ORTHOGONALIZATION SEPARATES SHARPNESS EXPLORATION FROM PERTURBED GRADIENT UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

In Sharpness-Aware Minimization (SAM), the perturbed gradient updates the model parameters, while the clean gradient determines the parameter perturbation. *This raises a natural question: does matrix orthogonalization have different effects when applied to the parameter perturbation versus the subsequent parameter update?* We thus investigate how SAM changes when matrix orthogonalization is introduced separately into these two stages: 1) for clean gradient, orthogonalization changes the first order sharpness term from a Frobenius norm quantity to a normalized nuclear norm quantity, providing a nuclear norm view in sharpness exploration; 2) for perturbed gradient, orthogonalization changes how gradient variation caused by the perturbation enters the final update. We further analyze the effects of finite step Newton–Schulz approximation and global normalization on these two mechanisms. With the above analysis, MuonSAM instantiates this construction by orthogonalizing both the clean and perturbed gradients. Experiments on vision and language tasks verify the theoretical predictions and provide evidence for the roles of orthogonalization in sharpness exploration and model updates.

1 INTRODUCTION

Background: Sharpness-Aware Minimization (SAM) modifies training through two sequential operations Foret et al. (2021): it constructs a perturbation to select where the loss is evaluated, then computes the gradient at that perturbed point and passes it to the outer optimizer. Existing variants modify different parts of this process. ASAM adapts the neighborhood to parameter scaling Kwon et al. (2021), while Fisher SAM defines it through information geometry Kim et al. (2022). Momentum-SAM changes the temporal source of the probe Becker et al. (2025). ED-SAM adjusts perturbations according to energy and corrects the update direction to change how clean and perturbed gradients enter the update Ye et al. (2026). These methods highlight two design choices in SAM: how to construct the perturbation neighborhood and how to form the outer update.

Observation: Muon transforms momentum matrices into approximate polar directions for matrix parameter updates Jordan et al. (2024); Kim & Oh (2026). Prior analyses connect this geometry to spectral norm optimization and implicit bias Chen et al. (2026); Fan et al. (2025). Unlike the raw matrix gradient, whose singular mode weights scale with the singular values, *the polar factor assigns equal directional weight to every active singular mode*. This suggests two places to apply matrix orthogonalization in SAM training: to the probe that selects the evaluation point and to the direction used to form the outer update. The probe determines *where the loss is evaluated*; the outer direction determines *how the gradient measured there becomes an update*. These roles affect different stages of SAM, so it remains unclear how matrix orthogonalization changes each stage and their interaction.

Question: *How does matrix orthogonalization affect SAM when used to select the sharpness probe and to form the outer update from the perturbed gradient?*

Motivation: We thus consider introducing Muon into SAM with two natural thoughts: 1) SAM uses the perturbed gradient to form the outer update, which may directly affect the optimization behavior; 2) Muon could modify this update through matrix orthogonalization. This makes MuonSAM theoretically feasible. Specifically, MuonSAM applies matrix orthogonalization at two stages: **1)** For the sharpness probe, it combines the clean gradient with the stored momentum state to form a temporary input, then uses the approximate polar direction of that input to perturb each Muon

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.

matrix block. This probe does not change the stored state, and non Muon blocks retain Euclidean perturbations; **2)** At the perturbed parameters, MuonSAM computes a new gradient, updates the momentum state, and forms a separate polar direction for the outer Muon update. Non Muon blocks receive AdamW updates Loshchilov & Hutter (2019).

Theoretical results: Our analysis establishes three properties: **1)** For a model with one matrix parameter and without momentum, exact polar exploration yields a first order loss term proportional to the gradient's nuclear norm divided by the sum of its Frobenius norm and a stabilizer (Theorem 1). A mixed parameter analysis extends this result and includes an explicit smoothness remainder (Corollary 1); **2)** A spectral characterization of finite Newton–Schulz iterations and a bound on the operator norm error describe approximation effects (Theorem 2); a separate condition identifies when momentum yields positive first order alignment with the clean gradient (Lemma 1); **3)** Global normalization produces *dynamic, coupled matrix block radii* (Eq. 14). An exact relation between the inner and outer momentum inputs expresses how a change in the gradient caused by the probe affects the outer direction (Proposition 1). An appendix trajectory bound accounts for outer approximation error (Theorem 3).

Experimental findings. The experiments assess predictive performance and the predicted geometry: **1)** MuonSAM *improves final test accuracy over matched Muon by* 0.20%–$2.75\% \uparrow$ on the evaluated image and text tasks and achieves *higher mean accuracy on every evaluated corruption suite*; **2)** Short run diagnostics show that realized matrix radii follow the global normalization formula and that the inner and outer directions differ. A timing audit measures the per step overhead relative to SAM.

Structured prospective: In MuonSAM, *the inner polar direction selects where the loss is evaluated, while the outer polar direction, formed from momentum updated with the perturbed gradient, determines the matrix update*.

The contributions are:

- **Theoretical insight.** The analysis shows how matrix geometry affects both probe sensitivity and the conversion of the perturbed gradient into an update. It characterizes exact polar sensitivity (Theorem 1), finite step approximation effects (Theorem 2), globally normalized matrix block radii (Eq. 14), and the relation between inner and outer momentum inputs (Proposition 1).
- **MuonSAM optimization.** The resulting mixed parameter optimizer combines globally normalized polar probes and perturbed gradient Muon updates for matrix blocks with Euclidean probes and AdamW updates for non Muon blocks.
- **Algorithm prototype.** The probe uses the orthogonalized clean gradient, while the update uses the orthogonalized gradient at the perturbed parameters.
- **Experimental Evidence.** For the evaluated image and text tasks, MuonSAM improves final test accuracy over matched Muon by 0.20%–$2.75\% \uparrow$ and achieves higher mean accuracy on every evaluated corruption suite. Additional diagnostics measure matrix radii, inner and outer direction changes, and per step overhead.

2 PRELIMINARIES

Training setup. We consider supervised training on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of examples, x_i is an input, and y_i is its target. The full parameter vector is $\theta \in \mathbb{R}^p$, where p is the total number of trainable scalar parameters, and $L(\theta)$ is the differentiable training loss at a given step. For a matrix block $W \in \mathbb{R}^{m \times n}$, let $G = \nabla_W L(\theta)$ denote its gradient with the other blocks held fixed. For higher order Muon tensors, W denotes the two dimensional matricization used by the implementation. This reshape preserves Frobenius norms and inner products; rank and singular values below refer to that matricization.

SAM perturbation. Standard SAM evaluates the loss within a Euclidean radius $\rho > 0$ around θ before computing the update gradient. Here $\| \cdot \|_2$ is the vector Euclidean norm, and ∇ denotes differentiation with respect to the displayed parameter. Under a first order approximation, the maximizing Euclidean perturbation is aligned with the raw gradient.

Matrix orthogonalization notation. Muon instead transforms matrix gradients with an orthogonalization map $\mathcal{O}(\cdot)$, approximated in practice by Newton–Schulz iteration. For a rank-r matrix with compact singular value decomposition (SVD) $G = U_r \Sigma_r V_r^\top$, $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{n\times r}$ have orthonormal columns, and $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains the positive singular values. The idealized polar direction is $\mathcal{O}(G) = U_r V_r^\top$, with $\mathcal{O}(0) = 0$; we also write $P = \mathcal{O}$ when discussing polar map derivatives and trajectory bounds. For equally shaped matrices, define the Frobenius inner product by $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$; write $\|A\|_F$ for the Frobenius norm, $\|A\|_{\mathrm{op}}$ for the largest singular value, and $\|A\|_* = \sum_i \sigma_i(A)$ for the nuclear norm. Here Tr is the matrix trace. MuonSAM uses this matrix geometry for neighborhood exploration and the perturbed gradient update.

Parameter blocks and iteration variables. Partition the trainable parameters into Muon matrix blocks $\mathcal{M}$ and non Muon blocks $\mathcal{A}$ (biases, normalization parameters, scalars, and vectors). At step t, let θ_t be the full parameter vector, $g_t = \nabla_\theta L(\theta_t)$ its clean gradient, $G_{W,t}$ the component for $W \in \mathcal{M}$, and $g_{v,t}$ the component for $v \in \mathcal{A}$. The saved momentum buffer for a matrix block is $B_{W,t}$, and $M_{W,t}$ denotes tentative clean gradient momentum formed without changing that buffer. Write $\mathcal{O}_{\mathrm{NS}}$ for the finite step Newton–Schulz map. When the iteration count matters, $Q_K(G)$ denotes the exact arithmetic output after K iterations; it is the K step version of this map, while the implementation uses its configured iteration count and finite precision arithmetic. Given a numerical stabilizer $\varepsilon > 0$, the common normalization factor is

$$Z_t = \|g_t\|_2 + \varepsilon = \left(\sum_{W\in\mathcal{M}} \|G_{W,t}\|_F^2 + \sum_{v\in\mathcal{A}} \|g_{v,t}\|_2^2 \right)^{1/2} + \varepsilon. \tag{1}$$

The full perturbation gives perturbed parameters $\widetilde{\theta}_t = \theta_t + \Delta_t$ and the corresponding gradient $g_t^{\mathrm{pert}} = \nabla_\theta L(\widetilde{\theta}_t)$. Here, $\|\Delta_t\|_2$ is the Euclidean norm after concatenating all parameter blocks.

3 Matrix Orthogonalization in Sharpness Aware Optimization

This section establishes three main results: the first order sensitivity of an exact polar probe (Theorem 1); the two-stage perturbation and its inner–outer momentum relation (Definition 1 and Proposition 1); and the effects of finite step orthogonalization and global normalization on sensitivity and matrix radii (Theorem 2, Corollary 1, and Eq. 14). Supporting error and alignment bounds appear in Appendix B.

3.1 Why Orthogonalize Sharpness Exploration?

Standard Euclidean SAM searches a Frobenius norm neighborhood for a matrix block. For a small perturbation, its first order loss increase is $\rho\|G\|_F$, and its probe follows $G/\|G\|_F$. Thus, its directional weight on singular mode i is proportional to σ_i. Muon's polar direction instead assigns unit directional weight to each active mode, and its inner product with G equals the nuclear norm $\|G\|_*$ by matrix Hölder inequality. To compare directions at the same realized Frobenius radius $\delta > 0$, let $r = \mathrm{rank}(G) > 0$ and define $\Delta_{\mathrm{pol}} = \delta\mathcal{O}(G)/\sqrt{r}$ and $\Delta_{\mathrm{Euc}} = \delta G/\|G\|_F$. Their first order loss changes satisfy

$$\langle G, \Delta_{\mathrm{pol}}\rangle = \delta\frac{\|G\|_*}{\sqrt{r}} \le \delta\|G\|_F = \langle G, \Delta_{\mathrm{Euc}}\rangle. \tag{2}$$

Since $\|G\|_* \le \sqrt{r}\|G\|_F$, at this matched radius the polar probe's first order loss change is no larger than that of the Euclidean direction. The comparison isolates the effect of reweighting active singular modes; it does not make the polar probe a maximizer over a Frobenius ball. MuonSAM instead uses a separate global scale, specified in Section 3.2.

The smoothness assumptions below are local analytical conditions: a Lipschitz gradient along the specific perturbation segment is enough to control the quadratic Taylor remainder, without requiring global smoothness of the training objective. Smoothness conditions of this kind support analyses of the perturb-and-evaluate step in the original SAM formulation and appear in convergence analyses of SAM variants Foret et al. (2021); Si & Yun (2023); Oikonomou & Loizou (2025). For Muon matrix blocks, the exact compact polar factor is the idealized orthogonalized update direction; practical

Muon implementations approximate this direction with Newton–Schulz iterations Jordan et al. (2024); Kim & Oh (2026).

Theorem 1 (Exact Polar Directional Sensitivity). *Consider a single matrix parameter W with gradient $G = \nabla_W L(W)$, no momentum, no non Muon parameter blocks, and an exact compact polar factor $\mathcal{O}(G)$. Assume $\nabla_W L$ is L_{sm}-Lipschitz along the perturbation segment. The MuonSAM perturbation is $\Delta = \rho\mathcal{O}(G)/(\|G\|_F + \varepsilon)$. Let $G = U_r\Sigma_r V_r^\top$ be its compact SVD, with $r = \mathrm{rank}(G)$, so that $\mathcal{O}(G) = U_r V_r^\top$. Here $\rho > 0$ is the perturbation scale, $\varepsilon > 0$ is the normalization stabilizer, L_{sm} is the gradient Lipschitz constant on the perturbation segment, and R_{pol} is the Taylor remainder.*

$$L(W + \Delta) = L(W) + \rho\frac{\|G\|_*}{\|G\|_F + \varepsilon} + R_{\mathrm{pol}}, \qquad |R_{\mathrm{pol}}| \le \frac{L_{\mathrm{sm}}\rho^2 r}{2(\|G\|_F + \varepsilon)^2}. \tag{3}$$

This theorem isolates the polar probe's first order loss contribution and bounds its curvature through the matrix rank.

The setup combines SAM's perturb-and-evaluate step with the compact polar direction underlying Muon matrix updates Foret et al. (2021); Jordan et al. (2024); Kim & Oh (2026); the theorem gives the resulting directional sensitivity and remainder for this probe.

The identity separates the probe's directional effect from curvature. The polar factor preserves the active singular vectors and replaces each positive singular value by one, so its inner product with G is the sum of those singular values, namely $\|G\|_*$. Its squared Frobenius norm is r, which gives the remainder bound that depends on rank under the stated smoothness condition. Thus, for a fixed gradient scale, the first order increase reflects how many singular modes participate, while ε stabilizes the normalization when $\|G\|_F$ is small. Appendix B gives the Taylor expansion and remainder calculation.

The idealized first order matrix quantity is the normalized polar directional sensitivity

$$\mathcal{S}_{\mathrm{pol}}(G) := \frac{\|G\|_*}{\|G\|_F + \varepsilon}. \tag{4}$$

For $G \neq 0$ and ignoring ε, it satisfies $1 \le \|G\|_*/\|G\|_F \le \sqrt{r}$, where $r = \mathrm{rank}(G) \le \min(m, n)$. This ratio is close to 1 when the gradient spectrum is concentrated in one singular direction and approaches $\sqrt{r}$ when the singular values participate more evenly. Away from the ε scale, it is invariant to multiplying G by a positive scalar and therefore describes spectral participation rather than gradient magnitude.

3.2 How MuonSAM Orthogonalizes

Globally normalized polar probe. MuonSAM uses a global gradient norm to scale all parameter blocks. Muon matrix blocks probe along approximate polar directions, while non Muon blocks retain Euclidean probes. The definition below uses a common ρ and no adaptive scaling; it does not define a fixed radius or an inner maximization over orthogonal matrices.

Definition 1 (Globally Normalized Muon Guided Perturbation). *Given the notation of Section 2 and a scale coefficient $\rho > 0$, form $Q^{\mathrm{in}}_{W,t} = \mathcal{O}_{\mathrm{NS}}(M_{W,t})$ for each $W \in \mathcal{M}$. MuonSAM perturbs the two block types by*

$$\Delta_{W,t} = \frac{\rho}{Z_t} Q^{\mathrm{in}}_{W,t}, \qquad \Delta_{v,t} = \frac{\rho}{Z_t} g_{v,t}. \tag{5}$$

The full perturbation Δ_t concatenates these block perturbations.

Remark 1 (Perturbation Scale Versus Realized Radius). ***ρ is a scale coefficient, not the realized perturbation radius.*** *The full radius varies with Z_t, block dimensions, momentum alignment, the finite step approximation, and non Muon gradients.*

Perturbed gradient update. The inner polar direction sets the probe; a separately computed outer polar direction forms the update. For one matrix block, $Q^{\mathrm{in}}_t = \mathcal{O}(M_t)$ gives $\Delta_t = (\rho/Z_t)Q^{\mathrm{in}}_t$ and the perturbed gradient $G^{\mathrm{pert}}_t = \nabla L(W_t + \Delta_t)$. The outer update uses $Q^{\mathrm{out}}_t = \mathcal{O}(M^{\mathrm{pert}}_t)$, so generally

$Q_t^{\text{out}} \neq Q_t^{\text{in}}$. The shared factor ρ/Z_t sets probe scale; the polar direction sets its matrix geometry. Algorithm 1 gives the full blockwise update.

Implementation via Newton–Schulz Iteration. Five Newton–Schulz iterations approximate each polar direction without an SVD. In Algorithm 1, the probe reads the momentum buffer without changing it; the outer step updates it with $G_{W,t}^{\text{pert}}$. Appendix D specifies the iteration coefficients and audits the approximation.

Algorithm 1: MuonSAM

Input: Parameters θ, loss L, perturbation scale ρ, normalization stabilizer ε, learning rate η
Output: Optimized θ

for *each batch* **do**
  $g_t \leftarrow \nabla_\theta L(\theta_t)$; `// First gradient pass`
  $Z_t \leftarrow \|g_t\|_2 + \varepsilon$; `// Global norm over all parameter groups`
  for *each Muon matrix block* $W \in \mathcal{M}$ **do**
    $M_{W,t} \leftarrow \text{TentativeMomentum}(G_{W,t}, B_{W,t})$; `// Read buffer only`
    $Q_{W,t}^{\text{in}} \leftarrow \mathcal{O}_{\text{NS}}(M_{W,t})$;
    $\widetilde{W}_t \leftarrow W_t + \frac{\rho}{Z_t} Q_{W,t}^{\text{in}}$;
  end
  for *each non Muon block* $v \in \mathcal{A}$ **do**
    $\widetilde{v}_t \leftarrow v_t + \frac{\rho}{Z_t} g_{v,t}$;
  end
  $g_t^{\text{pert}} \leftarrow \nabla_\theta L(\widetilde{\theta}_t)$; `// Second gradient pass`
  Restore θ_t from $\widetilde{\theta}_t$;
  for *each Muon matrix block* $W \in \mathcal{M}$ **do**
    $M_{W,t}^{\text{pert}} \leftarrow \text{UpdateMomentum}(G_{W,t}^{\text{pert}}, B_{W,t})$; `// Write buffer`
    $Q_{W,t}^{\text{out}} \leftarrow \mathcal{O}_{\text{NS}}(M_{W,t}^{\text{pert}})$;
    $W_{t+1} \leftarrow W_t - \eta Q_{W,t}^{\text{out}}$;
  end
  Update non Muon blocks with AdamW using g_t^{pert};
end

Inner–outer momentum relation. For one Muon matrix block W, suppress the block subscript and write $G_t := G_{W,t}$, $G_t^{\text{pert}} := G_{W,t}^{\text{pert}}$, and $B_t := B_{W,t}$. Thus $M_t^{\text{in}} = M_{W,t}$ and $M_t^{\text{out}} = M_{W,t}^{\text{pert}}$. Let $\mu \in [0, 1)$ be the momentum coefficient. Under Nesterov interpolation with decoupled (or zero) weight decay, the tentative clean and committed perturbed gradient momentum inputs are

$$M_t^{\text{in}} = \mu^2 B_t + (1-\mu^2) G_t, \qquad M_t^{\text{out}} = \mu^2 B_t + (1-\mu^2) G_t^{\text{pert}}. \tag{6}$$

Let $Q_t^{\text{in/out}} = \mathcal{O}_{\text{NS}}(M_t^{\text{in/out}})$ and $C_t^{\text{geo}} = Q_t^{\text{out}} - Q_t^{\text{in}}$. Here M_t^{in} and M_t^{out} are the tentative clean and committed perturbed gradient momentum inputs, $Q_t^{\text{in/out}}$ are their finite step polar directions, and C_t^{geo} is the outer minus inner direction.

Proposition 1 (Inner–Outer Momentum Relation)**.** *For the momentum inputs above and the full perturbation Δ_t from Definition 1,*

$$M_t^{\text{out}} - M_t^{\text{in}} = (1-\mu^2)(G_t^{\text{pert}} - G_t), \qquad C_t^{\text{geo}} = \mathcal{O}_{\text{NS}}(M_t^{\text{out}}) - \mathcal{O}_{\text{NS}}(M_t^{\text{in}}). \tag{7}$$

If the full gradient $\nabla_\theta L$ is L_{sm}-Lipschitz along the perturbation segment, then

$$\|M_t^{\text{out}} - M_t^{\text{in}}\|_F \leq (1-\mu^2) L_{\text{sm}} \|\Delta_t\|_2. \tag{8}$$

If the direction map $\mathcal{O}_{\text{NS}}$ has Lipschitz constant K_{NS} on the segment joining the two momentum matrices, then $\|C_t^{\text{geo}}\|_F \leq K_{\text{NS}}(1-\mu^2) L_{\text{sm}} \|\Delta_t\|_2$.

This proposition connects the probe induced gradient change to the outer direction change, with the latter bound requiring local Lipschitz continuity.

The shared buffer cancels in the difference of momentum inputs, leaving only the probe induced gradient change; smoothness therefore gives the stated input bound. Transferring it to the direction difference requires local Lipschitz continuity of $\mathcal{O}_{\mathrm{NS}}$. The exact arithmetic map has this property where its normalization denominator is positive. In finite precision, add the difference between the two direction errors; without Nesterov interpolation, replace $1-\mu^2$ by $1-\mu$.

For the exact polar map, full rank and a locally Lipschitz Hessian give the local response in Eq. 16 (Appendix B); the full-rank requirement is consistent with classical perturbation analysis of the polar factor Higham (1986). **Near rank deficiency, the direction response can grow, so the relation does not guarantee attenuation.** Weight decay and shape dependent learning rate also contribute to the practical parameter update.

3.3 What Finite Step Orthogonalization and Global Normalization Change

Mixed parameter sensitivity. The global normalization and mixed parameter groups determine the realized first order probe sensitivity. Assume that $\nabla_\theta L$ is L_{sm}-Lipschitz along the segment from θ_t to $\theta_t+\Delta_t$, and let $\|\Delta_t\|_2$ denote the Euclidean norm after concatenating all parameter blocks. Taylor expansion gives

$$L(\theta_t+\Delta_t)=L(\theta_t)+\frac{\rho}{Z_t}\left[\sum_{W\in\mathcal{M}}\langle G_{W,t},Q^{\mathrm{in}}_{W,t}\rangle+\sum_{v\in\mathcal{A}}\|g_{v,t}\|_2^2\right]+R_t \tag{9}$$

Here $R_t:=L(\theta_t+\Delta_t)-L(\theta_t)-\langle g_t,\Delta_t\rangle$ is the Taylor remainder for the full parameter perturbation, with $|R_t|\le L_{\mathrm{sm}}\|\Delta_t\|_2^2/2$. The linear term sums matrix-block alignments and non-Muon squared gradient norms. Its first-order interpretation requires the remainder to be controlled relative to that term; Z_t alone does not ensure it.

Finite step polar response. For the finite step map, let $G=U_r\,\mathrm{diag}(\sigma_1,\dots,\sigma_r)V_r^\top\neq 0$ be the compact SVD, where $r=\mathrm{rank}(G)$ and $\sigma_i>0$ are its nonzero singular values. Write $P(G)=U_rV_r^\top$ for the exact polar factor. Set $d_G=\|GG^\top\|_F^{1/2}+\varepsilon_{\mathrm{NS}}$, where $\varepsilon_{\mathrm{NS}}>0$ stabilizes the matrix normalization, and define the scalar recurrence

$$q_0(s)=s,\qquad q_{k+1}(s)=aq_k(s)+bq_k(s)^3+cq_k(s)^5,\quad (a,b,c)=(3.4445,-4.7750,2.0315). \tag{10}$$

Here q_k is the scalar response after k iterations, (a,b,c) are the fixed polynomial coefficients, and $\varepsilon_{\mathrm{NS}}$ is separate from the perturbation stabilizer ε in Z_t.

Theorem 2 (Finite Newton–Schulz Spectral Response)**.** *For the matrix G and recurrence above, the exact arithmetic output $Q_K(G)$ after K Newton–Schulz steps satisfies*

$$Q_K(G)=U_r\,\mathrm{diag}\left(q_K\left(\frac{\sigma_i}{d_G}\right)\right)_{i=1}^{r}V_r^\top. \tag{11}$$

The finite iteration reduces to scalar responses, making its alignment and worst-mode error explicit.

Each Newton–Schulz step is a matrix polynomial, so it preserves singular vectors and applies q_K to each normalized singular value; this characterizes the finite-step orthogonalization used by Muon Jordan et al. (2024); Kim & Oh (2026). Consequently, $\langle G,Q_K(G)\rangle/\|G\|_*$ is the singular value weighted average of $q_K(\sigma_i/d_G)$; the operator-norm error from $P(G)$ is $\max_i|q_K(\sigma_i/d_G)-1|$. The normalization couples the modes. These identities hold in exact arithmetic; approximation and rounding are handled by Proposition 2, Lemma 1, and the implementation audits in Appendix D. The proof is in Appendix B.

Momentum alignment and mixed-parameter sensitivity. For each Muon block, let $\mathcal{O}(M_{W,t})$ denote the exact polar direction of the tentative clean gradient momentum $M_{W,t}$, and measure its

alignment with the clean gradient and the finite step approximation error by

$$\alpha_{W,t} := \frac{\langle G_{W,t}, \mathcal{O}(M_{W,t})\rangle}{\|G_{W,t}\|_*} \quad (G_{W,t} \neq 0), \qquad \xi_{W,t} := \|\mathcal{O}_{\mathrm{NS}}(M_{W,t}) - \mathcal{O}(M_{W,t})\|_{\mathrm{op}}. \tag{12}$$

Set $\alpha_{W,t} = 0$ when $G_{W,t} = 0$.

Nuclear/operator norm duality bounds the implemented alignment by $[\alpha_{W,t} - \xi_{W,t}, \alpha_{W,t} + \xi_{W,t}]$. Thus, $\alpha_{W,t} > \xi_{W,t}$ guarantees positive first-order alignment; otherwise this bound gives no sign guarantee (Lemma 1).

Corollary 1 (Mixed Parameter Probe Sensitivity)**.** *Under the smoothness condition of the mixed parameter expansion above, define*

$$B_t^{\pm} := \frac{\rho}{Z_t}\left[\sum_{W\in\mathcal{M}} (\alpha_{W,t} \pm \xi_{W,t})\|G_{W,t}\|_* + \sum_{v\in\mathcal{A}} \|g_{v,t}\|_2^2\right],$$
$$B_t^- - \frac{L_{\mathrm{sm}}}{2}\|\Delta_t\|_2^2 \le L(\theta_t + \Delta_t) - L(\theta_t) \le B_t^+ + \frac{L_{\mathrm{sm}}}{2}\|\Delta_t\|_2^2. \tag{13}$$

Here B_t^- and B_t^+ are the lower and upper bounds on the first order probe contribution, obtained by summing the matrix block alignment bounds and the non Muon block gradient contributions.

This corollary combines blockwise alignment with curvature to bound the loss change under the full mixed parameter probe.

The interval sums the matrix alignment bounds and non-Muon gradient terms; the smoothness remainder accounts for curvature across blocks. It certifies loss increase only when B_t^- exceeds the curvature allowance; otherwise it gives no sign guarantee. See Appendix B for the derivation.

Dynamic effective matrix radius. For an $m_W \times n_W$ matrix block, let $r_{W,t} = \mathrm{rank}(G_{W,t})$. In the idealized no momentum case with an exact compact polar factor, its Frobenius norm is $\sqrt{r_{W,t}}$. The dynamic effective matrix radius is therefore:

$$\rho_{W,t}^{\mathrm{eff}} := \|\Delta_{W,t}\|_F = \frac{\rho\sqrt{r_{W,t}}}{Z_t} \tag{14}$$

Thus, the realized radius varies with rank and the global scale even when ρ is fixed. For the implemented direction, $\|\Delta_{W,t}\|_F = (\rho/Z_t)\|Q_{W,t}^{\mathrm{in}}\|_F$; because Z_t includes all active blocks, a change in one block rescales the perturbations of the others. Blockwise normalization would instead give $\rho\|Q_{W,t}^{\mathrm{in}}\|_F/(\|G_{W,t}\|_F + \varepsilon)$, while normalizing $Q_{W,t}^{\mathrm{in}}$ itself would fix each nonzero block radius at ρ; these choices define distinct neighborhoods.

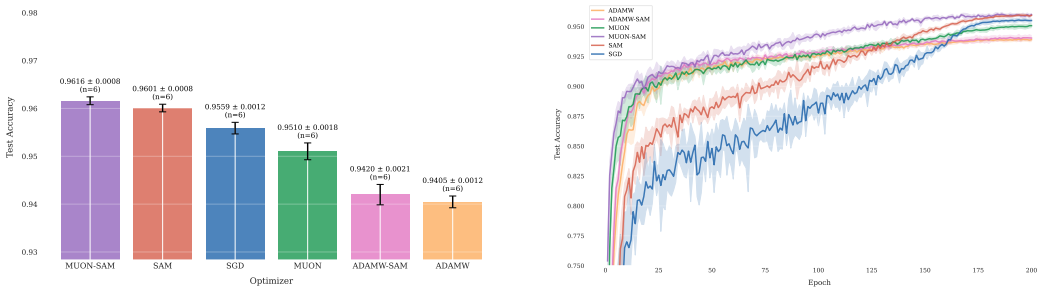
4 Experiments

Table 1: Mean final checkpoint test accuracy (%) across vision and language benchmarks; standard deviations are shown after ± when multiple runs are available. Best mean accuracy in each row is bold.

Dataset	Model	SGD	SAM	AdamW	AdamW-SAM	Muon	MuonSAM
CIFAR-10	ResNet-34	95.58 ± 0.13	96.31 ± 0.05	94.21 ± 0.23	94.41 ± 0.10	95.66 ± 0.08	95.86 ± 0.23
CIFAR-100	ResNet-34	78.91 ± 0.54	79.52 ± 0.12	74.97 ± 0.46	74.53 ± 0.56	78.81 ± 0.10	81.56 ± 0.24
ImageNet-100	ResNet-34	83.88 ± 0.20	84.22 ± 0.14	83.06 ± 0.27	83.50 ± 0.24	85.42 ± 0.21	87.92 ± 0.23
ImageNet-100	ResNet-50	84.66 ± 0.24	86.08 ± 0.21	84.04 ± 0.29	84.98 ± 0.22	85.84 ± 0.18	88.44 ± 0.21
CIFAR-10	ViT	75.15 ± 0.23	73.00 ± 0.02	88.30 ± 0.11	88.07 ± 0.44	87.39 ± 0.03	88.69 ± 0.30
SST-2	TinyBERT	86.81 ± 1.74	87.31 ± 4.30	82.42 ± 0.70	85.74 ± 1.08	88.00 ± 0.46	89.53 ± 0.37
SST-5	GPT-2	35.97 ± 8.38	47.76 ± 6.56	51.92 ± 0.99	52.83 ± 0.03	51.13 ± 1.03	53.21 ± 0.98

We first ask whether adding the orthogonalized sharpness probe and perturbed gradient update improves on the matched Muon optimizer under the same training budgets. Table 1 reports final

checkpoint test accuracy across seven vision and language settings. MuonSAM exceeds Muon in all seven, by 0.20%–2.75% ↑, and attains the highest mean among the listed optimizers in six settings. The gains span convolutional and transformer vision models and both language tasks. CIFAR-10/ResNet-34 is the exception to the comparison with the other baselines: SAM has the highest mean there, while the two Muon variants are close. On CIFAR-10/ResNet-18, a separate evaluation with six seeds reports $96.16\% \pm 0.08\%$ for MuonSAM and $95.10\% \pm 0.18\%$ for Muon (Figure 1). Appendix D specifies the training protocols, hyperparameters, timing setup, and implementation diagnostics.



(a) Final test accuracy, mean ± SD over six seeds.

(b) Test accuracy over 200 epochs.

Figure 1: CIFAR-10/ResNet-18 performance stability. MuonSAM reaches $96.16\% \pm 0.08\%$ final test accuracy across six seeds, compared with $95.10\% \pm 0.18\%$ for Muon. The right panel shows the mean and variability of test accuracy over training.

4.1 Corruption Robustness

We evaluate final checkpoints on fixed image corruptions and textual perturbations, comparing MuonSAM with matched Muon.

Finding 1: Higher mean corruption accuracy. MuonSAM improves mean accuracy in all seven evaluated settings. The gains are 1.8 percentage points on CIFAR-10/ResNet-18 (68.7% vs. 66.9%), 1.3 on CIFAR-10/ResNet-34 (68.9% vs. 67.6%), 2.0 on CIFAR-100/ResNet-34 (51.1% vs. 49.1%), 1.7 on CIFAR-10/ViT (70.6% vs. 68.9%), 8.2 on ImageNet-100/ResNet-34 (78.4% vs. 70.2%), 1.2 on SST-2/TinyBERT (81.4% vs. 80.2%), and 2.7 on SST-5/GPT-2 (43.2% vs. 40.5%). The gains span both vision and language tasks. Six settings improve by 1.2%–2.7%, while ImageNet-100 has a larger 8.2-point gain. This consistency in the direction of the mean change supports a benefit over matched Muon on these fixed suites, while the different magnitudes motivate inspecting the per-corruption results.

Table 2: Accuracy by corruption (%) on CIFAR-10/ResNet-18. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	G-Blur	H-Blur	Bright	Contr	Satur	Solar	Postr	Sharp	Elast	Rotat	G-Nois	S-Nois	Mean
ADAMW	15.1	12.4	93.4	93.6	94.0	76.3	92.3	89.2	90.5	87.1	**41.3**	10.3	66.3
ADAMW-SAM	19.3	15.7	93.5	93.7	94.1	76.0	92.3	89.2	90.6	86.6	39.8	11.7	66.9
MUON	20.2	15.8	94.8	94.8	95.2	77.4	92.9	91.0	92.1	87.6	27.2	13.3	66.9
MUONSAM	**27.0**	**21.6**	**95.6**	**95.8**	95.8	**83.2**	93.1	91.3	**94.1**	**90.8**	24.4	11.5	68.7
SAM	23.6	17.8	**95.6**	95.7	**95.9**	80.1	**94.1**	**92.6**	93.7	89.5	37.9	12.7	**69.1**
SGD	21.7	18.6	94.9	95.5	95.5	78.9	93.8	91.9	92.7	89.6	40.4	**14.9**	69.0

Finding 2: Gains vary across corruption types. The per-corruption results show where these mean changes arise. On CIFAR-10/ResNet-18, MuonSAM improves over Muon on blur, color, lighting, and geometric corruptions, but trails on Gaussian noise (24.4% vs. 27.2%) and shot noise (11.5% vs. 13.3%). On ImageNet-100/ResNet-34, it outperforms Muon on every listed corruption, with the largest gaps on Gaussian noise (70.9% vs. 41.2%) and solarization (86.4% vs. 62.3%). Thus, the CIFAR-10/ResNet-18 mean gain coexists with specific regressions, while the ImageNet-100 gain spans every listed condition. Tables 2 and 3 give the full per corruption accuracies for these settings.

Finding 3: The gain depends on the evaluation setting. The mean improvement ranges from 1.2% on SST-2 to 8.2% on ImageNet-100: the two language settings improve by 1.2% and 2.7%, while vision settings range from 1.3% to 8.2%. Since dataset and model vary together, these results do not isolate which factor explains the difference. Per-corruption tables for the other settings are in Appendix D.1.

Table 3: Accuracy by corruption (%) on ImageNet-100/ResNet-34. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	G-Blur	H-Blur	Bright	Contr	Satur	Solar	Postr	Sharp	Elast	Rotat	G-Nois	S-Nois	Mean
ADAMW	78.2	69.2	81.8	81.5	82.7	54.2	80.8	82.1	81.4	77.6	52.8	7.5	69.1
ADAMW-SAM	78.7	70.7	82.5	82.5	83.0	55.3	82.2	83.1	82.0	78.5	57.4	**14.6**	70.9
MUON	80.3	71.0	84.5	84.6	85.0	62.3	82.8	84.4	83.8	78.1	41.2	4.9	70.2
MUONSAM	**84.5**	**76.9**	**88.1**	**88.1**	**87.8**	**86.4**	**87.6**	**88.2**	**87.4**	**86.6**	**70.9**	8.2	**78.4**
SAM	79.9	71.0	83.0	83.3	83.7	57.8	83.1	83.7	82.8	79.5	60.0	13.2	71.7
SGD	79.6	71.2	82.4	82.7	83.3	57.4	82.1	83.1	82.6	78.6	59.3	13.4	71.3

Corruption categories. Vision perturbations cover blur (G-Blur, H-Blur), color and lighting shifts (Bright, Contr, Satur, Solar), digital artifacts (Postr, Sharp), geometric changes (Elast, Rotat), and sensor noise (G-Nois, S-Nois). Text perturbations cover character edits (C-Del, C-Ins, C-Swp, C-Sub), word edits (W-Del, W-Swp, W-Ins), truncation (Tr-50, Tr-80), padding (Pad), and formatting or spelling shifts (Upper, Lower, Spell). The remaining CIFAR-10/ResNet-34, CIFAR-100/ResNet-34, ViT, and language per condition tables appear in Appendix D.1.

4.2 Computational Overhead

To quantify the implementation cost, we measured per step time and peak memory on CIFAR-10/ResNet-18 with an NVIDIA RTX 4070Ti. MuonSAM takes 60.71 ms per step, 17.1% more than SAM under the reported protocol, with comparable peak memory (766.04 MB versus 766.25 MB). Compared with single-pass Muon (29.40 ms), it takes about $2.06\times$ as long per step, consistent with its two-pass protocol. These measurements use one model and hardware setup. Table 4 reports all measurements; Appendix D documents synchronization, warmup, repeats, and parameter groups.

Table 4: Optimizer overhead on CIFAR-10/ResNet-18. Measurements use FP32, batch size 128, 20 warmup steps, 100 measured steps, and five repeated runs. Step time and peak memory are reported as mean $\pm$ standard deviation across repeats.

Optimizer	Protocol	Step time (ms)	Peak memory (MB)	Relative note
SGD	Single pass	25.79 ± 0.02	721.20 ± 0.95	Fastest
SAM	Two pass	51.85 ± 0.04	766.25 ± 0.00	Reference
AdamW	Single pass	26.68 ± 0.00	762.75 ± 0.00	Baseline
AdamW-SAM	Two pass	52.61 ± 0.01	808.13 ± 0.00	Highest memory
Muon	Single pass + NS	29.40 ± 0.00	718.54 ± 0.00	$+14.0\%$ vs. SGD
MuonSAM	Two pass + NS	60.71 ± 0.00	766.04 ± 0.00	$+17.1\%$ vs. SAM

5 Discussion

MuonSAM orthogonalizes both stages: the inner probe reweights matrix singular modes to select the evaluation point, while the outer update recomputes its direction from the perturbed gradient. Analysis and trajectory diagnostics show positive alignment of the finite step direction with the clean gradient and similar but measurably different inner and outer directions. At equal radius, the polar probe changes singular mode weighting, while the raw gradient produces a larger immediate loss increase under a Frobenius constraint.

MuonSAM improves final accuracy over matched Muon across the evaluated vision and language settings and raises mean accuracy on every reported corruption suite. The size of the robustness gain varies across tasks and perturbations: it is largest on ImageNet-100, while the per corruption results show both gains and regressions on CIFAR-10/ResNet-18. On CIFAR-10/ResNet-18, its two-pass update uses 17.1% more step time than SAM with comparable peak memory.

AI USE STATEMENT

AI tools assisted with language revision and manuscript consistency checks. The authors are responsible for the scientific claims, mathematical arguments, results, and final manuscript.

ETHICS STATEMENT

This work evaluates optimization methods on existing image and text benchmarks and introduces no new data collection or for users deployment.

REPRODUCIBILITY STATEMENT

Appendix D records the datasets, model configurations, training schedules, optimizer hyperparameters, timing protocol, and geometric diagnostics used in the reported experiments. The anonymized code repository is available at `https://anonymous.4open.science/r/iclr2027-muonsam-7007`.

REFERENCES

Atish Agarwala and Yann Dauphin. SAM operates far from home: eigenvalue regularization as a dynamical phenomenon. In *International Conference on Machine Learning*, volume 202, pp. 152–168, 2023.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, volume 162, pp. 247–257, 2022.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *International Conference on Knowledge Discovery and Data Mining*, 2019.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 639–668, 2022.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47032–47051, 2023.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 948–1024, 2022.

Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7360–7371, 2022.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Marlon Becker, Frederick Altrock, and Benjamin Risse. Momentum-sam: Sharpness aware minimization without computational overhead. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 55336–55366, 2025.

Lizhang Chen, Jonathan Li, and qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=Blz4hjxLwU`.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67741–67770, 2023.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, 2025.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Khashayar Gatmiry, Zhiyuan Li, Sashank J Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. In *International Conference on Machine Learning*, pp. 15102–15129. PMLR, 2024.

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

Ting Han, Linara Adilova, Henning Petzka, Jens Kleesiek, and Michael Kamp. Flatness is necessary, Neural Collapse is not: Rethinking generalization via grokking. In *Advances in Neural Information Processing Systems*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

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.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt Kusner. When do flat minima optimizers work? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

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

Hyeongchan Kim. PyTorch_Optimizer: Optimizer & LR scheduler & loss function collections in PyTorch, 2021.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 11148–11161, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

Noam Itzhak Levi and Yaron Oz. The underlying universal statistical structure of natural datasets. In *International Conference on Machine Learning*, volume 267, pp. 34020–34043, 2025.

Jiacheng Li, Jianchao Tan, Hongtao Xu, Jiaqi Zhang, Yifan Lu, Yerui Sun, Yuchen Xie, and Xunliang Cai. Mona: Muon optimizer with nesterov acceleration for scalable language model training, 2026. URL https://arxiv.org/abs/2605.26842.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5631–5640. IEEE, 2024.

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, 2025.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12350–12360. IEEE, 2022.

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon^2: Boosting muon via adaptive second-moment preconditioning, 2026. URL https://arxiv.org/abs/2604.09967.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Thomas Möllenhoff and Mohammad Emtiyaz Khan. Sam as an optimal relaxation of bayes. In *International Conference on Machine Learning*, 2023.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17749–17761, 2021.

Hossein R. Nowdeh, Jie Ji, Xiaolong Ma, and Fatemeh Afghah. Modality-aware SAM: Sharpness-aware-minimization driven gradient modulation for harmonized multimodal learning. In *Advances in Neural Information Processing Systems*, 2025.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *International Conference on Learning Representations*, volume 2025, pp. 30071–30097, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon, 2026. URL https://arxiv.org/abs/2505.23737.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of SAM. In *International Conference on Machine Learning*, volume 267, pp. 55702–55719, 2025.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Jacob Springer, Vaishnavh Nagarajan, and Aditi Raghunathan. Sharpness-aware minimization enhances feature quality via balanced learning. In *International Conference on Learning Representations*, volume 2024, pp. 22534–22558, 2024.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794, 2020.

Tuan Truong, Hoang-Phi Nguyen, Haocheng Luo, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. Sharpness-aware teleportation on riemannian manifolds, 2025. URL https://arxiv.org/abs/2309.17215.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Mingwei Xu, Xiaofeng Cao, and Ivor Tsang. Sharpness-aware minimization activates the interactive teaching's understanding and optimization. In *Advances in Neural Information Processing Systems*, 2024.

Hailiang Ye, Xinyi Fang, Ming Li, and Feilong Cao. Ed-sam: Sharpness-aware minimization with energy-adjusted perturbations and direction-corrected updates. *Neural Networks*, pp. 109185, 2026.

Yun Yue, Jiadi Jiang, Zhiling Ye, Ning Gao, Yongchao Liu, and Ke Zhang. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3185–3194, 2023.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, volume 2025, pp. 20949–20980, 2025.

Jinping Zou, Xiaoge Deng, and Tao Sun. Sharpness-aware minimization with adaptive regularization for training deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

APPENDIX CONTENTS

A Related Work **14**

B Proofs and Auxiliary Results **15**

B.1 Proof of Theorem 1 15

B.2 Auxiliary Approximation and Alignment Bounds 15

B.3 Proofs of Theorem 2 and Proposition 2 16

B.4 Proof of Lemma 1 16

B.5 Proof of Corollary 1 16

B.6 Proof of Proposition 1 17

B.7 Derivation of the Local Polar Response 17

C Trajectory Control under Approximate Polar Updates **17**

D Additional experiments and details **18**

D.1 Per Corruption Robustness Results 19

D.2 Inner–Outer Direction Diagnostics 21

D.3 Timing Protocol and Parameter Groups 21

A RELATED WORK

Sharpness objectives and probes. SAM constructs a perturbation before evaluating the gradient used by the outer optimizer Foret et al. (2021). Its behavior has been studied across vision, language, and multimodal settings Bahri et al. (2022); Kaddour et al. (2022); Singh et al. (2025); Nowdeh et al. (2025); Xu et al. (2024), including links to representation structure, feature diversity, and training phase Andriushchenko et al. (2023); Springer et al. (2024); Zhou et al. (2025). ASAM and Fisher SAM change the neighborhood geometry Kwon et al. (2021); Kim et al. (2022), while recent teleportation methods study sharpness over a quotient manifold Truong et al. (2025). WSAM and SAMAR instead revisit the sharpness objective and its weighting Yue et al. (2023); Zou et al. (2025). Related analyses address implicit bias, normalization, stochasticity, convergence, and stability Andriushchenko & Flammarion (2022); Gatmiry et al. (2024); Möllenhoff & Khan (2023); Dai et al. (2023); Si & Yun (2023); Oikonomou & Loizou (2025); Agarwala & Dauphin (2023); Bartlett et al. (2023); Long & Bartlett (2024); Cohen et al. (2021); Ahn et al. (2022); Arora et al. (2022); Mulayoff et al. (2021). This work motivates asking how a matrix polar direction changes the probe relative to a Euclidean gradient direction.

Use of the perturbed gradient. Several methods change how the sharpness signal enters the outer update or how it is computed. ED-SAM adjusts perturbations according to energy and corrects the update direction Ye et al. (2026); Friendly-SAM studies stochastic probe components and their effect on the optimization path Li et al. (2024); and LookSAM, ESAM, and Momentum-SAM reduce the cost or frequency of obtaining a SAM direction through reuse, stochastic selection, or a probe based on momentum Liu et al. (2022); Du et al. (2022); Becker et al. (2025). These methods address update formation or probe cost, while the present work examines how matrix orthogonalization transforms the gradient measured at the probe.

Matrix orthogonalization. Muon transforms momentum matrices through a finite step Newton–Schulz approximation to a polar direction Jordan et al. (2024). Theory relates this geometry to spectral norm optimization, implicit bias, convergence, and approximation error Chen et al. (2026); Fan et al. (2025); Kim & Oh (2026); Shen et al. (2026); scaling studies and extensions examine Muon in language model training Liu et al. (2025); Li et al. (2026); Liu et al. (2026). These works

characterize orthogonalization within the optimizer. MuonSAM applies the same transformation to select a SAM perturbation and recomputes the outer direction from momentum updated with the perturbed gradient. Table 5 compares representative methods along these two stages; its entries summarize the comparison axes rather than every change made by each method.

Table 5: Representative SAM designs viewed through exploration and perturbed gradient update. "Base optimizer" denotes the optimizer selected by the method, not a shared update rule across rows.

Method	Exploration direction or neighborhood	Perturbed gradient update
SAM Foret et al. (2021)	Euclidean gradient	Base optimizer
ASAM Kwon et al. (2021)	Euclidean neighborhood adapted to parameter scaling	Base optimizer
Fisher SAM Kim et al. (2022)	Neighborhood defined by Fisher geometry	Base optimizer
Friendly-SAM Li et al. (2024)	Stochastic component of probe	Base optimizer
ED-SAM Ye et al. (2026)	Perturbation adjusted by energy	Clean/perturbed gradient direction correction
Momentum-SAM Becker et al. (2025)	Momentum probe	Base optimizer
MuonSAM	Approximate polar direction of clean momentum	Independently computed approximate polar direction of perturbed gradient momentum

B Proofs and Auxiliary Results

Proof roadmap. The proofs establish the main text claims in dependency order:

- Theorem 1 identifies the first order sensitivity of an exact polar probe.
- Theorem 2, Proposition 2, Lemma 1, and Corollary 1 extend that calculation to finite Newton–Schulz directions and mixed parameter groups.
- Proposition 1 relates the inner and outer momentum inputs; the final derivation gives the local response of the exact polar surrogate.

B.1 Proof of Theorem 1

Proof. The compact polar factor satisfies $\langle G, \mathcal{O}(G)\rangle = \|G\|_*$ and $\|\mathcal{O}(G)\|_F^2 = r$. Taylor expansion along the perturbation segment gives

$$L(W+\Delta) = L(W) + \langle G, \Delta\rangle + R_{\mathrm{pol}}, \qquad |R_{\mathrm{pol}}| \le \frac{L_{\mathrm{sm}}}{2}\|\Delta\|_F^2.$$

Substituting $\Delta = \rho\mathcal{O}(G)/(\|G\|_F + \varepsilon)$ yields the claimed expression and remainder bound. □

B.2 Auxiliary Approximation and Alignment Bounds

Proposition 2 (Approximate Polar Directional Sensitivity)**.** *Let $G = \nabla_W L(W) \neq 0$ have compact SVD $G = U_r\Sigma_r V_r^\top$, and let $P(G) = U_r V_r^\top$ be its exact polar factor. For any computed direction $\widehat{Q}(G)$, let $\xi = \|\widehat{Q}(G) - P(G)\|_{\mathrm{op}}$ be its operator norm error. For $\Delta = (\rho/Z)\widehat{Q}(G)$, where $\rho > 0$ is the perturbation scale, $Z > 0$ is the normalization factor, and ∇L is L_{sm}-Lipschitz along the perturbation segment, define $R_{\mathrm{NS}} := (\rho/Z)\langle G, \widehat{Q}(G) - P(G)\rangle$ and $R_L := L(W+\Delta) - L(W) - \langle G, \Delta\rangle$. Then*

$$\begin{aligned} L(W+\Delta) &= L(W) + \frac{\rho}{Z}\|G\|_* + R_{\mathrm{NS}} + R_L, \\ |R_{\mathrm{NS}}| &\le \frac{\rho}{Z}\xi\|G\|_*, \qquad |R_L| \le \frac{L_{\mathrm{sm}}\rho^2}{2Z^2}\|\widehat{Q}(G)\|_F^2. \end{aligned} \tag{15}$$

This bound separates the first order effect of direction error from curvature along the perturbation segment.

Lemma 1 (Momentum Polar Alignment Bound). *For a Muon block with $G_{W,t} \neq 0$, define $\alpha_{W,t} = \langle G_{W,t}, \mathcal{O}(M_{W,t})\rangle / \|G_{W,t}\|_*$ and $\xi_{W,t} = \|\mathcal{O}_{\mathrm{NS}}(M_{W,t}) - \mathcal{O}(M_{W,t})\|_{\mathrm{op}}$. Then*

$$\alpha_{W,t} - \xi_{W,t} \leq \frac{\langle G_{W,t}, \mathcal{O}_{\mathrm{NS}}(M_{W,t})\rangle}{\|G_{W,t}\|_*} \leq \alpha_{W,t} + \xi_{W,t}.$$

In particular, $\alpha_{W,t} > \xi_{W,t}$ suffices for the approximate polar matrix probe to have positive first order alignment with the clean gradient.

This lemma gives a sufficient condition for positive alignment of the approximate momentum direction with the clean gradient.

B.3 Proofs of Theorem 2 and Proposition 2

Proof of Theorem 2. For a nonzero input matrix X with compact SVD $X = U_r \operatorname{diag}(s_1, \ldots, s_r) V_r^\top$, where r is its rank, $s_i > 0$ are its singular values, and U_r, V_r contain the corresponding left and right singular vectors, one Newton–Schulz step is

$$aX + b(XX^\top)X + c(XX^\top)^2 X = U_r \operatorname{diag}(as_i + bs_i^3 + cs_i^5) V_r^\top.$$

Starting from $s_i = \sigma_i / d_G$ and applying this identity inductively gives the scalar recurrence and the stated spectral representation. The alignment expression follows by taking the Frobenius inner product with G; the operator error is the largest absolute singular coefficient of $Q_K(G) - P(G)$. □

Proof of Proposition 2. For an arbitrary computed direction, nuclear/operator norm duality gives

$$|\langle G, \widehat{Q}(G) - P(G)\rangle| \leq \|G\|_* \|\widehat{Q}(G) - P(G)\|_{\mathrm{op}} = \xi \|G\|_*.$$

Combining this estimate with $\langle G, P(G)\rangle = \|G\|_*$ bounds the directional term. Taylor's theorem for $\Delta = (\rho/Z)\widehat{Q}(G)$ gives the remaining smoothness bound. □

B.4 Proof of Lemma 1

Proof. Write $\mathcal{O}_{\mathrm{NS}}(M_{W,t}) = \mathcal{O}(M_{W,t}) + E_{W,t}$, where $\|E_{W,t}\|_{\mathrm{op}} = \xi_{W,t}$. Nuclear/operator norm duality gives

$$\left| \frac{\langle G_{W,t}, \mathcal{O}_{\mathrm{NS}}(M_{W,t})\rangle}{\|G_{W,t}\|_*} - \alpha_{W,t} \right| = \left| \frac{\langle G_{W,t}, E_{W,t}\rangle}{\|G_{W,t}\|_*} \right| \leq \xi_{W,t}.$$

This is the two sided bound after rearrangement. If $\alpha_{W,t} > \xi_{W,t}$, its lower bound is positive, proving positive first order alignment. □

B.5 Proof of Corollary 1

Proof. By Lemma 1, each Muon block's inner product with $G_{W,t}$ lies between $(\alpha_{W,t} - \xi_{W,t})\|G_{W,t}\|_*$ and $(\alpha_{W,t} + \xi_{W,t})\|G_{W,t}\|_*$. For each non Muon block, the first order term is exactly $\langle g_{v,t}, \Delta_{v,t}\rangle = (\rho/Z_t)\|g_{v,t}\|_2^2$. Summing these terms gives a linear contribution between B_t^- and B_t^+. The mixed parameter Taylor remainder is bounded in absolute value by $L_{\mathrm{sm}}\|\Delta_t\|_2^2/2$, proving both inequalities. □

B.6 Proof of Proposition 1

Proof. The saved buffer before the step is B_t. The tentative buffer formed from the clean gradient is $\mu B_t + (1-\mu)G_t$, and Nesterov interpolation gives

$$(1-\mu)G_t + \mu[\mu B_t + (1-\mu)G_t] = \mu^2 B_t + (1-\mu^2)G_t.$$

The outer pass uses the same saved buffer with G_t^{pert}, yielding the stated difference identity. Lipschitz continuity of the full gradient gives the momentum difference bound. Applying the assumed Lipschitz constant of the direction map gives the bound on C_t^{geo}. □

B.7 Derivation of the Local Polar Response

Taylor expansion of the block gradient along the full perturbation gives

$$G_t^{\mathrm{pert}} - G_t = H_{W,t}[\Delta_t] + O(\|\Delta_t\|_2^2).$$

The momentum identity in Proposition 1 then yields

$$M_t^{\mathrm{out}} - M_t^{\mathrm{in}} = (1-\mu^2)H_{W,t}[\Delta_t] + O(\|\Delta_t\|_2^2).$$

Under the full rank condition, the compact polar map is Fréchet differentiable along the segment. Its first order expansion at M_t^{in} gives

$$P(M_t^{\mathrm{out}}) - P(M_t^{\mathrm{in}}) = (1-\mu^2)DP_{M_t^{\mathrm{in}}}[H_{W,t}[\Delta_t]] + O(\|\Delta_t\|_2^2). \tag{16}$$

Local boundedness of the second derivative controls the quadratic remainder.

C Trajectory Control under Approximate Polar Updates

Theorem 3 (Approximate Polar Trajectory Bound)**.** *For the iterates above and every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}(1-\xi_t)\|G_t^{\mathrm{pert}}\|_* \leq \frac{L(W_0)-L^*}{\eta T} + \frac{L_{\mathrm{sm}}\sqrt{n}}{T}\sum_{t=0}^{T-1}(1+\xi_t)\delta_t \tag{17}$$

$$+ \frac{L_{\mathrm{sm}}\eta n}{2T}\sum_{t=0}^{T-1}(1+\xi_t)^2. \tag{18}$$

Proof. The proof uses three bounds. First, nuclear/operator norm duality and the polar trace identity give $\langle G_t^{\mathrm{pert}}, \widehat{Q}_t\rangle \geq (1-\xi_t)\|G_t^{\mathrm{pert}}\|_*$. Second, $\|P(G_t^{\mathrm{pert}})\|_F \leq \sqrt{n}$ and $\|E_t\|_F \leq \sqrt{n}\|E_t\|_{\mathrm{op}}$ imply $\|\widehat{Q}_t\|_F \leq \sqrt{n}(1+\xi_t)$. Third, smoothness gives $\|G_t - G_t^{\mathrm{pert}}\|_F \leq L_{\mathrm{sm}}\delta_t$. Applying these bounds to the smoothness inequality at W_t yields

$$L(W_{t+1}) \leq L(W_t) - \eta\langle G_t, \widehat{Q}_t\rangle + \frac{L_{\mathrm{sm}}\eta^2}{2}\|\widehat{Q}_t\|_F^2 \tag{19}$$

$$\leq L(W_t) - \eta(1-\xi_t)\|G_t^{\mathrm{pert}}\|_* + \eta L_{\mathrm{sm}}\sqrt{n}(1+\xi_t)\delta_t + \frac{L_{\mathrm{sm}}\eta^2 n}{2}(1+\xi_t)^2. \tag{20}$$

Summing over t telescopes the losses, and $L(W_T) \geq L^*$ gives the bound. □

The following deterministic single matrix result quantifies how the realized probe radius and outer polar error enter a descent bound. Let $W_t \in \mathbb{R}^{m\times n}$ with $m \geq n$, where m and n are the row and column counts, respectively. Let $G_t = \nabla L(W_t)$ be the gradient at the current iterate, $G_t^{\mathrm{pert}} = \nabla L(W_t + \Delta_t)$ the gradient at the perturbed point, and $\delta_t = \|\Delta_t\|_F$ the realized perturbation radius. Assume that L is bounded below by L^* and has an L_{sm}-Lipschitz gradient. These are standard deterministic first-order convergence hypotheses: the lower bound controls the total objective decrease, while smoothness supplies the one-step quadratic upper bound used in the telescoping argument Ghadimi & Lan (2013); SAM convergence analyses also use smoothness assumptions,

with guarantees depending on the precise SAM variant and setting Si & Yun (2023); Oikonomou & Loizou (2025). The transposed argument covers $m < n$. Write P for the exact polar map, E_t for the matrix error in the computed outer direction, and ξ_t for its operator norm upper bound. For learning rate $\eta > 0$, consider

$$W_{t+1} = W_t - \eta \widehat{Q}_t, \qquad \widehat{Q}_t = P(G_t^{\mathrm{pert}}) + E_t, \qquad \|E_t\|_{\mathrm{op}} \le \xi_t < 1. \tag{21}$$

The factor $1 - \xi_t$ discounts the perturbed gradient norm according to the outer direction error: as the computed direction departs from the exact polar factor, less alignment is guaranteed. The right hand side separates three effects. The initial gap $L(W_0) - L^*$ is amortized over the T steps, the realized radii δ_t contribute through the difference between clean and perturbed gradients, and the final sum accounts for curvature along the updates. The result applies to the deterministic single matrix trajectory specified above, with its realized radii and direction errors. The proof derives these terms from one step smoothness and sums the resulting inequality over the trajectory.

Corollary 2 (Uniform Polar Error Bound)**.** *Under Theorem 3, if the outer direction error satisfies $\xi_t \le \bar{\xi} < 1$ for every $t < T$, where $\bar{\xi}$ is a uniform error bound, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|G_t^{\mathrm{pert}}\|_* \le \frac{1}{1-\bar{\xi}} \left[\frac{L(W_0) - L^*}{\eta T} + \frac{L_{\mathrm{sm}}\sqrt{n}(1+\bar{\xi})}{T} \sum_{t=0}^{T-1} \delta_t + \frac{L_{\mathrm{sm}}\eta n (1+\bar{\xi})^2}{2} \right]. \tag{22}$$

When $\xi_t \le \bar{\xi} < 1$ at every step, each weight $1 - \xi_t$ is at least $1 - \bar{\xi}$. This lets us remove the time varying weights from the left side by dividing the trajectory bound by $1 - \bar{\xi}$, while replacing the error terms on the right by their uniform upper bounds. The resulting estimate controls the unweighted average nuclear norm; it becomes weaker as $\bar{\xi}$ approaches one, reflecting the loss of a useful alignment guarantee. This corollary is obtained directly from Theorem 3.

The one step inequality in the proof also gives a sufficient descent condition: if

$$(1 - \xi_t)\|G_t^{\mathrm{pert}}\|_* > L_{\mathrm{sm}}\sqrt{n}(1+\xi_t)\delta_t + \frac{L_{\mathrm{sm}}\eta n}{2}(1+\xi_t)^2, \tag{23}$$

then $L(W_{t+1}) < L(W_t)$. This condition makes the tradeoff among perturbed gradient alignment, realized perturbation radius, and finite polar error explicit.

Here $\rho > 0$ is the perturbation scale and $\varepsilon > 0$ is the normalization stabilizer defined in Section 2. For the no momentum, single matrix exact polar probe, $\Delta_t = \rho P(G_t)/(\|G_t\|_F + \varepsilon)$ and $\delta_t = \rho\sqrt{\mathrm{rank}(G_t)}/(\|G_t\|_F + \varepsilon)$. With a finite inner direction, use the measured $\delta_t = \rho\|Q_t^{\mathrm{in}}\|_F/(\|G_t\|_F + \varepsilon)$ instead. Thus the bound exposes both radius amplification and outer approximation error. It describes a deterministic gradient update without momentum, weight decay, stochastic batches, or mixed parameter groups; it makes no flat minimum or practical optimizer convergence claim.

D Additional experiments and details

Table 6: Accuracy by corruption (%) on CIFAR-10/ResNet-34. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	G-Blur	H-Blur	Bright	Contr	Satur	Solar	Postr	Sharp	Elast	Rotat	G-Nois	S-Nois	Mean
ADAMW	19.8	16.2	93.6	93.5	94.1	75.7	92.4	89.1	90.7	86.8	35.9	12.0	66.7
ADAMW-SAM	19.2	17.1	94.0	93.8	94.4	77.4	92.5	89.7	91.2	87.5	41.0	13.5	67.6
MUON	24.9	18.1	95.1	95.2	95.5	79.0	93.1	91.5	92.4	89.0	26.4	11.7	67.6
MUONSAM	24.1	19.5	95.4	95.5	95.7	**83.9**	94.1	92.3	93.6	90.7	28.2	14.2	68.9
SAM	21.9	14.5	**96.0**	**96.0**	**96.3**	82.1	**95.1**	**93.6**	**94.0**	**91.4**	**46.6**	**16.0**	**70.3**
SGD	**27.9**	**20.4**	95.1	95.2	95.4	80.2	93.6	92.0	92.4	89.1	41.0	12.4	69.5

Table 7: Accuracy by corruption (%) on CIFAR-100/ResNet-34. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	G-Blur	H-Blur	Bright	Contr	Satur	Solar	Postr	Sharp	Elast	Rotat	G-Nois	S-Nois	Mean
ADAMW	10.9	4.6	71.6	72.0	72.6	36.4	68.8	59.5	66.4	63.1	18.2	2.8	45.6
ADAMW-SAM	12.4	5.4	72.5	72.7	73.3	34.9	69.9	62.0	68.3	64.8	18.0	3.1	46.4
MUON	11.8	5.5	77.3	77.5	78.5	43.6	71.5	67.0	71.3	66.9	15.3	2.5	49.1
MUONSAM	11.7	4.1	**80.1**	**80.7**	**81.2**	**47.5**	74.2	**71.3**	**76.6**	**68.7**	14.5	2.5	**51.1**
SAM	14.6	**6.9**	77.9	78.2	79.1	43.5	**74.5**	69.2	73.3	68.6	19.3	3.8	50.7
SGD	**15.9**	5.9	77.4	77.2	78.0	42.0	73.4	68.3	72.0	67.7	**22.3**	**4.1**	50.3

D.1 Per Corruption Robustness Results

The remaining CIFAR-10/ResNet-34, CIFAR-100/ResNet-34, CIFAR-10/ViT, and language tables report accuracy for each fixed perturbation. The CIFAR-10/ResNet-18 and ImageNet-100/ResNet-34 tables are in Section 4.1.

Table 8: Accuracy by corruption (%) on CIFAR-10/ViT. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	G-Blur	H-Blur	Bright	Contr	Satur	Solar	Postr	Sharp	Elast	Rotat	G-Nois	S-Nois	Mean
ADAMW	65.1	42.9	87.9	88.1	88.2	**76.1**	**88.0**	**85.5**	**86.9**	85.4	**69.0**	**24.8**	**74.0**
ADAMW-SAM	**65.3**	**44.5**	87.9	87.9	88.2	75.1	87.8	84.9	86.2	85.4	**69.0**	24.5	73.9
MUON	61.4	40.2	87.0	87.0	87.2	74.7	86.1	82.5	85.2	84.2	39.7	11.7	68.9
MUONSAM	64.5	42.1	**88.3**	**88.3**	**88.4**	**76.1**	87.0	83.9	86.8	**85.9**	43.9	12.1	70.6
SAM	50.4	42.6	71.9	72.3	72.8	54.8	71.8	65.0	71.4	72.0	45.1	12.7	58.6
SGD	51.9	41.8	74.1	74.4	74.7	57.4	74.1	68.4	73.1	73.4	43.8	14.4	60.1

Table 9: Accuracy by perturbation (%) on SST-2. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	C-Del	C-Ins	C-Swp	C-Sub	W-Del	W-Swp	W-Ins	Tr-50	Tr-80	Pad	Upper	Lower	Spell	Mean
ADAMW	65.0	65.5	64.1	67.4	80.6	82.6	81.9	72.6	**64.3**	82.0	82.6	82.6	82.5	74.9
ADAMW-SAM	69.4	70.0	70.3	71.4	83.3	84.2	85.2	74.0	63.4	84.2	84.5	84.5	84.1	77.6
MUON	73.3	73.3	**72.7**	74.7	85.7	87.2	88.1	73.9	62.3	87.5	88.1	88.1	87.5	80.2
MUONSAM	**73.9**	**74.4**	70.5	**76.7**	**86.1**	**89.1**	**89.6**	**75.2**	64.0	**89.6**	**89.8**	**89.8**	**89.7**	**81.4**
SAM	67.2	66.1	65.7	69.2	82.2	82.9	82.5	74.2	63.2	82.5	82.3	82.3	82.2	75.6
SGD	72.6	70.0	68.3	72.5	85.6	88.3	89.1	74.4	62.2	87.7	88.8	88.8	88.1	79.7

Table 10: Accuracy by perturbation (%) on SST-5/GPT-2. The mean compares MuonSAM with matched Muon; behavior by category varies across optimizers. Best values in each category are bold.

Optimizer	C-Del	C-Ins	C-Swp	C-Sub	W-Del	W-Swp	W-Ins	Tr-50	Tr-80	Pad	Uppr	Lowr	Spel	Mean
ADAMW	36.9	36.5	35.3	37.1	46.0	47.7	46.9	30.7	24.9	46.6	40.0	47.8	47.7	40.3
ADAMW-SAM	39.6	**41.6**	**35.8**	39.2	45.9	46.0	47.7	33.8	25.7	47.5	**44.1**	47.0	47.2	41.6
MUON	33.2	37.5	**35.8**	37.9	45.4	46.0	45.9	**37.2**	28.6	47.0	37.7	47.0	46.6	40.5
MUONSAM	**39.8**	40.8	35.7	**42.0**	**48.9**	**49.7**	**50.3**	32.1	26.7	**51.6**	41.0	**51.5**	**52.0**	**43.2**
SAM	36.7	36.8	32.7	35.0	41.9	42.3	42.6	34.5	28.3	42.7	38.3	42.7	43.1	38.3
SGD	32.4	32.1	32.2	32.8	38.0	38.3	38.1	35.7	**29.9**	34.3	33.1	38.2	38.2	34.9

Datasets and Models. The evaluation spans computer vision (CV) and natural language processing (NLP), covering datasets and architectures with different data and model structures Levi & Oz (2025); Bahri et al. (2022); Kaddour et al. (2022):

- **Datasets**: CIFAR-10, CIFAR-100, and ImageNet-100 for CV tasks; SST-2 and SST-5 for NLP tasks Jiao et al. (2020); Socher et al. (2013); Wang et al. (2019); Radford et al. (2019); Socher et al. (2013).
- **Model Architectures**: We employed ResNet-18 and ResNet-34 (modified for CIFAR; original for ImageNet) He et al. (2016); Russakovsky et al. (2015); Tian et al. (2020), and ViT for vision tasks, alongside TinyBERT and GPT-2 for language tasks.

Training Protocols and Schedulers. Vision and language models use separate training budgets and schedules:

- **Computer vision:** ResNet-18, ResNet-34, ResNet-50, and ViT were trained from scratch for 200 epochs on CIFAR-10, CIFAR-100, and ImageNet-100. We used `CosineAnnealingLR`; the minimum learning rate was 10^{-7} for CIFAR and 10^{-6} for ImageNet-100.
- **Natural language processing:** Pretrained TinyBERT and GPT-2 were fine tuned for 50 epochs on SST-2 and SST-5. We used cosine learning rate decay with linear warmup over the first 10% of optimizer updates.

Unless otherwise stated, all reported accuracies are computed from the final checkpoint after the fixed training budget rather than a best-epoch checkpoint, giving all runs the same checkpoint rule.

ImageNet Training Protocol. To maintain a controlled experimental environment and focus on the intrinsic optimization dynamics, we applied only **standard basic data augmentation** for the ImageNet-100 training set, avoiding complex strategies such as AutoAugment or RandAugment. Specifically, the training pipeline included:

1. `RandomResizedCrop(224)` to introduce scale invariance;
2. `RandomHorizontalFlip()` for spatial mirroring;
3. `ToTensor()` followed by normalization using the ImageNet mean ($[0.485, 0.456, 0.406]$) and standard deviation ($[0.229, 0.224, 0.225]$).

The ImageNet-100 hyperparameter configuration is summarized in Table 11.

Implementation and Optimizers. We compared **SGD, SAM, AdamW, AdamW-SAM, Muon, and MuonSAM**. Optimizer and scheduler implementations use or follow the `pytorch_optimizer` library where applicable Kim (2021); Wolf et al. (2020).

Experiments used NVIDIA RTX 4070 Ti and RTX 4090 GPUs. Unless noted otherwise, runs use seeds 42, 10, and 3407. The additional CIFAR-10/ResNet-18 evaluation uses six seeds: 42, 10, 3407, 112916, 2026, and 777. Tables report the mean and, where available, standard deviation across the corresponding runs.

Table 11: Hyperparameter configurations for training ResNet-34 on the ImageNet-100 dataset. All configurations utilize a batch size of 128, a `CosineAnnealingLR` scheduler spanning 200 epochs, and a minimum learning rate ($\eta_{\min}$) of 1×10^{-6}.

Optimizer	Learning Rate	Weight Decay	Momentum	ρ	AdamW Scale
SGD	1×10^{-1}	1×10^{-4}	0.90	—	—
SAM	1×10^{-1}	1×10^{-4}	0.90	0.05	—
AdamW	1×10^{-3}	0.05	—	—	—
AdamW-SAM	1×10^{-3}	0.05	—	0.05	—
Muon	2×10^{-2}	0.025	0.95	—	2×10^{-2}
MuonSAM	2×10^{-2}	0.025	0.95	0.05	2×10^{-2}

Hyperparameter selection and interpretation. For vision tasks, Muon settings were selected using 50 `Optuna` trials per dataset–model pair Akiba et al. (2019). For all image dataset experiments, we use $\rho = 0.05$ for SAM, AdamW-SAM, and MuonSAM. The tables report optimizer-specific settings for ImageNet-100/ResNet-34, SST-2/TinyBERT, and SST-5/GPT-2; a dash marks a parameter that does not apply. For Muon and MuonSAM, `AdamW Scale` multiplies the Muon learning rate to set the learning rate for non Muon parameter groups. The ρ used by SAM and AdamW-SAM is a Euclidean radius, whereas MuonSAM's ρ is a global scale coefficient. Equal values therefore do not imply equal realized perturbation norms.

Table 12: Hyperparameter configurations for fine tuning TinyBERT on the SST-2 dataset.

Optimizer	Learning Rate	Weight Decay	Momentum	ρ	AdamW Scale	Batch Size
SGD	1×10^{-4}	1×10^{-4}	0.90	—	—	32
SAM	1×10^{-4}	1×10^{-4}	0.90	0.01	—	32
AdamW	5×10^{-5}	0.01	—	—	—	32
AdamW-SAM	5×10^{-5}	0.01	—	0.01	—	32
Muon	2×10^{-4}	0.10	0.95	—	2×10^{-2}	32
MuonSAM	2×10^{-4}	0.10	0.95	0.01	2×10^{-2}	32

Table 13: Hyperparameter configurations for fine tuning GPT-2 on the SST-5 dataset.

Optimizer	Learning Rate	Weight Decay	Momentum	ρ	AdamW Scale	Batch Size
SGD	1×10^{-3}	0.01	0.95	—	—	8
SAM	1×10^{-3}	0.01	0.95	0.05	—	8
AdamW	2×10^{-5}	0.01	—	—	—	8
AdamW-SAM	2×10^{-5}	0.01	—	0.05	—	8
Muon	1×10^{-4}	0.10	0.95	—	2×10^{-2}	32
MuonSAM	1×10^{-4}	0.10	0.95	0.05	2×10^{-2}	32

NLP Fine tuning Protocols. For NLP fine tuning, we used the configurations in Tables 12 and 13, following the flatness aware optimization protocol described by Han et al. (2025). The tables report batch size per optimizer; for SST-5, SGD, SAM, AdamW, and AdamW-SAM use batch size 8, whereas Muon and MuonSAM use batch size 32. All NLP runs use cosine decay with linear warmup over the first 10% of optimizer updates.

D.2 Inner–Outer Direction Diagnostics

We measure how the perturbed gradient changes the orthogonalized update direction. For each Muon matrix block W and step t, let $Q^{\mathrm{in}}_{W,t}$ be the inner approximate polar direction computed from clean gradient tentative momentum, and let $Q^{\mathrm{out}}_{W,t}$ be the outer approximate polar direction computed from perturbed gradient momentum. For nonzero directions, define their Frobenius cosine similarity $\Gamma_{W,t}$ and dissimilarity $D^{\mathrm{IO}}_{W,t}$ by

$$\Gamma_{W,t} = \frac{\langle Q^{\mathrm{in}}_{W,t}, Q^{\mathrm{out}}_{W,t} \rangle_F}{\|Q^{\mathrm{in}}_{W,t}\|_F \|Q^{\mathrm{out}}_{W,t}\|_F}, \qquad D^{\mathrm{IO}}_{W,t} = 1 - \Gamma_{W,t}. \tag{24}$$

Here $\langle A, B \rangle_F = \mathrm{Tr}(A^\top B)$, and $\Gamma_{W,t} \approx 1$ indicates that the direction is almost unchanged by the perturbation; lower values indicate stronger redirection. A complementary relative correction strength is

$$R^{\mathrm{geo}}_{W,t} = \frac{\|Q^{\mathrm{out}}_{W,t} - Q^{\mathrm{in}}_{W,t}\|_F}{\|Q^{\mathrm{in}}_{W,t}\|_F}. \tag{25}$$

The ratio $R^{\mathrm{geo}}_{W,t}$ normalizes the direction change by the inner direction norm. We record these quantities alongside the idealized effective radius $\rho^{\mathrm{eff}}_{W,t}$, clean training loss, and test accuracy. We also compare $\cos(G_{W,t}, G^{\mathrm{pert}}_{W,t})$ with $\cos(\mathcal{O}(G_{W,t}), \mathcal{O}(G^{\mathrm{pert}}_{W,t}))$, where $\cos(A, B) := \langle A, B \rangle_F / (\|A\|_F \|B\|_F)$ for nonzero matrices, to distinguish raw gradient changes from those visible after orthogonalization.

D.3 Timing Protocol and Parameter Groups

We measure implementation overhead separately from accuracy. The benchmark uses a fixed CIFAR-10/ResNet-18 batch on an NVIDIA RTX 4070Ti GPU, synchronizes CUDA before and after each measured step, performs 20 warmup steps, repeats each optimizer five times, and records peak CUDA memory. All SAM optimizers use two forward/backward passes: one gradient evaluation at the original point and one at the perturbed point, followed by the optimizer update. Ordinary SAM applies SGD to all trainable tensors returned by `model.parameters()` and perturbs this

same parameter set using the standard rule in gradient space. AdamW-SAM uses the corresponding AdamW base optimizer and the same convention for selecting parameters.

Muon and MuonSAM share a parameter partition: matrix parameters other than BatchNorm parameters use the Muon channel, while biases, BatchNorm parameters, and scalar/vector parameters use the non Muon AdamW channel. MuonSAM applies a Newton–Schulz orthogonalized SAM perturbation and Muon update to the Muon channel, and a standard SAM perturbation in gradient space followed by an AdamW update to the remaining parameters. Both use 5 Newton–Schulz iterations.

The benchmark results are summarized in Table 4 in the main text.

Newton–Schulz Approximation Error. Muon and MuonSAM use a finite step Newton–Schulz approximation rather than the exact polar decomposition computed by SVD. For each matrix update, the implementation flattens higher order tensors to a 2D matrix X, casts X to bfloat16 on GPU, normalizes by the conservative scale $(\|XX^\top\|_F^{1/2} + 10^{-7})$, and transposes the matrix when needed so that the iteration is applied to the smaller Gram side. We use five Newton–Schulz iterations with coefficients $(3.4445, -4.7750, 2.0315)$.

To quantify the approximation used in practice, we audited a CIFAR-10/ResNet-18 training batch with batch size 128. For each of the 21 Muon channel gradient matrices X, we compared the executed five step Newton–Schulz output $\mathcal{O}_{\mathrm{NS}}^{(5)}(X)$ with the exact polar factor $P(X) = U_r V_r^\top$, where $X = U_r \Sigma_r V_r^\top$ is its compact SVD. This audit applies the map to recorded gradient matrices; the optimizer applies it to momentum matrices. The superscript (5) denotes the iteration count, and the output reflects the implementation's bfloat16 arithmetic. For an output $O \in \mathbb{R}^{m\times n}$, where m and n are its row and column counts and I_k is the $k \times k$ identity matrix, we measure the normalized orthogonality residual as $\|OO^\top - I_m\|_F/\sqrt{m}$ when $m \le n$ and $\|O^\top O - I_n\|_F/\sqrt{n}$ otherwise. We also report the relative polar factor error $\|\mathcal{O}_{\mathrm{NS}}^{(5)}(X) - P(X)\|_F/\|P(X)\|_F$.

Table 14: **Newton–Schulz approximation audit on CIFAR-10/ResNet-18.** The audit compares the implemented Newton–Schulz direction after five steps against the exact SVD polar factor over 21 Muon channel gradient matrices from one training batch.

Metric	Mean	Median	Maximum
Normalized orthogonality residual	0.354	0.353	0.403
Relative polar factor error	0.209	0.202	0.383

Table 14 characterizes the Frobenius reconstruction error and orthogonality of Newton–Schulz outputs on raw gradient matrices. Theorem 2 instead characterizes the exact arithmetic operator error of a finite iteration. For any nonzero input X and computed direction $\widehat{Q}(X)$, write $\xi(X) := \|\widehat{Q}(X) - P(X)\|_{\mathrm{op}}$; in the raw gradient audit, $\widehat{Q}(X) = \mathcal{O}_{\mathrm{NS}}^{(5)}(X)$. With momentum, we also measure the alignment of the exact polar direction of $M_{W,t}$ with the clean gradient $G_{W,t}$. We therefore ran a separate short trajectory audit on the clean gradient tentative momentum matrices used by the inner probe.

Theory Linked Geometry Audit. The audit uses CIFAR-10/ResNet-18, batch size 64, three seeds (42, 10, and 3407), and 50 optimization steps per configuration. Each run uses a fixed subset of $1{,}024$ training examples and evaluates on a fixed subset of 256 examples. We audit six Muon matrix blocks at steps $1, 10, 20, 30, 40, 50$ with $K \in \{3, 5, 7\}$; the $K = 5$ result corresponds to the method's default. For each sampled block and iteration count K, let $Q_{W,t}^{(K),\mathrm{impl}}$ denote the executed bfloat16 Newton–Schulz direction computed from $M_{W,t}$ using K steps. The SVD polar factor $P(M_{W,t})$ uses that same input. We record $\xi_{W,t} = \|Q_{W,t}^{(K),\mathrm{impl}} - P(M_{W,t})\|_{\mathrm{op}}$, $\alpha_{W,t} = \langle G_{W,t}, P(M_{W,t})\rangle/\|G_{W,t}\|_*$, and $\widetilde{\alpha}_{W,t} = \langle G_{W,t}, Q_{W,t}^{(K),\mathrm{impl}}\rangle/\|G_{W,t}\|_*$. The probe comparison uses the same initialization and batch sequence. Its two directions perturb only Muon matrix blocks and are normalized to the same total Frobenius radius ρ.

Across the sampled blocks and steps, the executed map after five steps has positive clean gradient directional alignment, and the measured block radii follow the global normalization formula while varying across blocks. At equal Frobenius radius, the raw gradient probe produces a larger loss

Table 15: **Theory linked geometry diagnostics on CIFAR-10/ResNet-18.** Values are mean $\pm$ standard deviation across seeds unless stated otherwise. The table reports short run diagnostics, not a full training performance comparison.

Diagnostic	Result	Interpretation
Initial exact polar alignment α	1.000001 (18 blocks)	At the first step the tentative momentum is a positive scalar multiple of the clean gradient, giving the predicted unit alignment up to numerical error.
Operator error after five steps ξ	Median 0.324 ± 0.0002; maximum 0.904 ± 0.119	The median is stable across seeds; the worst sampled singular mode has larger approximation error.
Clean gradient alignment α; sufficient condition	0.744 ± 0.017; $\alpha > \xi$ in $90.7 \pm 1.6\%$	The condition in Theorem 2 holds for most audited block–step pairs. It is sufficient, so the remaining pairs are inconclusive.
Executed NS ascent alignment $\widetilde{\alpha} > 0$	100% of audited pairs	Every sampled direction after five steps has positive first order alignment with its clean gradient, including pairs where $\alpha > \xi$ does not certify it.
Realized total radius and matrix radius CV	$\|\Delta_t\|_F/\rho = 14.71 \pm 0.45$; 0.449 ± 0.001	The scale coefficient ρ is not the total parameter space radius; both the total and the block radii vary during training.
Radius formula discrepancy	Maximum relative error 3.5×10^{-6} (3,150 block–step records)	The observed block perturbations agree numerically with the global normalization radius formula.
Loss increase at the first step with equal radii	Polar 0.0812 ± 0.0040; raw gradient 0.1322 ± 0.0109	With only Muon matrices perturbed and equal Frobenius radius, the raw gradient gives larger immediate ascent, as expected for a Frobenius ball probe; the polar direction changes the singular mode weighting.
Inner–outer cosine Γ; relative correction R^{geo}	MuonSAM: 0.953 ± 0.001; 0.283 ± 0.002. No inner momentum: 0.454 ± 0.008; 1.036 ± 0.007	The directions are distinct, and the clean gradient momentum preview is associated with stronger inner–outer alignment in this short run. This diagnostic does not by itself test the momentum identity in Proposition 1.

increase at the first step than the polar probe, consistent with the Frobenius ball first order result in Section 3.1; the polar direction instead changes the weighting of active singular modes. Inner–outer statistics show that the two update directions differ, with higher alignment in the MuonSAM run than in the no inner momentum diagnostic.

Short Run Taylor and Outer Response Audits. We ran two additional CIFAR-10/ResNet-18 diagnostics with batch size 64, a fixed training subset of $8{,}192$ examples, 500 optimization steps, and seeds 42, 10, and 3407. The Taylor audit probes all Muon matrix blocks at steps $1, 50, 100, 200, 350,$ and 500. At each checkpoint, raw gradient, exact clean polar, exact momentum polar, and executed Newton–Schulz momentum directions after five steps are evaluated at matched global Frobenius radii 0.25ρ, 0.5ρ, ρ, and 2ρ. For probe Δ, the Taylor remainder is $R = L(\theta + \Delta) - L(\theta) - \langle \nabla L(\theta), \Delta \rangle$, and the relative remainder is $|R|/|\langle \nabla L(\theta), \Delta \rangle|$. Each direction contributes 72 checkpoint–radius probes across the three seeds.

Table 16: **Taylor remainder diagnostics on CIFAR-10/ResNet-18.** Statistics pool the 72 matched radius probes per direction from three seeds and six checkpoints. The mixed update row uses the actual full parameter MuonSAM perturbation at the same checkpoints.

Probe direction	mean $\lvert R\rvert$	median $\lvert R\rvert/\lvert\langle g, \Delta\rangle\rvert$	Interpretation
Raw matrix gradient	4.56×10^{-3}	4.44%	Larger nonlinear remainder under the matched global radius probes.
Exact clean polar	3.82×10^{-4}	0.38%	Smaller local remainder for the polar direction on these checkpoints.
Exact momentum polar	3.61×10^{-4}	0.41%	Similar remainder scale to the clean polar probe.
Five step NS momentum	3.56×10^{-4}	0.40%	Close to the exact momentum polar diagnostic at the tested radii.
Implemented mixed parameter update (18 probes)	1.83×10^{-2}	4.43%	The measured full parameter perturbation also has a finite Taylor remainder; its radius is not fixed to ρ.

At matched radii, the mean loss change is 0.0480 for the raw direction and 0.0190, 0.0171, and 0.0168 for clean polar, momentum polar, and NS directions, respectively. The smaller polar/NS remainders correspond to a closer local linear approximation in these probes. Implemented mixed

parameter probes have mean realized radius 0.310 (median 0.276), consistent with ρ acting as a scale coefficient rather than the total parameter space radius. The log–log remainder versus radius slopes are near two for raw probes (median 2.01 across 18 fits across checkpoints and seeds); polar and NS slopes vary across the 18 fits.

The outer response audit replays the clean to perturbed momentum relation and parameter update on six sampled matrix blocks at each step, yielding 3,000 records for each matrix and step per seed. Across seeds, the mean relative residual is $(3.51 \pm 0.03) \times 10^{-7}$ for the momentum difference identity and $(2.584 \pm 0.006) \times 10^{-7}$ for the outer update reconstruction. The mean inner–outer cosine is 0.9642 ± 0.0004, with mean relative direction change 0.2565 ± 0.0016. These measurements numerically verify the implementation identities on the sampled trajectory and quantify the direction change after recomputing the perturbed gradient update.

The geometry and Taylor diagnostics characterize local behavior on fixed CIFAR-10/ResNet-18 subsets. Taylor directions are compared at matched radii; the implemented mixed parameter updates retain their realized globally scaled radii. The momentum identity is established analytically and checked numerically on sampled blocks, while the Taylor probes provide a local empirical complement to the smoothness based remainder and direction response bounds.

Future Directions. The next step is to extend momentum input approximation measurements, realized radius tracking, and inner–outer direction comparisons across full training trajectories and additional vision and language settings from the benchmark.