keeps only the mn momentum entries, whereas stored-factor methods add $n^2 + m^2$ factor entries and their eigendecompositions.

4.2 Input Factor from Momentum

We first ask what the momentum Gram estimates. At fixed weights, let independent sequence gradients have mean μ and input-side covariance $\Sigma = \mathbb{E}[(G_s - \mu)^\top (G_s - \mu)]$. Writing the momentum as $M = \sum_s a_s G_s$ with weights $\sum_s a_s = 1$ and effective sample count $N_{\mathrm{eff}} = 1/\sum_s a_s^2$,

$$\mathbb{E}[M^\top M] = \mu^\top \mu + \frac{\Sigma}{N_{\mathrm{eff}}}. \tag{6}$$

This is the matrix analogue of familiar variance decompositions (Balles & Hennig, 2018; McCandlish et al., 2018). For a steady-state EMA over independent minibatches of B_s sequences, $N_{\mathrm{eff}} = B_s(1+\beta)/(1-\beta)$, a window of 19 minibatches at $\beta = 0.9$. The two terms have equal trace at $N_{\mathrm{eff}} = B_{\mathrm{noise}} = \mathrm{tr}(\Sigma)/\|\mu\|_F^2$, the simple noise scale of McCandlish et al. (2018).

The noise term carries curvature. With $C = \mathbb{E}[G_s^\top G_s]$ formed from observed labels, $\Sigma = C - \mu^\top \mu$ and

$$N_{\mathrm{eff}}\mathbb{E}[M^\top M] = C + (N_{\mathrm{eff}} - 1)\mu^\top \mu \approx R + (N_{\mathrm{eff}} - 1)\mu^\top \mu. \tag{7}$$

Only the last step assumes $C \approx R$. Sampled labels give R exactly in expectation, but observed labels need not (Kunstner et al., 2019), so we treat this correspondence as an assumption to check.

The signal term is also shaped by curvature. In a local Kronecker quadratic model of the expected loss, $\mathcal{L}(W) = \mathcal{L}_* + \frac{1}{2}\,\mathrm{vec}(\Delta)^\top (R \otimes L)\,\mathrm{vec}(\Delta)$ with minimizer W_* and residual $\Delta = W - W_*$, the mean gradient is $\mu = L\Delta R$, so

$$\mu^\top \mu = RAR, \qquad A = \Delta^\top L^2 \Delta \succeq 0. \tag{8}$$

With $C \approx R$ and the eigendecomposition $R = Q\Lambda Q^\top$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$,

$$Q^\top \big(N_{\mathrm{eff}} \mathbb{E}[M^\top M]\big) Q \approx \Lambda + (N_{\mathrm{eff}} - 1)\Lambda \widetilde{A} \Lambda, \qquad \widetilde{A} = Q^\top A Q. \tag{9}$$

The first term has the eigenvectors of R. The second weights direction i by $\lambda_i^2 \widetilde{A}_{ii}$, so it also emphasizes leading curvature directions unless the residual has been depleted there, consistent with gradients concentrating in the top Hessian subspace (Gur-Ari et al., 2018), although that component need not drive progress (Song et al., 2025). Its off-diagonal entries obey $|\widetilde{A}_{ij}| \le (\widetilde{A}_{ii}\widetilde{A}_{jj})^{1/2}$, so a leading subspace stays stable when its eigengap dominates the coupling to the rest, while directions of similar curvature mix easily. Even with the right basis, the spectrum follows R only if the signal energy $\lambda_i^2 \widetilde{A}_{ii}$ is roughly proportional to λ_i; a signal close to R^2 preserves eigenvectors but exaggerates anisotropy. Equation (9) is thus a conditional explanation of alignment, not a guarantee.

The same decomposition separates two errors. Small N_{eff} leaves realization noise in the single Gram that Muon forms, whereas large N_{eff} exposes the bias of the signal term. A stored factor averages many minibatch Grams: at stationary weights this reduces fluctuations while keeping the expectation at $N_{\mathrm{eff}} = B_s$, and its window can be chosen independently of the update. Moving weights add staleness, which eq. (6) omits, and the expected Gram does not determine the expected inverse root (Nayak et al., 2026).

These conditions yield three predictions, which section 5.3 tests. First, Muon's Gram should recover the leading eigenvectors best and mix directions of similar curvature. Second, at fixed weights its error should change from realization noise to signal bias as the number of averaged gradients passes B_{noise}. Third, because whitening strengths depend on eigenvalue ratios, spectral mismatch rather than subspace overlap should track which curvature window trains best.

4.3 Output Factor as the Identity

In eq. (3), the off-diagonal entries L_{ij} couple output neurons, and the diagonal entries L_{ii} set their relative curvature scales. Replacing L with a scaled identity therefore makes two approximations: negligible coupling and equal diagonal scales. Near-block-diagonal Hessians observed in Transformers (Zhang et al., 2024), and explained by large output-class counts (Dong et al., 2025), motivate the first.

Equal scales are the natural default for exchangeable neurons but are not guaranteed after training. Head specialization and heterogeneous MLP activity break the symmetry (Voita et al., 2019; 2023); elementwise nonlinearities distinguish coordinates, and RoPE distinguishes frequency pairs within query and key heads (Su et al., 2021). Residual-stream coordinates have no architecturally privileged basis (Elhage et al., 2021), although Adam-trained models can develop basis-aligned outliers (Elhage et al., 2023). We thus expect the most uniform scales in output and MLP down-projections and the least uniform ones in query and key projections.

Small pairwise coupling does not imply $L \approx cI$. For the equicorrelation matrix $L = (1-\alpha)I + \alpha \mathbf{1}\mathbf{1}^\top$, with $\mathbf{1}$ the all-ones vector, every normalized off-diagonal entry is α, but the leading eigenvalue is $1 + (m-1)\alpha$. The two failures also call for different corrections. Unequal diagonal scales need only a rowwise rescaling, whereas substantial coupling makes $L^{-1/2}$ mix output directions, which no diagonal scaling reproduces.

5 Analysis on Real Networks

5.1 Setup

Models and training. We train a modified GPT-2 small model (12 layers, width 768, 12 heads, RMSNorm, RoPE, untied embeddings) on FineWeb-10BT with sequence length 1024 (Penedo et al., 2024). Hidden matrices use Muon or SOAP; embeddings, readout, and normalization parameters use

Table 1: **Diagnostics.** U_k and Λ_k hold the top-k eigenvectors and eigenvalues λ_j of the reference R, V_k the top-k eigenvectors of an estimate $\widehat{R}$, $F_{ij}^{(k)} = U_k^\top F_{ij} U_k$, and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. κ averages over 64 sampled output neurons; ρ_{diag} uses all of them.

Diagnostic	Question	Definition	Ideal
Block separability sep_k	Do all neuron blocks share one R?	$\frac{\sum_{i,j} \langle F_{ij}^{(k)}, \Lambda_k \rangle_F^2}{\|\Lambda_k\|_F^2 \sum_{i,j} \|F_{ij}^{(k)}\|_F^2}$	1
Captured energy eff_k	Does $\widehat{R}$ find high curvature?	$\text{tr}(V_k^\top R V_k) / \sum_{j \le k} \lambda_j$	1
Subspace overlap O_k	Does $\widehat{R}$ recover eigenvectors of R?	$\|U_k^\top V_k\|_F^2 / k$	1
Spectral mismatch ρ_k	Does $\widehat{R}$ reproduce curvature ratios?	$\text{std}_j \log \lambda_j \left(\Lambda_k^{-1/2} U_k^\top \widehat{R} U_k \Lambda_k^{-1/2} \right)$	0
Pairwise coupling κ	Are output neurons coupled pairwise?	$\text{mean}_{i \neq j} \|F_{ij}\|_F^2 / (\|F_{ii}\|_F \|F_{jj}\|_F)$	0
Diagonal spread ρ_{diag}	Are output scales equal?	$\text{std}_i \log L_{ii}$	0

AdamW. The curvature runs use a 2^{21}-token batch (2M), 600 steps, and hyperparameters tuned at this batch. We store models and optimizer states at steps 150, 300, and 450 for two seeds, average layers 0, 6, and 11 within each module, and report the mean and standard deviation over these six checkpoints. `fc` and `proj` denote the MLP up- and down-projections. Window experiments instead use batch-specific tuned settings and about 2.49B training tokens, and vary one window with all else fixed (section A).

Reference curvature. We backpropagate sampled labels on 2048 sequences and form the sequence gradients G_s, retaining all uses of a weight within a sequence. Their partial traces estimate R and L of eq. (1), and their neuron blocks estimate F_{ij}. Products of independent halves remove self-noise from quadratic terms (section B).

Diagnostics. Table 1 lists six diagnostics and the question each answers. Block separability tests the Kronecker premise. Captured energy, overlap, and spectral mismatch compare an input-factor estimate $\widehat{R}$ ($M^\top M$ for Muon, the stored factor for SOAP) with R at increasing strictness, up to the eigenvalue ratios that set whitening strengths. Pairwise coupling and diagonal spread test the two parts of $L \approx cI$.

Table 2: **Kronecker Structure in GPT-2.** Diagnostics of table 1 at the 2M batch, as means and standard deviations over three checkpoints and two seeds after averaging layers 0, 6, and 11. The SOAP row uses SOAP trajectories; all other rows use Muon trajectories. Arrows mark the better direction.

Metric	q	k	v	o	fc	proj
$\text{sep}_{16} \uparrow$	$.86 \pm .03$	$.60 \pm .05$	$.91 \pm .01$	$.98 \pm .00$	$.92 \pm .01$	$.95 \pm .01$
$\text{eff}_{16} \uparrow$, Muon	$.86 \pm .01$	$.77 \pm .04$	$.94 \pm .02$	$.81 \pm .05$	$.94 \pm .02$	$.86 \pm .01$
$\text{eff}_{16} \uparrow$, SOAP	$.98 \pm .01$	$.95 \pm .01$	$.98 \pm .02$	$.94 \pm .06$	$.99 \pm .01$	$.95 \pm .04$
$\kappa \downarrow$	$.004 \pm .001$	$.006 \pm .001$	$.005 \pm .002$	$.005 \pm .003$	$.001 \pm .001$	$.008 \pm .004$
$\rho_{\text{diag}} \downarrow$	$1.80 \pm .17$	$1.90 \pm .12$	$.74 \pm .06$	$.21 \pm .04$	$.68 \pm .17$	$.21 \pm .03$

5.2 Kronecker Structure and Factor Accuracy

The shared-input-factor premise holds well in most modules (table 2). Block separability is 0.91 to 0.98 for value, output, and MLP matrices, but 0.86 for query and 0.60 for key. The weakness of key lies in coupling: off-diagonal key blocks have separability 0.41 (section B.1). This is consistent with the coupled query/key dependence of attention (Ormaniec et al., 2024): the backpropagated signal of a key is weighted by the attention it receives, so different key neurons can see differently weighted inputs. The measurement does not, however, rule out better Kronecker fits. The ordering across modules is unchanged from the 0.5M to the 8M batch. Where this premise fails, no Kronecker method can represent the curvature, stored factors included.

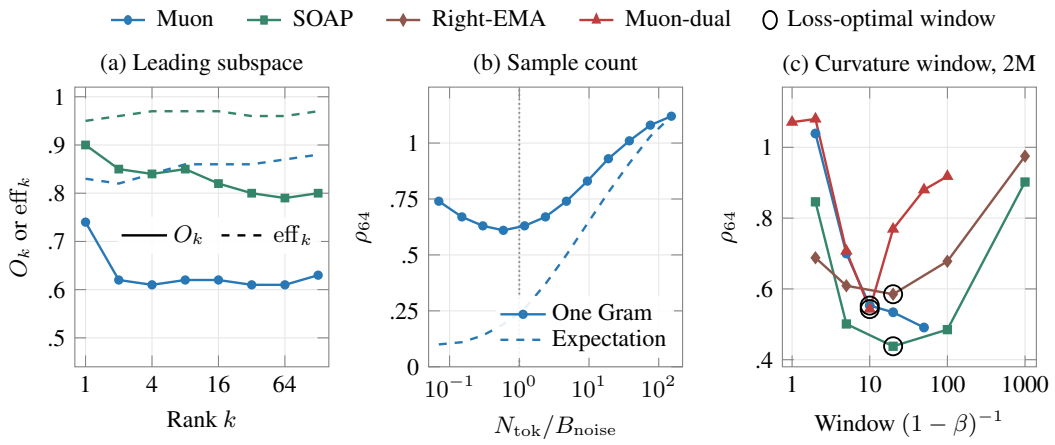


Figure 2: **Input-Factor Accuracy.** (a) Subspace overlap and captured energy of the top-k input directions at the 2M batch; each method is measured on its own trajectory. (b) Spectral mismatch at fixed SOAP weights for the Gram of one N_{tok}-token mean gradient and for its expectation in eq. (6); the dotted line marks $N_{\text{tok}} = B_{\text{noise}}$. (c) Spectral mismatch of each curvature estimate against its window at the 2M batch; circles mark the window with the lowest final loss. Muon-dual varies only the curvature window, with update momentum 0.9. Other batches are in table 6.

Muon's momentum Gram captures 0.77 to 0.94 of the best achievable leading energy, and SOAP's stored factor captures 0.94 to 0.99. The leading curvature directions are therefore present in the momentum without any accumulation. Since each method is measured on its own trajectory, the gap includes differences in the learned curvature and is not attributable to accumulation alone.

The output factor is nearly diagonal pairwise, with $\kappa = 0.001$ to 0.008 in every module. Its diagonal scales vary much more: ρ_{diag} is 0.21 in output and MLP down-projections but 1.80 to 1.90 in query and key, a one-standard-deviation spread of about a factor of six. This ordering matches the architectural prediction of section 4.3. Pairwise statistics nevertheless miss collective structure: summed over neurons at one checkpoint, off-diagonal entries hold 52% to 80% of the output factor's energy (section B.1).

The same pattern appears in other architectures (section E). In a pure Gated DeltaNet, key is again the least separable module (0.70), SOAP captures more energy than Muon in every module, and query and key spreads (0.42, 0.41) are smaller than in GPT-2, consistent with a role for RoPE. In ViT-Tiny, the fused qkv projection is the least separable (0.67) and has the largest spread (1.38).

5.3 Input-Factor Analysis

Figure 2a shows where Muon's input factor is accurate. Its overlap O_k is highest for the leading direction (0.74) and falls to about 0.61 by $k = 4$, while captured energy stays between 0.82 and 0.88 for all k. Muon thus recovers leading axes and mixes directions of similar curvature, as eq. (9) predicts, at little cost in energy. SOAP's stored factor has higher overlap on its own trajectory.

Figure 2b separates the two errors of section 4.2 at fixed weights. We form Grams of observed-label mean gradients over $N_{\text{tok}} = 16$K to 33.6M tokens and compare a single realization with the expectation of eq. (6), estimated from the same data. The expectation moves from R toward the signal Gram as the count grows (ρ_{64} from 0.10 to 1.12). The single realization is dominated by noise at small counts and reaches its lowest mismatch (0.61) near $N_{\text{tok}} \approx 0.6B_{\text{noise}}$, with B_{noise} converted to tokens; at large counts it approaches the expectation (1.08 versus 1.03 at $76B_{\text{noise}}$). Their agreement at the largest count is partly forced by the shared data stream. The gap between the two curves is the cost of forming one Gram without accumulation: a stored factor averages minibatch Grams, keeping the expectation at $N_{\text{eff}} = B_s$ while reducing this realization noise.

We then vary the curvature window during training (fig. 2c, table 6): the factor EMA β_2 in SOAP and Right-EMA, and the momentum β in Muon. For stored factors, the window with the lowest ρ_{64}

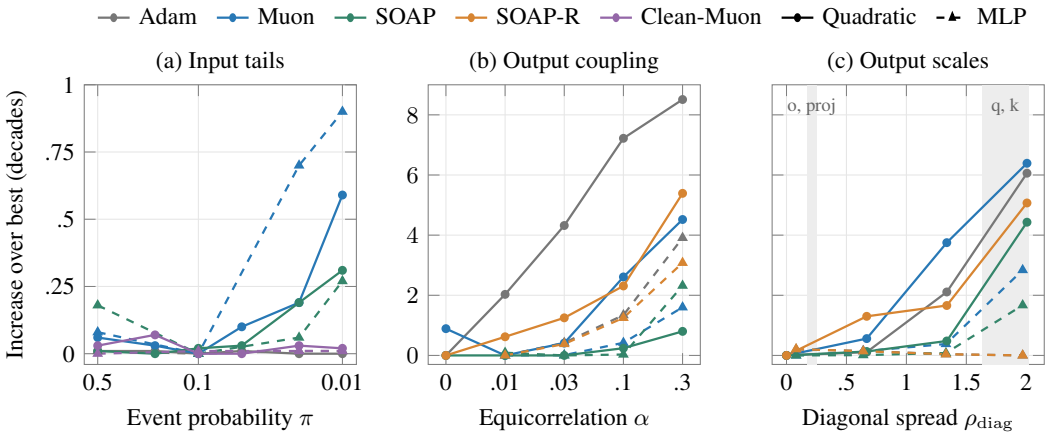


Figure 3: **Breaking One Approximation at a Time.** Increase in $\log_{10}(\mathcal{E}_T/\mathcal{E}_0)$ over each curve's best setting, where $\mathcal{E}_t$ is the excess loss at step t; lower is better. Solid lines are quadratic objectives (linear regression in (a), Kronecker quadratics in (b, c)) and dashed lines the nonlinear MLP. (a) Heavier input tails in low-variance directions at fixed input covariance, batch 16. (b) Equicorrelated output coupling at constant diagonal. (c) Unequal output scales indexed by ρ_{diag}; gray bands mark GPT-2 modules in table 2. Absolute values are in section D.

also has the lowest loss in all four 2M and 8M comparisons, although some neighboring losses lie within seed variation. The window with the highest overlap O_{64} does not; at 8M it costs 0.21 (SOAP) and 1.05 nats (Right-EMA). In standard Muon, one window sets both the curvature estimate and the update. There ρ_{64} keeps improving beyond the loss-optimal window: at 8M, raising β from 0.9 to 0.98 lowers ρ_{64} from 0.865 to 0.721 but raises the loss by 0.345 nats. At 0.5M, window differences for these methods mostly lie within seed variation (0.031 nats), so we do not rank them.

To decouple the two roles, Muon-dual keeps the update momentum at $\beta_u = 0.9$ and computes the whitening polynomial from a second momentum with window β_R (section C.1). At all three batches, $\beta_R = 0.9$ gives both the lowest ρ_{64} and the lowest loss, while the overlap-maximizing window loses 0.51, 0.37, and 1.34 nats at 0.5M, 2M, and 8M. Spectral accuracy thus again coincides with the loss optimum. Lengthening only the curvature window worsens both ρ_{64} and loss, so the long-window alignment of standard Muon is not a curvature-estimation gain. Because an update whitened by a different window is not an exact polar factor, mismatched windows may also disturb the trajectory; the coincidence is descriptive rather than causal.

6 Controlled Experiments

6.1 Setup

Problems. Each problem breaks one approximation while fixing the others. For the input factor, linear regression with output dimension 256, input dimension 64, and input covariance of condition number 10^4 fixes the curvature exactly; we change only the fourth moments of the eight lowest-variance input directions through a rare-event probability π, which alters gradient statistics but not the second moment. For the output factor, the quadratic $\mathcal{L}(\Delta) = \frac{1}{2}\operatorname{tr}(\Delta^\top L \Delta R)$ with $\Delta \in \mathbb{R}^{256\times 64}$ and fixed R of condition number 10^3 lets us vary either the diagonal spread of L or the equicorrelation α. A teacher-student MLP $f(x) = W_2 D\,\mathrm{GELU}(W_1 x)$ with dimensions $64 \to 256 \to 512$, trained only in W_1, provides nonlinear counterparts, with the output factor set through the fixed matrix D.

Methods and metric. We compare Adam, Muon, and SOAP, adding clean-Muon for the input experiments and SOAP-R for the output experiments. Clean-Muon applies Muon to the population gradient in linear regression and to a much larger batch in the MLP. SOAP-R rotates only the input coordinates and adapts each entry in that basis, so it can rescale rows but not mix outputs. All methods run 1000 cosine-decayed steps with a common Frobenius update norm, and learning rates

are selected per method and setting over three seeds (section D). We report $\log_{10}(\mathcal{E}_T/\mathcal{E}_0)$, where $\mathcal{E}_t$ is the excess loss at step t of $T = 1000$, and call its differences decades.

6.2 Input Estimation

With Gaussian inputs, Muon's excess loss is mostly estimation error. Its $\log_{10}$ loss ratio improves from -1.71 to -5.78 as the batch grows from 4 to 64 and reaches -8.10 with full gradients, while clean-Muon stays near -8.12 at every batch. In the MLP, the gap between Muon and clean-Muon shrinks from 6.4 to 0.7 decades as the batch grows from 4 to 1024. Since clean-Muon changes both the momentum and its Gram, these gaps do not isolate the preconditioner.

Heavier input tails affect optimization even at fixed curvature (fig. 3a). At batch 16 and $\pi = 0.01$, Muon lies 0.59 decades above its best setting in linear regression and 0.91 in the MLP, compared with 0.31 and 0.27 for SOAP; Adam and clean-Muon are unaffected. The sensitive estimator depends on the batch: at batch 64 in linear regression, Muon is unchanged while SOAP's ratio rises from -7.62 to -6.44. Gradient statistics therefore matter at fixed curvature, without a universal robustness ordering.

6.3 Output Scale and Coupling

Coupling defeats rowwise correction (fig. 3b). Raising the equicorrelation from 0 to 0.3 increases the condition number of L from 1 to about 111 at a constant diagonal. In the quadratic, Adam loses 8.5 decades and SOAP-R 5.4, while SOAP, which stores the full output factor, loses 0.8; Muon ends 4.5 decades above its best setting. The MLP shows the same failure of rowwise correction, ending 3.9 (Adam) and 3.1 (SOAP-R) decades above their best settings at $\alpha = 0.3$, although SOAP also degrades at this coupling (2.3) and Muon loses 1.6.

Unequal scales instead defeat Muon (fig. 3c). At $\rho_{\text{diag}} = 1.34$, a condition number of 100, Muon loses 3.7 decades in the quadratic, compared with 1.7 for SOAP-R and 0.5 for SOAP. In the MLP, Muon loses 2.8 decades at $\rho_{\text{diag}} = 1.97$, while SOAP-R and Adam lose none; Muon's selected rate lies at the edge of its quadratic grid in these cells (section D). On the same ρ_{diag} scale, GPT-2's query and key projections lie in the range where Muon degrades, and its output projections in the range where it does not. The toy losses do not transfer quantitatively, and the full spectrum, noise, and budget also matter.

7 Discussion and Conclusion

Muon admits a reading as an accumulation-free Kronecker-factored curvature approximation, with the input factor read from momentum and the output factor set to the identity. The signal-noise decomposition explains when the momentum Gram tracks leading curvature, and GPT-2 measurements show that it captures most leading curvature energy, with remaining errors in its spectrum and in the identity output factor. Controlled problems show that estimation noise, unequal output scales, and output coupling cause distinct failures. These results support a conditional view of Muon as a curvature method and indicate which correction each module needs.

Our analysis is static, but training dynamics matter. At one 0.5M seed, Muon's live momentum Grams have ρ_{64} of 0.54 to 0.87 across checkpoints, whereas Grams re-accumulated for 40 steps at frozen weights have 1.06 to 1.39 (section B.1). The momentum Gram along the trajectory is thus closer to the reference than a Gram formed at fixed weights, an effect outside the stationary analysis of section 4.2 that remains to be explained.

Limitations. Our reference is the GGN, not the exact Hessian. Stationarity, the empirical-Fisher correspondence, and the local quadratic are separate assumptions, and the fixed-checkpoint sweep does not establish the RAR mechanism along a trajectory. Methods are compared on their own trajectories, window experiments use one seed, and the main evidence comes from one small language model.

8 AI USE STATEMENT

In this work, we used generative AI tools to check mathematical derivations for potential errors and to assist with translation. We did not use them to develop theoretical frameworks, formulate hypotheses, design or provide feedback on experiments, implement methods, generate synthetic data, clean datasets, or interpret results. Proof assistance and qualitative data analysis were not applicable, as the paper contains no formal proofs or qualitative analyses.

Additionally, we used generative AI tools to assist with auxiliary code, including plotting, experiment scripts, log parsing, debugging, and job monitoring; the creation and refinement of figures and tables; and edits for grammar, readability, and formatting. The authors independently re-derived and verified all suggested mathematical corrections and reviewed all AI-assisted text, code, figures, and tables. We take full responsibility for the final content of this paper.

REFERENCES

Natalie Abreu, Nikhil Vyas, Sham Kakade, and Depen Morwani. The Potential of Second-Order Optimization for LLMs: A Study with Full Gauss-Newton. *arXiv preprint arXiv:2510.09378*, 2025. URL https://arxiv.org/abs/2510.09378.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, 1998. URL https://doi.org/10.1162/089976698300017746.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal Matrix Sign Methods and Their Application to the Muon Algorithm. *arXiv preprint arXiv:2505.16932*, 2025. URL https://arxiv.org/abs/2505.16932.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive Structured Gradient Optimization. *arXiv preprint arXiv:2503.20762*, 2025. URL https://arxiv.org/abs/2503.20762.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable Second Order Optimization for Deep Learning. *arXiv preprint arXiv:2002.09018*, 2020. URL https://arxiv.org/abs/2002.09018.

Artem Artemev, Rui Xia, Benjamin M. Boyd, Youjing Yu, Felix Dangel, Guillaume Hennequin, and Alberto Bernacchia. Exploiting weight-space symmetries for approximating curvature. *arXiv preprint arXiv:2606.00442*, 2026. URL https://arxiv.org/abs/2606.00442.

Lukas Balles and Philipp Hennig. Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients. In *International Conference on Machine Learning*, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Frederik Benzing. Gradient Descent on Neurons and its Link to Approximate Second-Order Optimization. *arXiv preprint arXiv:2201.12250*, 2022. URL https://arxiv.org/abs/2201.12250.

Jeremy Bernstein and Laker Newhouse. Modular Duality in Deep Learning. *arXiv preprint arXiv:2410.21265*, 2024a. URL https://arxiv.org/abs/2410.21265.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. *arXiv preprint arXiv:2409.20325*, 2024b. URL https://arxiv.org/abs/2409.20325.

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, Ruijie Zhang, Yao Lu, Yongxiang Liu, and Ganzhao Yuan. MuonEq: Balancing Before Orthogonalization with Lightweight Equilibration. *arXiv preprint arXiv:2603.28254*, 2026. URL https://arxiv.org/abs/2603.28254.

Tong Che. Second-Order Muon Done Right: A Principled Marriage of Spectral Geometry and Curvature. *arXiv preprint arXiv:2608.09763*, 2026. URL https://arxiv.org/abs/2608.09763.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon Optimizes Under Spectral Norm Constraints. *arXiv preprint arXiv:2506.15054*, 2025. URL https://arxiv.org/abs/2506.15054.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025. URL https://arxiv.org/abs/2512.04299.

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards Quantifying the Hessian Structure of Neural Networks. *arXiv preprint arXiv:2505.02809*, 2025. URL https://arxiv.org/abs/2505.02809.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. URL https://arxiv.org/abs/2010.11929.

Zhehang Du and Weijie Su. The Newton-Muon Optimizer. *arXiv preprint arXiv:2604.01472*, 2026. URL https://arxiv.org/abs/2604.01472.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Robert Lasenby, and Chris Olah. Privileged Bases in the Transformer Residual Stream. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/privileged-basis/index.html.

Runa Eschenhagen, Alexander Immer, Richard E. Turner, Frank Schneider, and Philipp Hennig. Kronecker-Factored Approximate Curvature for Modern Neural Network Architectures. *arXiv preprint arXiv:2311.00636*, 2023. URL https://arxiv.org/abs/2311.00636.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying Shampoo: Adapting Spectral Descent to Stochasticity and the Parameter Trajectory. *arXiv preprint arXiv:2602.09314*, 2026. URL https://arxiv.org/abs/2602.09314.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast Approximate Natural Gradient Descent in a Kronecker-factored Eigenbasis. *arXiv preprint arXiv:1806.03884*, 2018. URL https://arxiv.org/abs/1806.03884.

GLM-5-Team. GLM-5: from Vibe Coding to Agentic Engineering. *arXiv preprint arXiv:2602.15763*, 2026. URL https://arxiv.org/abs/2602.15763.

Wenbo Gong, Meyer Scetbon, Chao Ma, and Edward Meeds. Towards Efficient Optimizer Design for LLM via Structured Fisher Approximation with a Low-Rank Extension. *arXiv preprint arXiv:2502.07752*, 2025. URL https://arxiv.org/abs/2502.07752.

Diego Granziol and Khurshid Juarev. Hessian Spectral Analysis at Foundation Model Scale. *arXiv preprint arXiv:2602.00816*, 2026. URL https://arxiv.org/abs/2602.00816.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. In *International Conference on Machine Learning*, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient Descent Happens in a Tiny Subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL https://arxiv.org/abs/1812.04754.

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/.

Kimi Team. Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

Kimi Team. Kimi K3: Open Frontier Intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

Dmitry Kovalev. Understanding Gradient Orthogonalization for Deep Learning via Non-Euclidean Trust-Region Optimization. *arXiv preprint arXiv:2503.12645*, 2025. URL https://arxiv.org/abs/2503.12645.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the Empirical Fisher Approximation for Natural Gradient Descent. *arXiv preprint arXiv:1905.12558*, 2019. URL https://arxiv.org/abs/1905.12558.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise First, Orthogonalize Later: Understanding Momentum in Muon via Spectral Filtering. *arXiv preprint arXiv:2606.03899*, 2026. URL https://arxiv.org/abs/2606.03899.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025. URL https://arxiv.org/abs/2510.05491.

Wu Lin, Scott C. Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger B. Grosse. Understanding and Improving Shampoo and SOAP via Kullback-Leibler Minimization. *arXiv preprint arXiv:2509.03378*, 2025. URL https://arxiv.org/abs/2509.03378.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, et al. Muon is Scalable for LLM Training. *arXiv preprint arXiv:2502.16982*, 2025a. URL https://arxiv.org/abs/2502.16982.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. COSMOS: A Hybrid Adaptive Optimizer for Memory-Efficient Training of LLMs. *arXiv preprint arXiv:2502.17410*, 2025b. URL https://arxiv.org/abs/2502.17410.

Yifeng Liu, Angela Yuan, and Quanquan Gu. MARS-M: When Variance Reduction Meets Matrices. *arXiv preprint arXiv:2510.21800*, 2025c. URL https://arxiv.org/abs/2510.21800.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 2020. URL https://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International Conference on Machine Learning*, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An Empirical Model of Large-Batch Training. *arXiv preprint arXiv:1812.06162*, 2018. URL https://arxiv.org/abs/1812.06162.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A New Perspective on Shampoo's Preconditioner. *arXiv preprint arXiv:2406.17748*, 2024. URL https://arxiv.org/abs/2406.17748.

Hiroki Naganuma, Shagun Gupta, Youssef Briki, Ioannis Mitliagkas, Irina Rish, Parameswaran Raman, and Hao-Jun Michael Shi. Adaptive Batch Sizes Using Non-Euclidean Gradient Noise Scales for Stochastic Sign and Spectral Descent. *arXiv preprint arXiv:2602.03001*, 2026. URL https://arxiv.org/abs/2602.03001.

Nikhil Nayak, Julia White, Urchade Zaratiana, Kelton Zhang, Henrijs Princis, Dhruv Atreja, Henry Fawcett, Matthew Thomas, George Hurn-Maloney, and Ash Lewis. Correcting Stochastic Update Bias in Preconditioned Language Model Optimizers. *arXiv preprint arXiv:2605.20756*, 2026. URL https://arxiv.org/abs/2605.20756.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral Flattening Is All Muon Needs: How Orthogonalization Controls Learning Rate and Convergence. *arXiv preprint arXiv:2605.13079*, 2026. URL https://arxiv.org/abs/2605.13079.

Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What Does It Mean to Be a Transformer? Insights from a Theoretical Hessian Analysis. *arXiv preprint arXiv:2410.10986*, 2024. URL https://arxiv.org/abs/2410.10986.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*, 2024. URL https://arxiv.org/abs/2406.17557.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training Deep Learning Models with Norm-Constrained LMOs. *arXiv preprint arXiv:2502.07529*, 2025. URL https://arxiv.org/abs/2502.07529.

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). *arXiv preprint arXiv:2505.13416*, 2025. URL https://arxiv.org/abs/2505.13416.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence Bound and Critical Batch Size of Muon Optimizer. *arXiv preprint arXiv:2507.01598*, 2025. URL https://arxiv.org/abs/2507.01598.

Nicol N. Schraudolph. Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent. *Neural Computation*, 14(7):1723–1738, 2002. URL https://doi.org/10.1162/08997660260028683.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via Stochastic Frank-Wolfe under Heavy-Tailed Noise. *arXiv preprint arXiv:2506.04192*, 2025. URL https://arxiv.org/abs/2506.04192.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the Convergence Analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025. URL https://arxiv.org/abs/2505.23737.

Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A Distributed Data-Parallel PyTorch Implementation of the Distributed Shampoo Optimizer for Training Neural Networks At-Scale. *arXiv preprint arXiv:2309.06497*, 2023. URL https://arxiv.org/abs/2309.06497.

Chongjie Si, Debing Zhang, and Wei Shen. AdaMuon: Adaptive Muon Optimizer. *arXiv preprint arXiv:2507.11005*, 2025. URL https://arxiv.org/abs/2507.11005.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1757af1fe1429801bdf3abf5600f8bba-Abstract-Conference.html.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Weijie Su. Isotropic Curvature Model for Understanding Deep Learning Optimization: Is Gradient Orthogonalization Optimal? *arXiv preprint arXiv:2511.00674*, 2025. URL https://arxiv.org/abs/2511.00674.

Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3503–3513, 2020. URL https://proceedings.mlr.press/v108/thomas20a.html.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *arXiv preprint arXiv:1905.09418*, 2019. URL https://arxiv.org/abs/1905.09418.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in Large Language Models: Dead, N-gram, Positional. *arXiv preprint arXiv:2309.04827*, 2023. URL https://arxiv.org/abs/2309.04827.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and Stabilizing Shampoo using Adam for Language Modeling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e988664070e9591f93fdcf605f7dc623-Abstract-Conference.html.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon Outperforms Adam: A Curvature Perspective. *arXiv preprint arXiv:2606.04662*, 2026. URL https://arxiv.org/abs/2606.04662.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting Hessian: Understanding Common Structure of Hessian in Neural Networks. *arXiv preprint arXiv:2010.04261*, 2020. URL https://arxiv.org/abs/2010.04261.

Chenrui Xu, Wenjing Yan, and Ying-Jun Angela Zhang. FISMO: Fisher-Structured Momentum-Orthogonalized Optimizer. *arXiv preprint arXiv:2601.21750*, 2026. URL https://arxiv.org/abs/2601.21750.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. *arXiv preprint arXiv:2412.06464*, 2024. URL https://arxiv.org/abs/2412.06464.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George E. Dahl, Christopher J. Shallue, and Roger Grosse. Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Quadratic Model. *arXiv preprint arXiv:1907.04164*, 2019. URL https://arxiv.org/abs/1907.04164.

Hongtao Zhang, Wenjie Zhou, Wei Chen, and Xueqi Cheng. When and Why Grouping Attention Heads Accelerates Muon Optimization. *arXiv preprint arXiv:2605.08933*, 2026a. URL https://arxiv.org/abs/2605.08933.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the Geometry of Muon with Curvature-Aware Preconditioning. *arXiv preprint arXiv:2603.09697*, 2026b. URL https://arxiv.org/abs/2603.09697.

Yikuan Zhang, Ning Yang, and Yuhai Tu. On the Superlinear Relationship between SGD Noise Covariance and Loss Landscape Curvature. *arXiv preprint arXiv:2602.05600*, 2026c. URL https://arxiv.org/abs/2602.05600.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use Fewer Learning Rates To Gain More. *arXiv preprint arXiv:2406.16793*, 2024. URL https://arxiv.org/abs/2406.16793.

A Experimental Details

Language model. The GPT-2-style model has 12 layers, width 768, 12 heads, a 50304-token vocabulary, RoPE with base 10000, RMSNorm ($\epsilon = 10^{-5}$), no linear biases, and untied embeddings; it is trained from scratch on the FineWeb 10BT sample. All 72 hidden matrices use the compared

optimizer, and we measure 18 of them (six module types in layers 0, 6, and 11). The input embedding, readout, and normalization parameters use AdamW with learning rate 0.01, betas $(0.95, 0.95)$, and weight decay 0, 0.001, and 0, respectively, fixed across methods. Gradients are clipped at norm 1, and batches are accumulated from per-GPU microbatches of 32 sequences on eight GPUs. The batches denoted 0.5M, 2M, and 8M are 2^{19}, 2^{21}, and 2^{23} tokens.

Curvature runs. These use 600 steps, 12 warmup steps, and cosine decay to 10% of the peak rate. Muon uses learning rate 0.03619, weight decay 0.0562, and momentum 0.9; SOAP uses learning rate 0.005, weight decay 0.562, momentum and factor decay 0.95, and a basis refresh every two steps. Learning rate and weight decay were tuned on multiplicative grids around these values. Seeds 1339 and 2024 are stored at steps 150, 300, and 450.

A.1 Optimizer Implementations

Muon uses the spectral-descent path of Distributed Shampoo (Shi et al., 2023): after Frobenius normalization $X_0 = M/\|M\|_F$, it applies five iterations $X_{j+1} = 3.4445X_j - 4.7750X_jX_j^\top X_j + 2.0315(X_jX_j^\top)^2X_j$, transposing tall matrices. This polynomial approximates the polar factor on the nonzero singular support; bias correction rescales the momentum and leaves the polar direction unchanged. Right-EMA is the input-only, factored-eigenvalue path of SOAP: it stores an EMA of $G^\top G$, rotates only the input basis, and adapts separate one-dimensional scale vectors for the rotated input and unrotated output coordinates. It therefore restores output scales but no output rotation. The controlled SOAP-R instead keeps a full matrix of elementwise second moments in the input eigenbasis.

A.2 Effective Sample Count

For independent $G_s = \mu + \varepsilon_s$ with centered noise, every cross-noise term of $\mathbb{E}[M^\top M]$ vanishes, giving $\mathbb{E}[M^\top M] = \mu^\top\mu + (\sum_s a_s^2)\Sigma$ and hence eq. (6); substituting $\Sigma = C - \mu^\top\mu$ gives eq. (7). With bias-corrected EMA over t equal minibatches, $N_{\text{eff}}(t) = B_s\frac{1+\beta}{1-\beta}\frac{1-\beta^t}{1+\beta^t}$, which tends to the steady-state value. Correlated gradients and minibatch clipping break this independent-averaging assumption.

B Curvature Estimation

Reference. The reference uses 256 batches of eight sequences (2^{21} tokens). Labels are sampled from the model's predictive distribution at every position, and each sequence contributes one G_s summing all uses of the weight. Summing per-position outer products instead would drop cross-position terms. No independence between activations and backpropagated gradients is assumed.

Estimators. Alternating batches form two independent halves. Quadratic block energies use products across the halves, which removes self-noise; separability uses one half for its numerator and the other for the basis U_k. Ratios, logarithms, and estimated bases retain finite-sample error, so corrected separability can slightly exceed one. Eigenvalues are floored at 10^{-30}, and standard deviations use denominator $k-1$. Chance overlap of random k-subspaces is k/n. Captured energy is normalized by the best rank-k energy, not by $\operatorname{tr}(R)$. Spectral mismatch is restricted to the leading reference subspace and does not measure the full inverse-root action. Under an exact Kronecker model, the summand of κ equals $L_{ij}^2/(L_{ii}L_{jj})$.

B.1 Additional Structure Measurements

Table 3 repeats the main diagnostics at other batches and restricts separability to off-diagonal blocks. The module ordering is unchanged across batches, and key remains the weakest module. Pairwise coupling can hide aggregate coupling: at one 0.5M Muon checkpoint (step 450, seed 1339), off-diagonal entries hold $0.63, 0.80, 0.59, 0.68, 0.52$, and 0.63 of $\|L\|_F^2$ for q, k, v, o, fc, and proj, with normalized effective ranks $(\operatorname{tr} L)^2/(m\|L\|_F^2)$ of $0.17, 0.07, 0.31, 0.31, 0.39$, and 0.36. Optimizer trajectories also matter: at the same seed, Muon's stored momentum Grams have ρ_{64} of 0.54 to 0.87

Table 3: **Structure Across Batch Sizes.** The aggregation follows table 2. Off-diagonal separability restricts the block sum to distinct neurons. All runs use the fixed 600-step curvature protocol, so token budgets differ across batches.

Metric (batch)	q	k	v	o	fc	proj
sep_{16} (2M)	$.86 \pm .03$	$.60 \pm .05$	$.91 \pm .01$	$.98 \pm .00$	$.92 \pm .01$	$.95 \pm .01$
sep_{16} (0.5M)	$.84 \pm .03$	$.53 \pm .05$	$.91 \pm .02$	$.97 \pm .01$	$.91 \pm .02$	$.93 \pm .01$
sep_{16} (8M)	$.87 \pm .01$	$.59 \pm .05$	$.92 \pm .01$	$.99 \pm .01$	$.92 \pm .01$	$.97 \pm .01$
sep_{16}, off-diagonal (2M)	$.70 \pm .04$	$.41 \pm .05$	$.83 \pm .04$	$.87 \pm .04$	$.76 \pm .08$	$.76 \pm .09$
sep_{16}, off-diagonal (0.5M)	$.69 \pm .04$	$.34 \pm .05$	$.85 \pm .05$	$.88 \pm .01$	$.74 \pm .11$	$.75 \pm .06$
sep_{16}, off-diagonal (8M)	$.67 \pm .05$	$.41 \pm .07$	$.82 \pm .02$	$.85 \pm .05$	$.75 \pm .06$	$.80 \pm .07$
eff_{16}, Muon (2M)	$.86 \pm .01$	$.77 \pm .04$	$.94 \pm .02$	$.81 \pm .05$	$.94 \pm .02$	$.86 \pm .01$
eff_{16}, Muon (0.5M)	$.88 \pm .01$	$.77 \pm .04$	$.95 \pm .02$	$.84 \pm .03$	$.95 \pm .02$	$.85 \pm .02$
eff_{16}, Muon (8M)	$.86 \pm .01$	$.80 \pm .02$	$.93 \pm .03$	$.77 \pm .05$	$.93 \pm .03$	$.86 \pm .03$
eff_{16}, SOAP (2M)	$.98 \pm .01$	$.95 \pm .01$	$.98 \pm .02$	$.94 \pm .06$	$.99 \pm .01$	$.95 \pm .04$
eff_{16}, SOAP (0.5M)	$.98 \pm .01$	$.95 \pm .01$	$.98 \pm .02$	$.95 \pm .05$	$.98 \pm .02$	$.94 \pm .05$
eff_{16}, SOAP (8M)	$.98 \pm .01$	$.96 \pm .01$	$.98 \pm .02$	$.94 \pm .06$	$.98 \pm .01$	$.95 \pm .03$

Table 4: **Accuracy by Rank.** Columns give rank k at the 2M batch. Means and standard deviations use six checkpoint/seed combinations, after averaging the 18 measured matrices.

Statistic	1	2	4	8	16	32	64	128
O_k Muon	$.74 \pm .07$	$.62 \pm .08$	$.61 \pm .05$	$.62 \pm .03$	$.62 \pm .01$	$.61 \pm .00$	$.61 \pm .01$	$.63 \pm .02$
O_k SOAP	$.90 \pm .04$	$.85 \pm .09$	$.84 \pm .08$	$.85 \pm .08$	$.82 \pm .08$	$.80 \pm .09$	$.79 \pm .09$	$.80 \pm .08$
eff_k Muon	$.83 \pm .03$	$.82 \pm .04$	$.84 \pm .03$	$.86 \pm .02$	$.86 \pm .01$	$.86 \pm .01$	$.87 \pm .00$	$.88 \pm .01$
eff_k SOAP	$.95 \pm .02$	$.96 \pm .03$	$.97 \pm .03$	$.97 \pm .02$	$.97 \pm .03$	$.96 \pm .03$	$.96 \pm .03$	$.97 \pm .02$

across checkpoints, versus 1.06 to 1.39 after 40 steps of re-accumulation at frozen weights, and SOAP's frozen re-accumulation raises O_{64} from 0.72 to 0.84 at step 150.

C Input-Factor Experiments

Sample sweep. At a SOAP checkpoint (0.5M, step 2000), we compute 2048 observed-label gradients of 16384-token microbatches and average them hierarchically to $N_{\mathrm{tok}} = 16384 \cdot 2^j$ tokens, $j = 0, \ldots, 11$. Single-realization values average up to 12 disjoint realizations. The expectation is $S_{\mathrm{sig}} + \Sigma_{16K}/2^j$, with signal Gram S_{sig} and microbatch covariance Σ_{16K} from the same stream. Counts are normalized by $B_{\mathrm{noise}} = 16384\,\mathrm{tr}(\Sigma_{16K})/\,\mathrm{tr}(S_{\mathrm{sig}})$ tokens, the median over matrices. A Muon checkpoint (0.5M, step 450) gives the same pattern.

C.1 Window Experiments

The 0.5M, 2M, and 8M batches train for 4750, 1188, and 297 steps (about 2.49B tokens) with 2% warmup, cosine decay to 10%, and seed 1339; we report the final validation loss over 80 minibatches. Window diagnostics use a 1024-sequence sampled-label reference. Base learning rate and weight decay are, for 0.5M/2M/8M, Muon $0.0192/0.0562$, $0.03619/0.0562$, $0.0384/0.0562$; SOAP $0.00186/0.562$, $0.005/0.562$, $0.005/1.78$; and Right-EMA $0.0025/0.316$, $0.005/1.0$, $0.003751/3.16$. Basis refresh periods are 10, 2, and 1 steps, and stored-factor update momentum is 0.95 at 0.5M and 2M and 0.9 at 8M. Standard Muon varies $\beta \in \{0.5, 0.8, 0.9, 0.95, 0.98\}$; SOAP and Right-EMA vary $\beta_2 \in \{0.5, 0.8, 0.95, 0.99, 0.999\}$, and Muon-dual varies $\beta_R \in \{0, 0.5, 0.8, 0.9, 0.95, 0.98, 0.99\}$. Repeated Muon seeds differ in loss by about 0.031, 0.011, and 0.049 to 0.065 nats at 0.5M, 2M, and 8M, so smaller differences are not ranked.

Muon-dual keeps an update momentum M_u ($\beta_u = 0.9$) and a curvature momentum M_R (β_R). Its update is proportional to $\frac{M_u}{\|M_u\|_F}\,\psi\big(M_R^\top M_R/\|M_R\|_F^2\big)$ with $\psi(\lambda) = \phi(\sqrt{\lambda})/\sqrt{\lambda}$, where ϕ is the scalar five-step Newton-Schulz polynomial, and its Frobenius norm matches the standard update of M_u. It reduces to standard Muon when $\beta_R = \beta_u$, and the two runs differ by at most 0.002 nats. For $\beta_R \neq \beta_u$ the update is not an exact polar factor.

Table 5: **Fixed-Checkpoint Sample Sweep.** SOAP checkpoint at 0.5M batch, step 2000. Single-realization entries average up to 12 disjoint samples; expectation estimates use the same finite gradient stream. Normalized counts are medians across matrices.

N_{tok} (M)	$N_{\text{tok}}/B_{\text{noise,tok}}$	O_{64} single	O_{64} expected	ρ_{64} single	ρ_{64} expected
0.016	0.07	.58	.97	.74	.10
0.033	0.15	.60	.97	.67	.11
0.066	0.30	.62	.96	.63	.14
0.131	0.59	.64	.96	.61	.19
0.262	1.19	.65	.95	.63	.26
0.524	2.37	.66	.93	.67	.37
1.049	4.74	.67	.90	.74	.50
2.097	9.48	.69	.86	.83	.64
4.194	18.96	.70	.82	.93	.78
8.389	37.92	.71	.77	1.01	.91
16.777	75.85	.71	.74	1.08	1.03
33.554	151.70	.72	.72	1.12	1.12

Table 6: **Curvature-Window Sweep.** Entries are final-loss gap to the best window of the same method and batch (nats) / ρ_{64} / O_{64}, measured at three quarters of training with one seed and a 1024-sequence reference. Muon-dual varies β_R at update momentum $\beta_u = 0.9$.

Method	Window	0.5M	2M	8M
Muon	.5	+.074 / .693 / .597	+.161 / 1.039 / .615	+.379 / 1.458 / .688
Muon	.8	+.015 / .582 / .585	+.024 / .700 / .596	+.066 / 1.118 / .636
Muon	.9	+.002 / .561 / .585	.000 / .553 / .599	.000 / .865 / .613
Muon	.95	.000 / .553 / .586	+.002 / .534 / .594	+.051 / .761 / .602
Muon	.98	+.010 / .524 / .587	+.050 / .491 / .610	+.345 / .721 / .576
SOAP	.5	+.007 / .470 / .711	+.449 / .846 / .698	+3.425 / 2.652 / .633
SOAP	.8	.000 / .347 / .789	+.023 / .501 / .810	+.210 / .827 / .746
SOAP	.95	+.003 / .273 / .875	.000 / .438 / .869	.000 / .763 / .699
SOAP	.99	+.008 / .261 / .914	+.004 / .485 / .883	+.045 / .928 / .519
SOAP	.999	+.015 / .428 / .856	+.020 / .902 / .653	+.070 / 1.075 / .475
Right-EMA	.5	+.001 / .472 / .727	+.068 / .688 / .783	+1.054 / 1.116 / .845
Right-EMA	.8	+.001 / .408 / .798	+.006 / .609 / .828	.000 / .884 / .813
Right-EMA	.95	+.002 / .311 / .880	.000 / .585 / .871	+.054 / .934 / .816
Right-EMA	.99	.000 / .292 / .930	+.005 / .678 / .875	+.158 / 1.080 / .698
Right-EMA	.999	+.004 / .475 / .841	+.021 / .975 / .666	+.175 / 1.221 / .645
Muon-dual	0	+.077 / .744 / .646	+.261 / 1.071 / .664	+1.304 / 2.055 / .832
Muon-dual	.5	+.100 / .807 / .663	+.373 / 1.080 / .682	+1.341 / 2.001 / .835
Muon-dual	.8	+.034 / .638 / .650	+.063 / .707 / .671	+.515 / 1.110 / .719
Muon-dual	.9	.000 / .552 / .590	.000 / .543 / .597	.000 / .912 / .611
Muon-dual	.95	+.512 / .681 / .678	+.394 / .769 / .571	+.603 / 1.212 / .637
Muon-dual	.98	+.627 / .828 / .593	+.414 / .880 / .510	+.756 / 1.159 / .655
Muon-dual	.99	+.531 / .850 / .545	+.339 / .918 / .515	+.780 / 1.272 / .638

D CONTROLLED-PROBLEM DETAILS

Problems. Linear regression uses 50000 training samples and a randomly rotated input covariance of condition number 10^4. In the eight lowest-variance directions, Gaussian coordinates are multiplied by Bernoulli-π masks and divided by $\sqrt{\pi}$, and each input set is then transformed to match the covariance exactly, so the objective is identical across π. The quadratics use R with condition number 10^3 and a random basis, diagonal L with log-spaced eigenvalues of condition number 1 to 10^3, or equicorrelation $\alpha \in \{0, 0.01, 0.03, 0.1, 0.3\}$, whose condition number is $(1 + 255\alpha)/(1 - \alpha)$. Quadratic gradients receive Gaussian noise with per-entry standard deviation $0.1\|\nabla\mathcal{L}\|_F/\sqrt{mn}$. The MLP perturbs the teacher's W_1 (W_2 and D are fixed and shared) with Gaussian noise; it uses 50000 training and 4096 held-out inputs, input condition number 10^3 for the tail experiment (batches

Table 7: **Input Tails.** Numerical values for fig. 3; entries are $\log_{10}(\mathcal{E}_T/\mathcal{E}_0)$ with lower better. Selection and normalization follow section D.

Problem	π	Adam	Muon	SOAP	Clean-Muon
Linear	0.5	−1.25	−4.07	−4.72	−8.11
Linear	0.2	−1.25	−4.1	−4.73	−8.07
Linear	0.1	−1.26	−4.13	−4.71	−8.14
Linear	0.05	−1.25	−4.03	−4.7	−8.14
Linear	0.02	−1.26	−3.94	−4.54	−8.11
Linear	0.01	−1.26	−3.54	−4.42	−8.12
MLP	0.5	−3.63	−6.78	−5.83	−10.98
MLP	0.1	−3.64	−6.86	−6.01	−10.97
MLP	0.02	−3.63	−6.16	−5.95	−10.97
MLP	0.01	−3.63	−5.96	−5.74	−10.97

Table 8: **Output Coupling.** Numerical values for fig. 3; entries are $\log_{10}(\mathcal{E}_T/\mathcal{E}_0)$ with lower better. Selection and normalization follow section D.

Problem	α	Adam	Muon	SOAP	SOAP-R
Quadratic	0	−9.78	−8.27	−9.66	−10.92
Quadratic	0.01	−7.75	−9.16	−9.66	−10.3
Quadratic	0.03	−5.46	−8.75	−9.66	−9.67
Quadratic	0.1	−2.56	−6.55	−9.42	−8.61
Quadratic	0.3	−1.27	−4.64	−8.86	−5.53
MLP	0.01	−8.91	−9	−9.06	−8.95
MLP	0.03	−8.49	−8.99	−9.15	−8.57
MLP	0.1	−7.57	−8.58	−9.12	−7.7
MLP	0.3	−5	−7.4	−6.83	−5.87

4 to 1024) and 10 for the output experiments (batch 64), with D of diagonal spread 1 to 30 or equicorrelated $D^\top D$. We report the measured ρ_{diag} of its output GGN factor.

Optimization. All methods use heavy-ball momentum 0.9, 1000 steps, cosine decay to zero, no weight decay, and update Frobenius norm $\sqrt{\min(m,n)}$. Second moments decay at 0.99 (0.95 for quadratic SOAP and SOAP-R), and bases refresh every 20 steps (10 in the quadratics). Learning rates come from grids of 8 (linear), 5 (quadratic), and 7 (MLP) log-spaced values; linear and MLP results use the rate with the lowest median over seeds 0 to 2, whereas quadratic results select a rate per seed before taking the median, an optimistic convention. In the quadratics, Muon selects the largest rate multiplier at $\alpha \geq 0.1$ and at diagonal condition numbers of at least 10, so its losses there may overstate the degradation; at condition number 1000 all methods select the largest rate. Clean-Muon in the MLP samples $\min(64B, 50000)$ examples per step and thus retains some noise. Toy SOAP and SOAP-R do not transport second moments across basis refreshes and are simplified variants.

E Additional Architectures

ViT-Tiny. The model (Dosovitskiy et al., 2020) has width 192 and 12 layers and uses 64-pixel Tiny-ImageNet-200 images with patch size 4 (257 tokens). It is trained for 300 epochs at batch 256 with RandAugment, Mixup/CutMix, and label smoothing; Muon uses momentum 0.95, and SOAP uses momentum and factor decay 0.95 with refresh every 10 steps. The checkpoints (iterations 8000, 30000, and 90000, one seed) lack optimizer states, so we resume training for 100 steps to rebuild them. The reference uses 16384 sampled-label images with training augmentation. The qkv projection is one fused 576×192 matrix and is not directly comparable to separate q, k, and v.

Gated DeltaNet. The pure Gated DeltaNet (Yang et al., 2024) has width 768 and 12 layers and is trained on FineWeb with 2^{23} tokens per step for 294 steps. Muon follows the original implementation (Nesterov momentum without bias correction, bf16 Newton-Schulz) with learning rate 0.0384, weight decay 0.562, and momentum 0.95, and its captured energy is measured on the momentum buffer;

Table 9: **Output Scales.** Numerical values for fig. 3; entries are $\log_{10}(\mathcal{E}_T/\mathcal{E}_0)$ with lower better. Selection and normalization follow section D.

Problem	ρ_{diag}	Adam	Muon	SOAP	SOAP-R
Quadratic	0	-9.78	-8.27	-9.66	-10.92
Quadratic	0.669	-9.67	-7.71	-9.53	-9.62
Quadratic	1.34	-7.67	-4.52	-9.18	-9.26
Quadratic	2.01	-3.72	-1.88	-5.23	-5.85
MLP	0.08	-8.87	-8.93	-8.98	-8.91
MLP	0.64	-8.92	-8.79	-8.97	-8.98
MLP	1.33	-9.03	-8.55	-8.9	-9.08
MLP	1.97	-9.07	-6.09	-7.31	-9.12

Table 10: **Kronecker Structure in ViT-Tiny.** Metrics follow table 2. Means and standard deviations over three checkpoints of one seed, with fused qkv and reconstructed optimizer states.

Metric	qkv	proj	fc1	fc2
sep_{16}	$.67 \pm .10$	$.88 \pm .04$	$.82 \pm .01$	$.75 \pm .12$
eff_{16}, Muon	$.90 \pm .03$	$.88 \pm .04$	$.96 \pm .00$	$.77 \pm .07$
eff_{16}, SOAP	$.99 \pm .01$	$.98 \pm .01$	$.99 \pm .00$	$.99 \pm .01$
κ	$.035 \pm .010$	$.017 \pm .007$	$.011 \pm .003$	$.026 \pm .012$
ρ_{diag}	$1.38 \pm .27$	$.28 \pm .09$	$.87 \pm .10$	$.13 \pm .01$

SOAP uses 0.004, 3.16, betas $0.95/0.95$, and refresh every two steps. States are stored at steps 74, 147, and 221 for two seeds. As in GPT-2, SOAP captures more energy on its own trajectory, and output scales are least uniform in fc. Query and key spreads are smaller than in GPT-2; these mixers use headwise normalization without RoPE, but several architectural features change at once.

Table 11: **Kronecker Structure in Pure Gated DeltaNet.** Metrics follow table 2. Means and standard deviations over three checkpoints and two seeds. The model uses only gated delta-rule mixers.

Metric	q	k	v	o	fc	proj
sep_{16}	$.88 \pm .01$	$.70 \pm .06$	$.84 \pm .03$	$.96 \pm .02$	$.87 \pm .03$	$.90 \pm .01$
eff_{16}, Muon	$.84 \pm .01$	$.83 \pm .02$	$.86 \pm .02$	$.65 \pm .03$	$.89 \pm .03$	$.73 \pm .03$
eff_{16}, SOAP	$.96 \pm .01$	$.96 \pm .01$	$.95 \pm .01$	$.86 \pm .06$	$.96 \pm .02$	$.92 \pm .03$
κ	$.005 \pm .002$	$.002 \pm .001$	$.002 \pm .001$	$.006 \pm .001$	$.002 \pm .001$	$.008 \pm .001$
ρ_{diag}	$.42 \pm .10$	$.41 \pm .06$	$.34 \pm .03$	$.09 \pm .01$	$.67 \pm .04$	$.08 \pm .01$