THE ILLUSION OF SPEED: UNCOVERING THE TRADE-OFFS OF MUON'S SPECTRAL FLATTENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has become a promising optimizer for accelerating neural network training by flattening the momentum update spectrum (momentum orthogonalization). Because it often improves model capabilities, the literature has focused on its theoretical understanding and applications while largely overlooking the underlying costs and drawbacks of the acceleration mechanism behind Muon. Our analysis reveals a fundamental tradeoff behind Muon's speed: this mechanism alters implicit biases that help neural networks generalize in non-synthetic, noisy data settings, degrading performance in real-world environments. In practice, by equalizing the singular values of the update step, momentum orthogonalization assigns equal weight to update directions driven by the "clean" data distribution and those driven by rare, weak, or purely noisy data examples. We develop a theoretical account of this tradeoff, showing that spectral equalization encourages high-rank learning (while gradient-based methods are usually known as low-rank learners), removes the natural separation between signal and noise during training, and can move optimization away from simple solutions. We further identify asymmetric label corruption as a realistic failure mode: corrupted examples create coherent but weak directions that Muon promotes relative to Stochastic Gradient Descent (SGD). Experiments with Vision Transformers on CIFAR-100 support this prediction. Muon matches AdamW on clean data but underperforms SGD under label corruption, with a larger gap under asymmetric than symmetric noise. These findings motivate a new direction for optimizer design that retains the speed of spectral equalization while preserving the robustness of amplitude-sensitive learning.

1 INTRODUCTION

As neural networks continue to grow in size, optimization has become a major computational bottleneck, driving the development of optimizers that achieve strong solutions in fewer training steps. Muon (Jordan et al., 2024) has recently emerged as a promising approach, demonstrating substantial speedups and scalability in large-scale language model training (Liu et al., 2025). However, optimization speed alone does not determine an optimizer's practical value: the transformations it applies to gradients also shape which patterns a model learns and how it responds to imperfect data. This distinction matters especially beyond large, curated text corpora, where training data may contain label errors, measurement noise, outliers, and structured corruptions. We therefore ask a broader question: *does Muon's distinctive momentum transformation preserve its optimization advantages when learning from noisy data, or does faster optimization come at the cost of robustness and generalization?*

This question follows directly from Muon's main mechanism, which differs from stochastic gradient descent in one sharply localized way: before applying an update, Muon replaces the gradient (or momentum) matrix with its nearest orthonormal factor, setting every nonzero singular value to one. We call this operation *spectral equalization* and write it $\mathrm{Ortho}(\cdot)$. Its origin is most transparent through a variational reading. Modeling the loss locally around a weight matrix W with gradient $G = \nabla_W \mathcal{L}(W)$ by, $\mathcal{L}(W + \Delta) \approx \mathcal{L}(W) + \langle G, \Delta \rangle + \frac{\lambda}{2}\|\Delta\|^2$, the minimizer is the steepest-descent step under the chosen norm,

$$\Delta = -\frac{\|G\|_*}{\lambda} \operatorname*{arg\,max}_{\|T\|=1} \langle G, T \rangle, \tag{1}$$

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

where $\|\cdot\|_*$ is the dual norm. Optimizers differ mainly in this norm: Adam corresponds to a coordinate-wise (max-of-max) norm, whereas Muon adopts the spectral norm on each matrix (Bernstein & Newhouse, 2024a;b). Writing the singular value decomposition $G = U\Sigma V^\top$, the dual of the spectral norm is the nuclear norm, and the maximizer in equation 1 is $UV^\top$. The Muon update is therefore proportional to $UV^\top$: it preserves the singular *directions* of the gradient but replaces every nonzero singular value by one, mapping every active singular value to the same scale. This is the precise sense in which Muon is *amplitude-blind*: it discards the spectrum Σ that encodes the relative strength of gradient directions, and with it the amplitude hierarchy that SGD preserves. The immediate payoff is a speedup: because all nonzero singular values become one, the condition number of the update is always $\kappa(UV^\top) = 1$. SGD is condition-number limited: stability requires $\eta = \mathcal{O}(1/\lambda_{\max}(H))$, so the slowest Hessian mode contracts only at $1 - \Theta(1/\kappa(H))$. Muon instead equalizes nonzero update singular values ($\kappa = 1$ on its active gradient subspace), increasing weak-mode update energy, but does not by itself guarantee faster objective convergence.

This acceleration is not always benign. On clean, heavy-tailed data such as text, promoting weak modes can be beneficial, because rare but informative patterns may correspond to weak spectral modes that would otherwise be learned slowly. The same mechanism becomes risky when weak modes arise from noise rather than rare signal, since Muon promotes both indiscriminately. Complementary work has shown that pure orthogonalization may be suboptimal (Su, 2025) and that Muon can alter feature-learning dynamics (Dragutinović et al., 2026).

Erasing the singular-value hierarchy therefore creates a fundamental tradeoff. SGD preserves the gradient spectrum, prioritizing dominant modes while suppressing weak ones; this hierarchy supports incremental learning (Kalimeris et al., 2019) and the implicit bias toward low-rank solutions (Gunasekar et al., 2017; Arora et al., 2019). Muon instead equalizes all active singular values, accelerating weak modes whether they represent signal or noise, and thereby trading optimization speed for selectivity between robust and spurious features. We make this cost precise by analyzing the setting in which gradients contain strong signal modes and weak noise modes, and we trace a single mechanism, the removal of the amplitude hierarchy, through four consequences. Because the equalized update spreads its energy as uniformly as possible across the active singular directions, it *maximizes* the nuclear norm of the update and so opposes low-rank bias (Theorem 3.1). Because it discards the magnitudes that keep weak noise subordinate to strong signal, it inflates the noise-to-signal share of the update energy, assigning the noise subspace at least α^2 more relative update energy than SGD under a signal-to-noise spectral gap $\alpha > 1$ (Theorem 3.2). Because it forces strong and weak modes to a common learning rate, it collapses the time-scale separation that lets robust structure be fit before weak or corrupted structure, closing the window that early stopping exploits (Lemma 3.3). And because it converts a small residual gradient into a rank-scale step, it induces a nonvanishing drift of magnitude $\sqrt{l}$ near a clean optimum, removing the local braking that lets amplitude-sensitive optimizers settle into clean solutions (Theorem 3.4). Whether erasing the hierarchy is benign thus depends on the data: on clean signals, equalization mainly forfeits inductive bias, but when the data contain a coherent weak noise mode, it actively promotes that mode to parity with the signal. We show that asymmetric label corruption produces exactly such a mode, and our analysis connects Muon's update geometry to reduced robustness under structured label noise (Section 4).

Contributions. We derive the following theoretical and empirical results: **Loss of low-rank bias.** We prove that, for a fixed step size and active rank, spectral equalization maximizes the nuclear norm of the update. Muon therefore produces the least spectrally concentrated update on the active subspace, opposing the low-rank bias of amplitude-sensitive optimizers (Theorem 3.1). **Loss of spectral rate separation.** We analyze a two-mode model with amplitude-dependent versus equal rates and show that equal rates remove the separation between strong and weak modes. This removes the early-stopping window during which robust structure can be learned before weaker patterns (Lemma 3.3). **Noise amplification and drift.** We prove that, under a signal-to-noise spectral gap $\alpha > 1$, Muon assigns the noise subspace at least α^2 more update energy relative to the signal than SGD (Theorem 3.2). We further show that a residual noise gradient of rank l induces a nonvanishing drift of magnitude $\sqrt{l}$ near a clean optimum (Theorem 3.4). **Structured label noise.** We derive how asymmetric label corruption creates coherent weak spectral modes in both a Gaussian mixture model and the last layer gradient of a deep linear network (Section 4). Experiments confirm that Muon promotes these modes and is less robust than SGD as structured label noise increases. Together, these results characterize a fundamental cost of Muon's acceleration. Spectral equalization speeds

the learning of weak modes, which can help recover rare signal in clean data, but it also promotes coherent noise when the data are corrupted.

2 Related Work

Gradient preconditioning and orthogonalization. Preconditioning is central to the efficiency of modern optimizers. Notable strategies include signed gradients (Bernstein et al., 2018; Crawshaw et al., 2022; Chen et al., 2023; Kunstner et al., 2023), gradient clipping (Zhang et al., 2020), normalization (Zhang et al., 2020; You et al., 2019), and gradient whitening (Jordan et al., 2024; Yang & Laaksonen, 2008; Kingma & Ba, 2015; Hwang, 2024; Bernstein & Newhouse, 2024a;b; Carlson et al., 2015). A recent line of work (Jordan et al., 2024; Tuddenham et al., 2022) uses gradient *orthogonalization* to accelerate training: constraining updates to directions of uniform magnitude (spectral radius one) prevents any single direction from dominating and mitigates instabilities from ill-conditioned gradients. Our analysis isolates the flip side of this design: removing amplitude information also removes the spectral hierarchy on which amplitude-sensitive implicit biases rely.

Theory of Muon. Muon admits convergence and spectral-norm steepest-descent/non-Euclidean trust-region views (Bernstein & Newhouse, 2024a; Li & Hong, 2025; Kovalev, 2025; Sato et al., 2025), analyses of practical inexact Newton–Schulz updates (Shulgin et al., 2025) and spectral-norm implicit regularization (Li et al., 2025), low-rank Hessian exploitation (Shen et al., 2025), and spectral-norm max-margin implicit bias on separable data (Fan et al., 2025). Unlike prior work on converged directions, we study update geometry: equalization reallocates signal/noise energy, erases rate separation, and harms learning under coherent label noise, complementing Su (2025); Dragutinović et al. (2026).

Learning with noisy labels. Deep networks can fit arbitrary label noise (Zhang et al., 2017; Arpit et al., 2017), yet they tend to learn clean patterns before memorizing corrupted ones, a "clean-first" dynamic that makes early stopping and related regularizers effective defenses (Li et al., 2020; Liu et al., 2020). Robustness methods further exploit noise-specific structure, including asymmetric (class-conditional) corruption (Patrini et al., 2017; Han et al., 2018; Song et al., 2022). Our contribution to this literature is on the optimizer side: we show that the clean-first ordering is a *rate-based* phenomenon that amplitude-sensitive optimizers preserve but that spectral equalization removes, and that asymmetric label noise generates exactly the coherent, low-rank gradient mode that equalization promotes to parity with the signal.

Further related work discussion appears in Appendix A.15.

3 Theoretical Results: Spectral Equalization and Its Costs

Equalization maximizes the nuclear norm and encourages high-rank bias. Among all updates of a fixed Frobenius norm supported on the same active subspace, the equalized update spreads its energy as uniformly as possible across the active singular directions.

Theorem 3.1 (Singular-value equalization maximizes the nuclear norm)**.** *Let $G \in \mathbb{R}^{m\times d}$ have rank $r \geq 1$ and singular value decomposition $G = U\Sigma V^\top$. For a fixed $\rho > 0$, define the normalized gradient and equalized updates $\tilde{G}_{\mathrm{grad}} = -\rho \frac{G}{\|G\|_F}$, and $\tilde{G}_{\mathrm{eq}} = -\rho \frac{UV^\top}{\|UV^\top\|_F}$. Then, $\|\tilde{G}_{\mathrm{grad}}\|_F = \|\tilde{G}_{\mathrm{eq}}\|_F = \rho, \quad \|\tilde{G}_{\mathrm{grad}}\|_* \leq \|\tilde{G}_{\mathrm{eq}}\|_* = \rho\sqrt{r}$. The inequality is strict whenever the nonzero singular values of G are not all equal.*

Note that all proofs of all theorems and lemmas are relegated to Appendix A.

Two quantities are in tension here. The Frobenius norm fixes the total energy of the update; the nuclear norm measures how that energy is spread across singular directions, and for fixed energy and rank it is largest exactly when all singular values are equal. Equalization sits at that extreme, producing the least spectrally concentrated update the active subspace admits. Since low-rank bias is precisely the tendency to concentrate energy in a few dominant directions, equalization pushes the update the other way: it spreads energy uniformly and, with it, dissolves the low-rank preference that amplitude-sensitive methods inherit from the gradient spectrum. Because low-rank-biased updates

naturally concentrate energy in a few dominant directions, Theorem 3.1 establishes that equalization drives the update in the exact opposite direction.

Equalization amplifies the noise-to-signal update energy. We now focus on the setting most relevant for learning under noise. When the active subspace splits into orthogonal signal and noise directions, the magnitude hierarchy that equalization erases is exactly what keeps weak noise subordinate to strong signal

Theorem 3.2 (Spectral equalization amplifies noise)**.** *Let G_k be the complete matrix to which exact spectral equalization is applied, such as a layer gradient or momentum matrix, and suppose it admits a signal plus noise decomposition $G_k = G_k^s + G_k^n$, $G_k^s = \sum_{i=1}^{r_s} \sigma_i^s u_i v_i^\top$, $G_k^n = \sum_{j=1}^{r_n} \sigma_j^n a_j b_j^\top$, where the signal and noise singular subspaces are orthogonal on both sides. Assume $r_s, r_n > 0$ and a spectral gap $\alpha\sigma_{\max}^n \le \sigma_{\min}^s$ for some $\alpha > 1$. Let*

$$\Delta_{\mathrm{SGD}} = -\eta\, G_k, \quad \Delta_{\mathrm{Muon}} = -\eta\, \mathrm{Ortho}(G_k) = -\eta \left(\sum_{i=1}^{r_s} u_i v_i^\top + \sum_{j=1}^{r_n} a_j b_j^\top \right).$$

Then the noise-to-signal update-energy ratio satisfies

$$\frac{\|\Delta_{\mathrm{Muon}}^n\|_F^2}{\|\Delta_{\mathrm{Muon}}^s\|_F^2} \ge \alpha^2\, \frac{\|\Delta_{\mathrm{SGD}}^n\|_F^2}{\|\Delta_{\mathrm{SGD}}^s\|_F^2}.$$

In particular, Muon assigns the noise subspace a noise-to-signal update-energy ratio at least α^2 times larger than SGD.

The factor α^2 makes the cost concrete: whatever spectral gap separates signal from noise in the gradient, Muon suppresses its effect on the update by that gap squared. Read distributionally, this is intuitive. SGD preserves the amplitude structure of the empirical gradient, so directions carried by frequent, robust examples dominate those carried by rare, weak, or noisy ones. Equalization discards those magnitudes, so a weak but nonzero direction, including one produced by rare or corrupted examples, can draw the same per-direction update energy as a dominant signal direction. Muon does not change the data distribution; it changes the distribution as seen through the optimizer, upweighting the tail relative to the bulk.

Equalization removes rate-based separation and early stopping. While Theorem 3.2 describes the geometric impact on a single update, these effects compound dynamically over training time. Good generalization is often attributed to learning strong, robust features before weak, spurious, or noisy ones, creating a window in which early stopping retains the former and discards the latter. The following lemma isolates this mechanism spectrally and shows that equalization actively destroys it.

Lemma 3.3 (Equalization removes the early-stopping separation)**.** *Consider two decoupled spectral modes with singular values $\sigma_{\max} > \sigma_{\min} > 0$. Let $e_{\max}(t)$ and $e_{\min}(t)$ denote their residual errors, with $|e_{\max}(0)| = |e_{\min}(0)| = e_0 > \varepsilon > 0$, and define the hitting time $\tau_i := \inf\{t \ge 0 : |e_i(t)| \le \varepsilon\}$. Consider two flows:*

1. *amplitude-sensitive flow $\dot{e}_i(t) = -\rho(\sigma_i)\, e_i(t)$, where $\rho(\sigma) > 0$ is strictly increasing in σ.*
2. *equalized flow $\dot{e}_i(t) = -c\, e_i(t)$, where $c > 0$ is independent of σ_i.*

Then, for the amplitude-sensitive flow, $\tau_i^{\mathrm{amp}} = \frac{1}{\rho(\sigma_i)} \log\!\left(\frac{e_0}{\varepsilon}\right)$, so that $\tau_{\max}^{\mathrm{amp}} < \tau_{\min}^{\mathrm{amp}}$, whereas for the equalized flow $\tau_{\min}^{\mathrm{eq}} = \tau_{\max}^{\mathrm{eq}} = \frac{1}{c} \log\!\left(\frac{e_0}{\varepsilon}\right)$.

The two flows differ in exactly one respect, and it is decisive. Under amplitude sensitivity, the strong mode reaches tolerance strictly before the weak one, opening an interval $(\tau^{\mathrm{amp}}\max, \tau^{\mathrm{amp}}\min)$ in which robust structure has been fit but weak, spurious, or noisy structure has not, the very window early stopping exploits. Equalization imposes a common rate, forcing $\tau^{\mathrm{eq}}\min = \tau^{\mathrm{eq}}\max$ and collapsing that window: strong and weak active modes are now learned together, erasing the rate-based simplicity bias that SGD-like dynamics enjoy. This is the early-stopping consequence of the equal-rate feature learning reported by Dragutinović et al. (2026).

Equalization destabilizes clean optima. We examine behavior at a stationary point of the clean objective. Near such a point, SGD naturally slows as the residual gradient shrinks; equalization removes this stabilizing effect, converting a small residual noise gradient into a rank-scale update whose size is independent of the noise magnitude.

Theorem 3.4 (Amplitude-blind updates cause non-vanishing drift at clean optima)**.** *Let the noisy objective decompose as $\mathcal{L}(W) = \mathcal{L}^s(W) + \mathcal{L}^n(W)$, where $\mathcal{L}^s$ is the clean-signal loss and $\mathcal{L}^n$ the noise-induced loss. Let W^* be a stationary point of the clean loss, $\nabla\mathcal{L}^s(W^*) = 0$, and suppose the residual noise gradient is nonzero with rank $l > 0$: $G^n(W^*) := \nabla\mathcal{L}^n(W^*) \neq 0, \quad \operatorname{rank}\big(G^n(W^*)\big) = l$. Consider the SGD update $\Delta W_{\mathrm{SGD}} = -\eta_{\mathrm{SGD}}\, \nabla\mathcal{L}(W^*)$ and the equalized (amplitude-blind) update $\Delta W_{\mathrm{AB}} = -\eta_{\mathrm{AB}}\, \mathrm{Ortho}(\nabla\mathcal{L}(W^*))$. Then $\|\Delta W_{\mathrm{SGD}}\|_F = \eta_{\mathrm{SGD}}\, \|G^n(W^*)\|_F, \quad \|\Delta W_{\mathrm{AB}}\|_F = \eta_{\mathrm{AB}}\sqrt{l}$. Therefore, the amplitude-blind update moves farther from the clean stationary point than SGD whenever $\eta_{\mathrm{AB}}\sqrt{l} > \eta_{\mathrm{SGD}}\, \|G^n(W^*)\|_F$.*

The gap is starkest when the residual noise gradient is spectrally small but occupies many thin directions. Its Frobenius norm $|G^n(W^*)|_F$ can be made arbitrarily small, whereas $\sqrt{l}$ depends only on how many directions are active, not on their size. SGD, whose step scales with the residual gradient, then drifts only slightly from the clean optimum; the equalized update drifts by a rank-scale amount that does not vanish with the noise. Equalization thus removes the local braking that lets amplitude-sensitive optimizers settle into clean solutions, the stationary-point counterpart of the energy reallocation in 3.2.

A single mechanism: Taken together, Theorems 3.1, 3.2, 3.4 and Lemma 3.3 trace *one* mechanism, the removal of the gradient's singular-value hierarchy, through the geometry of a single update, the allocation of signal and noise energy, the dynamics of feature learning, and behavior at clean optima. That hierarchy supports amplitude-sensitive implicit regularization, low-rank bias, rate-based feature selection, robustness to rare and corrupted examples, and local stability near clean optima. Whether erasing it is benign depends on the data: on clean signals, equalization mainly forfeits inductive bias, but when the data contain a coherent weak noise mode, it actively promotes that mode to parity with the signal. The next section shows that asymmetric label corruption produces exactly such a mode.

4 A Concrete Failure Mode: Coherent Label Noise

The results of Section 3 hinge on one structural assumption: that noise enters the gradient as a *coherent* spectral mode, weak relative to the signal but not an isotropic, zero-mean perturbation. Isotropic noise would spread its energy across all directions and average away, leaving equalization little coherent noise to amplify. We now show this assumption is not a convenience but a generic feature of label noise: under asymmetric corruption, mislabeled examples induce a coherent, low-rank supervised component, exactly the weak mode that equalization promotes to parity.

The argument follows one template in four settings of increasing generality. Corruption produces a coherent supervised component whenever a subpopulation is consistently observed with the wrong label; this component becomes a *distinct singular mode* precisely when the corrupted subpopulation has a feature or representation direction separated from the dominant one; and once a distinct weak mode with a spectral gap exists, Theorem 3.2 and Lemma 3.3 apply unchanged. We fix the corruption model and the separation condition in §4, record the shared conclusion once in §4, and then instantiate it for Gaussian-mixture data (§4), for any linear readout (§4), for deep linear networks (§4), and for deep homogeneous ReLU networks (§4).

The corruption model and the coherence condition. Subpopulation-dependent label corruption: A semantic class need not correspond to a single feature distribution. A class may contain canonical examples alongside atypical ones: rare poses, unusual backgrounds, ambiguous instances, or examples close to another class. We model this by letting class a contain a dominant subpopulation with mean μ_a together with a smaller atypical subpopulation with mean $\nu_{a\to b}$. Both are semantically class a, but the atypical subpopulation is more likely to be *observed* with label e_b, because its features are more easily confused with class b. The corruption is therefore instance-dependent rather than uniform over the class. The model is trained only on the observed pairs (x_i, y_i); the semantic description merely motivates the data-generating process, and what matters for optimization is the observed supervised signal.

Assumption 4.1 (Structured subpopulation label noise). Fix two labels $a \neq b$; all examples considered have true semantic label e_a. The class-a distribution contains two subpopulations: (1) $i \in I_a$: $x_i = \mu_a + z_i,\ y_i = e_a,$ with $|I_a| = n_a$; (2) $i \in I_{a\to b}:\ x_i = \nu_{a\to b} + z_i,\ y_i = e_b$ with $|I_{a\to b}| = n_b$ where $z_i \sim \mathcal{N}(0, I_d)$, and $0 < n_b < n_a$. Here I_a is the dominant, correctly observed subpopulation and $I_{a\to b}$ the smaller atypical subpopulation observed with the noisy label e_b, with feature means μ_a and $\nu_{a\to b}$ respectively. Assume the observed label directions and the two feature directions are separated, $e_a^\top e_b = 0, \quad \mu_a^\top \nu_{a\to b} = 0, \quad \mu_a \neq 0, \quad \nu_{a\to b} \neq 0,$ and that the noisy subpopulation is spectrally weaker, $0 < n_b\|\nu_{a\to b}\|_2 < n_a\|\mu_a\|_2$.

The coherence condition: The orthogonality $\mu_a^\top \nu_{a\to b} = 0$ is a transparent idealization, not the essential requirement, but some separation is genuinely needed. Had the two observed groups shared the feature direction μ_a, their contribution would factorize as $n_a e_a \mu_a^\top + n_b e_b \mu_a^\top = (n_a e_a + n_b e_b)\mu_a^\top$, which is rank one: the corrupted labels would create a coherent supervised bias but not a separate singular mode, and the amplification below would not apply. More generally, writing $u_a := \mu_a/\|\mu_a\|_2$, $\nu_{a\to b} = \nu_\parallel + \nu_\perp$ with $\nu_\parallel = (u_a^\top \nu_{a\to b})u_a$, and $\nu_\perp = (I - u_a u_a^\top)\nu_{a\to b}$, $n_a e_a \mu_a^\top + n_b e_b \nu_{a\to b}^\top = \big(n_a\|\mu_a\|_2 e_a + n_b(u_a^\top \nu_{a\to b})e_b\big)u_a^\top + n_b e_b \nu_\perp^\top$, so the parallel component merges into the dominant direction while $\nu_\perp$ creates a second right-side feature direction. Assuming $\mu_a^\top \nu_{a\to b} = 0$ sets $\nu_\parallel = 0$ and makes the two-mode calculation exact; the essential requirement is only $\nu_\perp \neq 0$.

The shared conclusion Each setting below exhibits a layer gradient with two orthogonal components, a dominant observed mode and a corrupted-label mode, separated by a spectral gap. The consequence is always the same, so we state it once and instantiate it thereafter.

Proposition 4.2 (Coherent label noise triggers amplification and rate collapse)**.** *Let a layer gradient decompose as $G = G^s + G^n$, where the singular directions of G^s and G^n are mutually orthogonal with singular values σ_a and σ_b, and suppose $\alpha\,\sigma_b \leq \sigma_a$ for some $\alpha > 1$. Then the idealized Muon update assigns the corrupted-label mode a larger relative update-energy ratio than SGD,*

$$\frac{\|\Delta^n_{\mathrm{Muon}}\|_F^2}{\|\Delta^s_{\mathrm{Muon}}\|_F^2} \;\geq\; \alpha^2\,\frac{\|\Delta^n_{\mathrm{SGD}}\|_F^2}{\|\Delta^s_{\mathrm{SGD}}\|_F^2}, \tag{2}$$

and if the corresponding residual dynamics satisfy Lemma 3.3, then amplitude-sensitive dynamics admit an early-stopping window in which the dominant observed mode is learned before the weak corrupted-label mode, whereas equalized dynamics remove this rate separation.

Instantiation I: Gaussian-mixture data. At the data level the corrupted subpopulation appears directly in the observed data-label matrix.

Lemma 4.3 (Subpopulation label corruption creates orthogonal observed components)**.** *Consider C-class Gaussian-mixture data with observed labels $y_i \in \{e_1, \ldots, e_C\}$, and let Assumption 4.1 hold. Define the observed data-label matrix $M := \sum_{i\in I_a \cup I_{a\to b}} y_i x_i^\top$. Then $\mathbb{E}[M] = n_a e_a \mu_a^\top + n_b e_b \nu_{a\to b}^\top$. Moreover, the two displayed rank-one terms are orthogonal on both their left and right sides. Consequently, they are singular components of $\mathbb{E}[M]$.*

Corollary 4.4. *Suppose $0 < n_b\|\nu_{a\to b}\|_2 < n_a\|\mu_a\|_2$ in Lemma 4.3, then the observed subpopulation labeled e_b creates a weak active singular mode when $\sigma_a = n_a\|\mu_a\|_2$, and $\sigma_b = n_b\|\nu_{a\to b}\|_2$.*

Corollary 4.5 (Muon amplifies the weak corrupted-label singular mode)**.** *In the setting of Lemma 4.3 and Corollary 4.4, suppose the two singular components of $\mathbb{E}[M]$ appear in the layer gradient as signal and corrupted-label components $G^s = n_a e_a \mu_a^\top, \quad G^n = n_b e_b \nu_{a\to b}^\top$, and assume a spectral gap $\alpha\sigma_b \leq \sigma_a$, $\alpha > 1$, where $\sigma_a = n_a\|\mu_a\|_2$ and $\sigma_b = n_b\|\nu_{a\to b}\|_2$. Then the hypotheses of Proposition 4.2 hold, and the idealized Muon update assigns the corrupted-label mode a larger relative update-energy ratio than SGD, follows equation 2. Moreover, if the corresponding residual dynamics satisfy the assumptions of Lemma 3.3, then amplitude-sensitive dynamics admit an early-stopping window in which the dominant observed mode is learned before the weak corrupted-label mode, whereas equalized dynamics remove this rate separation.*

Instantiation II: any linear readout. The mechanism does not require the data to be a raw Gaussian mixture, and holds for any network with a linear readout. The final layer sees the training data only through the learned representation $\phi_\theta(x)$, so observed label corruption creates a structured last-layer gradient component whenever a group of examples is consistently observed with an incorrect label. The existence of this coherent component is architecture-independent; whether it becomes a distinct singular mode depends on the geometry of the learned representations.

Lemma 4.6 (Subpopulation corruption creates a distinct last-layer mode)**.** *Consider a deep network whose final layer is linear, $f_\Theta(x) = W_L\phi_\theta(x)$, where $\phi_\theta(x) \in \mathbb{R}^m$ is the representation produced by all layers before the final layer, trained with squared loss $\mathcal{L}(\Theta) = \frac{1}{2}\sum_{i=1}^n \|W_L\phi_\theta(x_i) - y_i\|_2^2$ on observed labels $y_i \in \{e_1, \dots, e_C\}$. Let Assumption 4.1 hold and define the representation means $m_a(\theta) := \mathbb{E}[\phi_\theta(x_i) \mid i \in I_a]$ and $m_b(\theta) := \mathbb{E}[\phi_\theta(x_i) \mid i \in I_{a\to b}]$. Then the expected label-dependent contribution of the two subpopulations is $H_L^{(a,b)} = -n_a e_a m_a(\theta)^\top - n_b e_b m_b(\theta)^\top$. If $m_a(\theta)^\top m_b(\theta) = 0$, $m_a(\theta) \neq 0$, and $m_b(\theta) \neq 0$, then the two terms are singular components of $H_L^{(a,b)}$, with singular values $\sigma_a = n_a\|m_a(\theta)\|_2, \sigma_b = n_b\|m_b(\theta)\|_2$.*

Corollary 4.7 (Conditional amplification of the last-layer corrupted-label mode)**.** *In the setting of Lemma 4.6, let G_L denote the complete last-layer matrix to which idealized Muon equalization is applied. Assume that the two components of $H_L^{(a,b)}$ remain singular components of G_L. If $0 < \alpha\sigma_b \le \sigma_a$ for some $\alpha > 1$, then the relative-amplification conclusion of Proposition 4.2 applies at the last layer. If the corresponding residuals additionally obey the flows assumed in Lemma 3.3, its hitting-time conclusion also applies.*

Instantiation III: deep linear networks. The same two components appear at every layer, modulated by the surrounding weight products.

Lemma 4.8 (Layerwise corrupted-label modes under transformed orthogonality)**.** *Consider the depth-$L \ge 2$ linear network $f_\Theta(x) = W_L W_{L-1}\cdots W_1 x$ trained with squared loss on observed labels. For $1 \le \ell \le L$, define $A_{\ell-1} := W_{\ell-1}\cdots W_1$, $B_{\ell+1} := W_L\cdots W_{\ell+1}$, with $A_0 = I$ and $B_{L+1} = I$. Let Assumption 4.1 hold. Assume additionally that, for every $1 \le \ell \le L$, $e_a^\top B_{\ell+1}B_{\ell+1}^\top e_b = 0, \mu_a^\top A_{\ell-1}^\top A_{\ell-1}\nu_{a\to b} = 0$, and that $B_{\ell+1}^\top e_a$, $B_{\ell+1}^\top e_b$, $A_{\ell-1}\mu_a$, $A_{\ell-1}\nu_{a\to b}$ are nonzero. Then the expected label-dependent contribution of the two subpopulations I_a and $I_{a\to b}$ is $H_\ell^{(a,b)} = -n_a B_{\ell+1}^\top e_a(A_{\ell-1}\mu_a)^\top - n_b B_{\ell+1}^\top e_b(A_{\ell-1}\nu_{a\to b})^\top$. The two displayed rank-one terms are mutually bi-orthogonal singular components of $H_\ell^{(a,b)}$, with singular values $\sigma_{a,\ell} = n_a\|B_{\ell+1}^\top e_a\|_2\|A_{\ell-1}\mu_a\|_2$, $\sigma_{b,\ell} = n_b\|B_{\ell+1}^\top e_b\|_2\|A_{\ell-1}\nu_{a\to b}\|_2$.*

Corollary 4.9. *In the setting of Lemma 4.8, suppose that, for every ℓ, the two components of $H_\ell^{(a,b)}$ remain singular components of the complete matrix G_ℓ to which idealized Muon equalization is applied. Suppose further that $0 < \alpha\sigma_{b,\ell} \le \sigma_{a,\ell}$ for some $\alpha > 1$. Then the hypotheses and the relative-amplification conclusion of Proposition 4.2 hold at every layer. If the corresponding residuals additionally obey the flows assumed in Lemma 3.3, its hitting-time conclusion also holds layerwise.*

Remark 4.10 (Layerwise assumptions). Corollary 4.9 is a conditional layerwise statement, not a global analysis of deep-linear training. The products $B_{\ell+1}$ and $A_{\ell-1}$, as well as the remaining terms in G_ℓ, can rotate or mix the two components. Therefore, the singular-mode and spectral-gap conditions must hold at every layer and, for a training trajectory, at every relevant iterate. The early-stopping conclusion is additionally conditional on the residual dynamics assumed in Lemma 3.3.

Instantiation IV: deep homogeneous networks. Consider a depth-L ReLU network without biases, $f_\Theta(x) = W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_1x)))$, $\sigma = \mathrm{ReLU}$, trained with squared loss on observed labels. Write the gates $D_\ell(x) = \mathrm{diag}(\mathbf{1}[a_\ell(x) > 0])$ with $D_L := I$. Since σ is positively 1-homogeneous, the ReLU identity $\sigma(a) = D(a)a$ makes the forward map exactly linear once the gates are fixed. Defining the gated products: $A_{\ell-1}(x) := D_{\ell-1}(x)W_{\ell-1}\cdots D_1(x)W_1$, where $A_0 := I$, and $B_{\ell+1}(x) := W_L D_{L-1}(x)\cdots D_{\ell+1}(x)W_{\ell+1}$, where $B_{L+1} := I$, gives $f_\Theta(x) = B_{\ell+1}(x)D_\ell(x)W_\ell A_{\ell-1}(x)x$. These play the same role as the layerwise linear products; replacing all gates by the identity recovers the deep-linear case exactly.

Corollary 4.11 (Layerwise corrupted-label modes in deep homogeneous ReLU)**.** *Fix an activation pattern, assume the gates are common to all examples within each subpopulation, and assume that Lemma 4.8 holds for the corresponding gated products with orthogonal left factors $D_\ell B_{\ell+1}^\top e_a$ and $D_\ell B_{\ell+1}^\top e_b$. If its two modes remain singular components of the complete gradient G_ℓ and $0 < \sigma_{b,\ell} < \sigma_{a,\ell}$ for all ℓ, then Proposition 4.2 applies at every layer.*

5 EXPERIMENTS

Three experiments test three distinct predictions: loss of low-rank bias (Theorem 3.1), noise amplification (Theorem 3.2, Proposition 4.2), and collapse of rate separation (Lemma 3.3, Theorem 3.4). The first two use standard baselines (AdamW, SGD) on CIFAR-100 where scale and data regime are well characterized. The third adds an exploratory MNIST comparison with SUMO, a low-rank subspace variant of Muon (k=16) that equalizes only the top-k singular directions. Protocols, measurement definitions, and additional figures are in Appendix B.

Coherent noise selectively harms equalized updates. We train a ViT-B/16 classifier on CIFAR-100 under two noise models at matched corruption rates r: *asymmetric* label noise, the setting modeled by Lemmas 4.3, 4.6, and 4.8, and *white* (symmetric) label noise. On clean data, SGD trails AdamW and Muon by roughly one point, yet under asymmetric noise the ordering inverts: by r=0.3 SGD leads Muon and AdamW by roughly 4 and 5 points (Table 1). Under white noise, the gap is smaller, and all three optimizers degrade gracefully. The larger SGD–Muon gap under asymmetric corruption is consistent with the proposed mechanism (§4), although these experiments do not verify the required gradient-subspace assumptions.

Optimizer	r=0	r=0.1	r=0.2	r=0.3	r=0.4
			Asymmetric label noise		
AdamW	**0.8859**$\pm$0.0021	0.8603$\pm$0.0039	0.8383$\pm$0.0066	0.7991$\pm$0.0071	0.7413$\pm$0.0058
Muon	0.8852$\pm$0.0018	0.8677$\pm$0.0029	0.8420$\pm$0.0056	0.8119$\pm$0.0114	0.7461$\pm$0.0102
SGD	0.8735$\pm$0.0010	**0.8698**$\pm$0.0021	**0.8661**$\pm$0.0013	**0.8518**$\pm$0.0018	**0.7967**$\pm$0.0048
			White label noise		
AdamW	**0.8859**$\pm$0.0021	0.8479$\pm$0.0052	0.8281$\pm$0.0042	0.8210$\pm$0.0035	0.8012$\pm$0.0060
Muon	0.8852$\pm$0.0018	0.8497$\pm$0.0016	0.8401$\pm$0.0041	0.8312$\pm$0.0042	0.8214$\pm$0.0042
SGD	0.8735$\pm$0.0010	**0.8653**$\pm$0.0028	**0.8569**$\pm$0.0025	**0.8494**$\pm$0.0023	**0.8416**$\pm$0.0031

Table 1: Clean test accuracy of a ViT-B/16 classifier on CIFAR-100, mean $\pm$ standard deviation. Bold entries mark the best optimizer per column. SGD leads from r=0.1 onward under asymmetric noise and is also the most robust under white noise; AdamW leads on clean data (r=0).

Equalized updates are maximally spread; weights stay high-rank. Theorem 3.1 bounds the update spread $\|\Delta\|_*/\|\Delta\|_F \in [1, \sqrt{r}]$. Training two compact ViTs on CIFAR-100 that differ *only* in the update applied to the 16 two-dimensional hidden matrices, this quantity separates sharply: SGD attains 2.64, Muon 10.56, essentially saturating $\sqrt{r} = 11.31$ (Table 2). Empirical gradients are therefore far from equalized, and equalization moves the update to the opposite extreme of the admissible range. The effect propagates to the weights: SGD contracts the mean stable rank from 44.29 at initialization to 7.61, retaining 17%, whereas Muon retains 72%. Muon's weights travel 9.3$\times$ further in Frobenius norm yet retain far more initial stable rank: equalized updates fail to acquire low-rank structure *despite* the larger displacement, itself the signature of Theorem 3.4. Muon also achieves lower training loss (3.06 vs. 3.59) and higher clean accuracy: this is not learning failure but the acceleration of §1 bought at the cost of Theorem 3.1's bias. The separation is a property of the update rule, unchanged under 20% asymmetric and white corruption (Appendix B.1).

Label-noise robustness: MNIST results. We also train Adam, SGD, Muon, and the low-rank subspace variant SUMO (k=16) on MNIST with coherent and symmetric label noise at rates 0.3–0.5. Tables 3 and 4 give the quantitative summary. At r=0.3, Muon fits the noisy training labels above 98% and its clean-test accuracy falls by roughly 18–20 points from its peak, whereas Adam and SUMO fit them at 69–76%. Under symmetric noise, SUMO outperforms Adam by 5–9 points. These exploratory MNIST comparisons illustrate different memorization behavior under the reported configurations; unequal tuning and implementation differences prevent attributing the gaps to spectral treatment alone.

	SGD	Muon
Update spread $\|\Delta\|_*/\|\Delta\|_F$ (max $\sqrt{r} = 11.31$)	2.64	**10.56**
Stable rank $\|W\|_F^2/\|W\|_2^2$	**7.61**	31.87
retained vs. initialization (44.29)	**17%**	72%
Displacement $\|W - W_0\|_F$	3.67	**34.09**
Final training loss	3.59	**3.06**
Clean test accuracy	0.1315	**0.1685**

Table 2: Update geometry and weight rank for a compact ViT on CIFAR-100, averaged over the 16 hidden matrices and, for the update spread, over all 624 iterations. Bold marks the better value; initialization stable rank is 44.29. Muon attains a markedly higher weight rank *despite* a 9.3× larger displacement. Accuracies are low because this is a small model on a 10k-image subset; not comparable to Table 1. Single seed; per-layer values and protocol in Appendix B.1.

Table 3: Final clean-test accuracy (%, epoch 20, mean ± std over 3 seeds) on MNIST with injected label noise (r=0.3, 0.4; Pair 0.5 omitted—at that rate all methods converge to ≈ 49%). SGD with momentum is the most robust optimizer. † Muon's learning rate was tuned on clean labels by test accuracy; robustness under Muon is strongly learning-rate dependent.

	Coherent (pair) noise		Symmetric (white) noise		
Optimizer	0.3	0.4	0.3	0.4	0.5
Adam	85.7 ± 1.8	70.7 ± 1.0	89.5 ± 1.0	87.2 ± 0.3	83.4 ± 1.4
Muon†	73.1 ± 0.9	60.7 ± 1.8	77.0 ± 0.1	68.1 ± 0.6	56.9 ± 1.6
low-rank SUMO (k=16)	86.6 ± 3.5	71.8 ± 2.8	95.0 ± 0.3	93.7 ± 0.5	92.3 ± 0.1
SGD	$\mathbf{93.8 \pm 1.8}$	$\mathbf{85.2 \pm 1.9}$	$\mathbf{96.4 \pm 0.2}$	$\mathbf{95.9 \pm 0.0}$	$\mathbf{95.3 \pm 0.2}$

Table 4: Memorization (final noisy-train accuracy, %, noise rate 0.3, mean over 3 seeds). Muon fits corrupted labels almost perfectly at this learning rate. Muon's clean-test accuracy then falls from its peak: coherent 91.7% → 73.1%, symmetric 96.8% → 77.0%; Adam falls by 10.8 and 7.7 points, SUMO by 8.1 and 1.6.

Optimizer	Coherent 0.3	Symmetric 0.3
Adam	75.6	∼75
Muon	**98.7**	**98.4**
SUMO (k=16)	69.8	68.8

6 Conclusion

Muon's spectral equalization enforces a unit condition number on every update, and this carries structural costs we have made precise in this work. First, Muon produces maximally spread update spectra at fixed Frobenius norm and active rank, and in our compact-ViT experiment it also yields higher final stable rank; when coherent signal in the data happens to align with a low-rank mode (as with asymmetric label noise), the optimizer that preserves this structure (SGD) leads Muon from r=0.1 onward. Second, erasing the singular-value hierarchy inflates the noise-to-signal energy share in each update; in our MNIST runs Muon fit corrupted labels almost perfectly at its selected learning rate. Third, the rate separation that drives the crossover from optimization advantage to generalization deficit is itself a function of the spectral gap, suggesting that data geometry, not just architecture or learning rate, determines when equalization helps or hurts. Taken together, these results argue that future optimizers should decouple equalization (the condition-number benefit) from spectrum erasure (the noise cost), and we hope this analysis provides a principled starting point for that design.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, pp. 233–242. PMLR, 2017.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024a.

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, 2023.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

Sara Dragutinović, Yedi Zhang, and Rajesh Ranganath. To use or not to use muon: How simplicity bias in optimizers matters, 2026.

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

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Dongseong Hwang. Fadam: Adam is a natural gradient optimizer using diagonal empirical fisher information. *arXiv preprint arXiv:2405.12807*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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. Blog post.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023.

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

Jonathan Li et al. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4313–4324. PMLR, 2020.

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

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, et al. Muon is scalable for llm training, 2025.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944–1952, 2017.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610. IEEE, 2007.

Naoki Sato et al. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Wei Shen et al. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

Egor Shulgin et al. Beyond the ideal: Analyzing the inexact muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.

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

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

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory (COLT)*, pp. 3635–3673. PMLR, 2020.

Zhirong Yang and Jorma Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 21(2-3):232–240, 2008.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020.

A Proofs

A.1 Proof of Theorem 3.1

Proof. Both updates are normalized to Frobenius norm ρ, so $\|\tilde{G}_{\mathrm{grad}}\|_F = \|\tilde{G}_{\mathrm{eq}}\|_F = \rho$.

Since $UV^\top$ has r nonzero singular values, all equal to 1,

$$\|UV^\top\|_F = \sqrt{r}, \qquad \|UV^\top\|_* = r,$$

and therefore

$$\|\tilde{G}_{\mathrm{eq}}\|_* = \rho\,\frac{\|UV^\top\|_*}{\|UV^\top\|_F} = \rho\,\frac{r}{\sqrt{r}} = \rho\sqrt{r}.$$

On the other hand, writing the nonzero singular values of G as $\sigma_1, \dots, \sigma_r$,

$$\|\tilde{G}_{\mathrm{grad}}\|_* = \rho\,\frac{\|G\|_*}{\|G\|_F} = \rho\,\frac{\sum_{i=1}^r \sigma_i}{\left(\sum_{i=1}^r \sigma_i^2\right)^{1/2}}.$$

By the Cauchy–Schwarz inequality, $\sum_{i=1}^r \sigma_i \le \sqrt{r}\left(\sum_{i=1}^r \sigma_i^2\right)^{1/2}$, so that

$$\|\tilde{G}_{\mathrm{grad}}\|_* \le \rho\sqrt{r} = \|\tilde{G}_{\mathrm{eq}}\|_*.$$

Equality in Cauchy–Schwarz holds if and only if $\sigma_1 = \cdots = \sigma_r$; hence the inequality is strict whenever the nonzero singular values of G are not all equal. □

A.2 Proof of Theorem 3.2

Proof. For an update Δ with signal and noise components Δ^s and Δ^n, define the noise-to-signal update-energy ratio $R(\Delta) = \|\Delta^n\|_F^2 / \|\Delta^s\|_F^2$.

SGD. Since $\Delta_{\mathrm{SGD}} = -\eta G_k = -\eta G_k^s - \eta G_k^n$, its components are $\Delta^s_{\mathrm{SGD}} = -\eta G_k^s$ and $\Delta^n_{\mathrm{SGD}} = -\eta G_k^n$. Using the singular decompositions of G_k^s and G_k^n,

$$\|\Delta^s_{\mathrm{SGD}}\|_F^2 = \eta^2 \sum_{i=1}^{r_s} (\sigma_i^s)^2, \qquad \|\Delta^n_{\mathrm{SGD}}\|_F^2 = \eta^2 \sum_{j=1}^{r_n} (\sigma_j^n)^2,$$

hence

$$R(\Delta_{\mathrm{SGD}}) = \frac{\sum_{j=1}^{r_n} (\sigma_j^n)^2}{\sum_{i=1}^{r_s} (\sigma_i^s)^2}.$$

By the spectral-gap assumption $\alpha\,\sigma^n_{\max} \le \sigma^s_{\min}$, for all i, j we have $\sigma^n_j \le \sigma^n_{\max} \le \frac{1}{\alpha}\sigma^s_{\min} \le \frac{1}{\alpha}\sigma^s_i$. Summing,

$$\sum_{j=1}^{r_n}(\sigma^n_j)^2 \le r_n(\sigma^n_{\max})^2 \le \frac{r_n}{\alpha^2}(\sigma^s_{\min})^2 \le \frac{r_n}{\alpha^2}\frac{1}{r_s}\sum_{i=1}^{r_s}(\sigma^s_i)^2,$$

so that

$$R(\Delta_{\mathrm{SGD}}) \le \frac{1}{\alpha^2}\frac{r_n}{r_s}.$$

Muon. By definition $\Delta^s_{\mathrm{Muon}} = -\eta\sum_{i=1}^{r_s} u_i v_i^\top$ and $\Delta^n_{\mathrm{Muon}} = -\eta\sum_{j=1}^{r_n} a_j b_j^\top$. Because the signal and noise singular directions are mutually orthogonal, the rank-one terms $u_i v_i^\top$ and $a_j b_j^\top$ are orthonormal in the Frobenius inner product, giving

$$\|\Delta^s_{\mathrm{Muon}}\|_F^2 = \eta^2 r_s, \qquad \|\Delta^n_{\mathrm{Muon}}\|_F^2 = \eta^2 r_n,$$
$$R(\Delta_{\mathrm{Muon}}) = \frac{r_n}{r_s}.$$

Combining the two displays,

$$\begin{aligned}\frac{\|\Delta^n_{\mathrm{Muon}}\|_F^2}{\|\Delta^s_{\mathrm{Muon}}\|_F^2} &= R(\Delta_{\mathrm{Muon}})\\ &\ge \alpha^2 R(\Delta_{\mathrm{SGD}})\\ &= \alpha^2\frac{\|\Delta^n_{\mathrm{SGD}}\|_F^2}{\|\Delta^s_{\mathrm{SGD}}\|_F^2}.\end{aligned}$$

□

A.3 Proof of Lemma 3.3

Proof. For the amplitude-sensitive flow $\dot{e}_i(t) = -\rho(\sigma_i)e_i(t)$, solving the linear ODE gives $e_i(t) = e_i(0)\exp(-\rho(\sigma_i)t)$. Since $|e_i(0)| = e_0$, the stopping condition $|e_i(t)| \le \varepsilon$ is equivalent to $e_0\exp(-\rho(\sigma_i)t) \le \varepsilon$, so

$$\tau^{\mathrm{amp}}_i = \frac{1}{\rho(\sigma_i)}\log\Big(\frac{e_0}{\varepsilon}\Big).$$

As ρ is strictly increasing and $\sigma_{\max} > \sigma_{\min}$, we have $\rho(\sigma_{\max}) > \rho(\sigma_{\min})$, and therefore $\tau^{\mathrm{amp}}_{\max} < \tau^{\mathrm{amp}}_{\min}$.

For the equalized flow $\dot{e}_i(t) = -c\,e_i(t)$, we get $e_i(t) = e_i(0)\exp(-ct)$ and hence $\tau^{\mathrm{eq}}_i = \frac{1}{c}\log(e_0/\varepsilon)$, which is independent of i. Thus $\tau^{\mathrm{eq}}_{\min} = \tau^{\mathrm{eq}}_{\max}$. □

A.4 Proof of Theorem 3.4

Proof. Since W^* is stationary for the clean loss, $\nabla\mathcal{L}^s(W^*) = 0$, and therefore

$$\nabla\mathcal{L}(W^*) = \nabla\mathcal{L}^s(W^*) + \nabla\mathcal{L}^n(W^*) = G^n(W^*).$$

The SGD update is $\Delta W_{\mathrm{SGD}} = -\eta_{\mathrm{SGD}}G^n(W^*)$, so $\|\Delta W_{\mathrm{SGD}}\|_F = \eta_{\mathrm{SGD}}\|G^n(W^*)\|_F$.

For the amplitude-blind update, $\Delta W_{\mathrm{AB}} = -\eta_{\mathrm{AB}}\mathrm{Ortho}(G^n(W^*))$. Since $\mathrm{rank}(G^n(W^*)) = l$, the matrix $\mathrm{Ortho}(G^n(W^*))$ has exactly l nonzero singular values, all equal to 1, so $\|\mathrm{Ortho}(G^n(W^*))\|_F = \sqrt{l}$ and thus $\|\Delta W_{\mathrm{AB}}\|_F = \eta_{\mathrm{AB}}\sqrt{l}$. Comparing the two norms, $\|\Delta W_{\mathrm{AB}}\|_F > \|\Delta W_{\mathrm{SGD}}\|_F$ if and only if $\eta_{\mathrm{AB}}\sqrt{l} > \eta_{\mathrm{SGD}}\|G^n(W^*)\|_F$. □

A.5 Proof of Proposition 4.2

Proof. Orthogonality of the singular directions of G^s and G^n is exactly the singular-direction assumption required by Theorem 3.2, and the spectral gap $\alpha\sigma_b \le \sigma_a$ is exactly the signal-to-noise separation condition in Theorem 3.2. Therefore Theorem 3.2 gives equation 2. The explicit ratios follow because SGD preserves the singular values σ_a, σ_b, while idealized Muon replaces both active singular values by equal values. The early-stopping claim follows directly from Lemma 3.3: amplitude-sensitive dynamics learn larger singular-value modes faster, while equalized dynamics learn all active modes at the same rate. □

A.6 Proof of Lemma 4.3

Proof of Lemma 4.3. For $i \in I_a$, we have $y_i = e_a$ and $x_i = \mu_a + z_i$, with $\mathbb{E}[z_i] = 0$. Hence

$$\mathbb{E}\left[\sum_{i \in I_a} y_i x_i^\top\right] = n_a e_a \mu_a^\top.$$

Similarly, for $i \in I_{a\to b}$, we have $y_i = e_b$ and $x_i = \nu_{a\to b} + z_i$, so

$$\mathbb{E}\left[\sum_{i \in I_{a\to b}} y_i x_i^\top\right] = n_b e_b \nu_{a\to b}^\top.$$

Adding the two identities gives the claimed expectation. □

A.7 Proof of Corollary 4.4

Proof of Corollary 4.4. From Lemma 4.3 it holds that $e_a^\top e_b = 0$ and $\mu_a^\top \nu_{a\to b} = 0$, so the left and right factors of the two rank-one matrices are mutually orthogonal. Therefore the normalized vectors

$$e_a, \quad e_b, \qquad \frac{\mu_a}{\|\mu_a\|_2}, \quad \frac{\nu_{a\to b}}{\|\nu_{a\to b}\|_2}$$

form valid left and right singular directions of $\mathbb{E}[M]$. The corresponding singular values are $n_a\|\mu_a\|_2$, and $n_b\|\nu_{a\to b}\|_2$. Thus the e_b-labeled subpopulation is a weak active singular mode since $0 < n_b\|\nu_{a\to b}\|_2 < n_a\|\mu_a\|_2$. □

A.8 Proof of Corollary 4.5

Proof of Corollary 4.5. By Lemma 4.3, the two observed components have orthogonal left and right singular directions. Thus, combining with Corollary 4.4, they satisfy the singular-direction assumption required by Theorem 3.2. The spectral gap assumption $\alpha\sigma_b \le \sigma_a$ is exactly the signal-to-noise separation condition in Theorem 3.2. Therefore Theorem 3.2 gives

$$\frac{\|\Delta^n_{\mathrm{Muon}}\|_F^2}{\|\Delta^s_{\mathrm{Muon}}\|_F^2} \ge \alpha^2 \frac{\|\Delta^n_{\mathrm{SGD}}\|_F^2}{\|\Delta^s_{\mathrm{SGD}}\|_F^2}.$$

The early-stopping claim follows directly from Lemma 3.3: amplitude-sensitive dynamics learn larger singular-value modes faster, while equalized dynamics learn all active modes at the same rate. □

A.9 Proof of Lemma 4.6

Proof of Lemma 4.6. For squared loss and a linear last layer, the gradient holds

$$\nabla_{W_L}\mathcal{L}(\Theta) = \sum_{i=1}^{n}(W_L\phi_\theta(x_i) - y_i)\phi_\theta(x_i)^\top.$$

The part depending directly on the observed labels is

$$-\sum_{i=1}^{n} y_i\phi_\theta(x_i)^\top.$$

Restricting this term to the dominant observed subpopulation I_a, where $y_i = e_a$, gives

$$-\sum_{i \in I_a} e_a\phi_\theta(x_i)^\top.$$

Taking expectations over $i \in I_a$, we obtain

$$\mathbb{E}\left[-\sum_{i \in I_a} e_a\phi_\theta(x_i)^\top\right] = -n_a e_a m_a(\theta)^\top.$$

Similarly, taking expectations over $i \in I_{a\to b}$, the corrupted observed subpopulation $I_{a\to b}$, where $y_i = e_b$, gives

$$\mathbb{E}\left[-\sum_{i\in I_{a\to b}} e_b \phi_\theta(x_i)^\top\right] = -n_b e_b m_b(\theta)^\top.$$

Combining the two contributions gives

$$H_L^{(a,b)} = -n_a e_a m_a(\theta)^\top - n_b e_b m_b(\theta)^\top.$$

Since $e_a^\top e_b = 0$ and, by assumption, $m_a(\theta)^\top m_b(\theta) = 0$, the two terms have orthogonal left and right factors. They therefore form singular components of $H_L^{(a,b)}$, with singular values $\sigma_a = n_a\|m_a(\theta)\|_2, \sigma_b = n_b\|m_b(\theta)\|_2$. □

A.10 Proof of Corollary 4.7

Proof of Corollary 4.7. By Lemma 4.6, the two components of $H_L^{(a,b)}$ have orthogonal left and right singular directions, with singular values σ_a and σ_b. By assumption, they remain singular components of the complete matrix G_L.

Together with $0 < \alpha\sigma_b \le \sigma_a$, this verifies the hypotheses of Proposition 4.2, so its relative-amplification conclusion applies at the last layer. Under the additional residual-flow assumptions, the hitting-time conclusion follows from Lemma 3.3. □

A.11 Proof of Lemma 4.8

Proof of Lemma 4.8. Let

$$h_{\ell-1,i} := A_{\ell-1}x_i.$$

At a fixed parameter value Θ, the gradient with respect to W_ℓ is

$$\nabla_{W_\ell}\mathcal{L}(\Theta) = \sum_{i=1}^n B_{\ell+1}^\top\big(f_\Theta(x_i) - y_i\big)h_{\ell-1,i}^\top.$$

Its label-dependent part is

$$-\sum_{i=1}^n B_{\ell+1}^\top y_i h_{\ell-1,i}^\top.$$

For $i \in I_a$, we have $y_i = e_a$ and $x_i = \mu_a + z_i$. Using $\mathbb{E}[z_i] = 0$,

$$\mathbb{E}\left[-\sum_{i\in I_a} B_{\ell+1}^\top e_a (A_{\ell-1}x_i)^\top\right] = -n_a B_{\ell+1}^\top e_a (A_{\ell-1}\mu_a)^\top.$$

Similarly, for $i \in I_{a\to b}$, we have $y_i = e_b$ and $x_i = \nu_{a\to b} + z_i$, and hence

$$\mathbb{E}\left[-\sum_{i\in I_{a\to b}} B_{\ell+1}^\top e_b (A_{\ell-1}x_i)^\top\right] = -n_b B_{\ell+1}^\top e_b (A_{\ell-1}\nu_{a\to b})^\top.$$

Define

$$p_{a,\ell} := B_{\ell+1}^\top e_a, \qquad p_{b,\ell} := B_{\ell+1}^\top e_b,$$

and

$$r_{a,\ell} := A_{\ell-1}\mu_a, \qquad r_{b,\ell} := A_{\ell-1}\nu_{a\to b}.$$

The transformed-orthogonality assumptions give

$$p_{a,\ell}^\top p_{b,\ell} = e_a^\top B_{\ell+1}B_{\ell+1}^\top e_b = 0,$$
$$r_{a,\ell}^\top r_{b,\ell} = \mu_a^\top A_{\ell-1}^\top A_{\ell-1}\nu_{a\to b} = 0.$$

Let

$$u_{a,\ell} := -\frac{p_{a,\ell}}{\|p_{a,\ell}\|_2}, \qquad u_{b,\ell} := -\frac{p_{b,\ell}}{\|p_{b,\ell}\|_2},$$

and

$$v_{a,\ell} := \frac{r_{a,\ell}}{\|r_{a,\ell}\|_2}, \qquad v_{b,\ell} := \frac{r_{b,\ell}}{\|r_{b,\ell}\|_2}.$$

Then $\{u_{a,\ell}, u_{b,\ell}\}$ and $\{v_{a,\ell}, v_{b,\ell}\}$ are orthonormal sets, and

$$H_\ell^{(a,b)} = \sigma_{a,\ell} u_{a,\ell} v_{a,\ell}^\top + \sigma_{b,\ell} u_{b,\ell} v_{b,\ell}^\top,$$

where

$$\sigma_{a,\ell} = n_a \|B_{\ell+1}^\top e_a\|_2 \|A_{\ell-1}\mu_a\|_2,$$
$$\sigma_{b,\ell} = n_b \|B_{\ell+1}^\top e_b\|_2 \|A_{\ell-1}\nu_{a\to b}\|_2.$$

Thus the two terms form a valid singular-value decomposition of $H_\ell^{(a,b)}$. □

A.12 Proof of Corollary 4.9

Proof of Corollary 4.9. By Lemma 4.8, the two components of $H_\ell^{(a,b)}$ have orthogonal left and right singular directions, with singular values $\sigma_{a,\ell}$ and $\sigma_{b,\ell}$. By assumption, they remain singular components of the complete matrix G_ℓ to which idealized Muon equalization is applied.

Together with $0 < \alpha\sigma_{b,\ell} \le \sigma_{a,\ell}$, this verifies the hypotheses of Proposition 4.2 at every layer. Its relative-amplification conclusion therefore applies layerwise.

Under the additional residual-flow assumptions, the hitting-time conclusion follows directly from Lemma 3.3. □

A.13 Proof of Corollary 4.11

Proof. Fix a layer ℓ. Because σ is positively 1-homogeneous, the ReLU identity $\sigma(a) = D(a)a$ makes the forward map exactly linear once the gates are fixed, and $f_\Theta(x) = B_{\ell+1}(x)D_\ell(x)W_\ell A_{\ell-1}(x)x$. The gradient computation of Lemma 4.8 therefore applies with $A_{\ell-1}$ and $B_{\ell+1}$ replaced by the gated products $A_{\ell-1}(x)$ and $B_{\ell+1}(x)$, evaluated on each subpopulation. Taking expectations separately over $i \in I_a$, where $y_i = e_a$ and $x_i = \mu_a + z_i$, and over $i \in I_{a\to b}$, where $y_i = e_b$ and $x_i = \nu_{a\to b} + z_i$, gives the two displayed label-dependent components of G_ℓ.

Unlike the deep-linear case, the gate matrices differ between the two subpopulations, so $A_{\ell-1}^{(a)} \neq A_{\ell-1}^{(b)}$ and $B_{\ell+1}^{(a)} \neq B_{\ell+1}^{(b)}$ in general, and the orthogonality of the transformed left and right factors does not follow from $e_a^\top e_b = 0$ and $\mu_a^\top u_{a\to b} = 0$ alone. Under the layerwise orthogonality hypothesis, the two components are distinct orthogonal singular components, with singular values

$$\sigma_{a,\ell} = n_a \|D_\ell^{(a)} B_{\ell+1}^{(a)\top} e_a\|_2 \|A_{\ell-1}^{(a)}\mu_a\|_2, \qquad \sigma_{b,\ell} = n_b \|D_\ell^{(b)} B_{\ell+1}^{(b)\top} e_b\|_2 \|A_{\ell-1}^{(b)} u_{a\to b}\|_2,$$

since each is a rank-one matrix whose spectral norm is the product of the norms of its factors. Together with the spectral gap $\alpha\sigma_{b,\ell} \le \sigma_{a,\ell}$, the hypotheses of Proposition 4.2 hold at layer ℓ, which gives the conclusion. □

A.14 Auxiliary Claims

Corollary A.1 (No convergence at a fixed step size). *Fix $\eta > 0$ and consider the equalized iteration*

$$W_{t+1} \;=\; W_t \;-\; \eta\,\mathrm{Ortho}(\nabla\mathcal{L}(W_t))\,, \qquad t = 0, 1, 2, \ldots$$

Suppose $\nabla\mathcal{L}(W_t) \neq 0$ for every t, and write $l_t := \mathrm{rank}\big(\nabla\mathcal{L}(W_t)\big) \ge 1$. Then

$$\|W_{t+1} - W_t\|_F \;=\; \eta\sqrt{l_t} \;\ge\; \eta \;>\; 0 \qquad \textit{for every } t, \tag{3}$$

so $(W_t)_{t\ge0}$ is not a Cauchy sequence and therefore does not converge. Moreover, for every $\bar{W}$ and every t,

$$\max\left\{\left\|W_t - \bar{W}\right\|_F,\ \left\|W_{t+1} - \bar{W}\right\|_F\right\} \;\ge\; \tfrac{\eta}{2}\sqrt{l_t} \;\ge\; \tfrac{\eta}{2}, \tag{4}$$

so of any two consecutive iterates, at least one lies at distance at least $\eta/2$ from any fixed candidate limit: the equalized iteration has a residual floor of order $\Omega(\eta)$, independent of the magnitude of the residual gradient.

Proof. Equation equation 3 is the computation of theorem 3.4 applied at W_t: $\mathrm{Ortho}(\nabla\mathcal{L}(W_t))$ has exactly l_t nonzero singular values, all equal to 1, hence Frobenius norm $\sqrt{l_t}$, and $l_t \geq 1$ because $\nabla\mathcal{L}(W_t) \neq 0$. A convergent sequence is Cauchy, so $\|W_{t+1} - W_t\|_F \to 0$; this contradicts the uniform lower bound η, so (W_t) cannot converge. For equation 4, the triangle inequality gives $\eta\sqrt{l_t} = \|W_{t+1} - W_t\|_F \leq \|W_{t+1} - \bar{W}\|_F + \|W_t - \bar{W}\|_F \leq 2\max\{\|W_t - \bar{W}\|_F, \|W_{t+1} - \bar{W}\|_F\}$. □

Remark A.2 (Scope and a testable prediction). theorem A.1 concerns a fixed step size. With $\eta_t \to 0$ the bound equation 3 degrades to $\eta_t\sqrt{l_t} \to 0$ and convergence is no longer excluded; what the corollary predicts is a residual floor proportional to the *current* step size rather than to the residual gradient. The plateau of fig. 3 is consistent with this floor; the corollary predicts that its height scales linearly in η, which we have not tested. The same mechanism is familiar for sign- and normalized-gradient methods (Bernstein et al., 2018); the content here is the $\sqrt{l}$ rank scaling specific to spectral equalization.

A.15 Additional related works discussion

Implicit bias and low-rank learning. Gradient-based methods on overparameterized models select among many zero-loss solutions, and this implicit bias often favors low-rank or low-complexity structure. Matrix factorization and deep linear models exhibit an implicit bias toward low nuclear norm and low rank (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2018; Razin & Cohen, 2020). Our Theorem 3.1 shows that spectral equalization sits at the opposite extreme: among updates of fixed energy and rank it *maximizes* the nuclear norm, spreading energy as uniformly as possible and thereby opposing precisely this low-rank preference.

Theory of Muon. Beyond its empirical success, Muon has attracted a fast-growing theoretical literature. Several works establish convergence guarantees and read Muon as a steepest-descent or non-Euclidean trust-region method under the spectral norm (Bernstein & Newhouse, 2024a; Li & Hong, 2025; Kovalev, 2025; Sato et al., 2025), analyze the inexact Newton–Schulz update actually used in practice (Shulgin et al., 2025), and connect Muon's implicit regularization to a constraint on the spectral norm of the weights (Li et al., 2025). Closest to us, Shen et al. (2025) show that Muon can exploit *low-rank* Hessian structure, and Fan et al. (2025) characterize the *implicit bias* of Muon and spectral descent, proving convergence to a max-margin solution with respect to the spectral norm on separable data. These results concern the direction of the converged separator; we instead study the geometry and dynamics of the update itself, how spectral equalization reallocates energy between signal and noise modes and collapses their rate separation, and show that the same mechanism becomes actively harmful once the data contain coherent label noise. Our findings are complementary to Su (2025), who argue that pure orthogonalization is directionally correct but not strictly optimal, and to Dragutinović et al. (2026), who observe that Muon weakens simplicity bias.

B Additional Experiments and Further Details

B.1 Rank experiment: measures, protocol, and additional results

Additional rank measures. Beyond the stable rank of Table 2, we record two further measures at the end of training, averaged over the 16 hidden matrices:

	Initialization	SGD	Muon
Effective (entropy) rank	115.49	100.59	109.02
Numerical rank ($\sigma_i > 10^{-2}\sigma_1$)	127.56	125.56	127.25

The numerical rank is saturated for all three configurations and does not distinguish the optimizers. This is precisely why we rely on continuous measures: equalization does not change *how many*

directions are active, only how energy is distributed across them, and only a measure sensitive to that distribution detects it.

Invariance to label corruption. Repeating the experiment under 20% asymmetric and 20% white label corruption leaves both the update geometry and the weight rank essentially unchanged (Table 5): the mean update spread varies by at most 0.07 and the mean stable rank by at most 0.7 across all three data conditions, for both optimizers. Spectral equalization therefore imposes its high-rank update geometry, and forfeits the low-rank preference of amplitude-sensitive training, irrespective of whether the training signal is clean or corrupted.

	Update spread $\|\Delta\|_*/\|\Delta\|_F$			Stable rank $\|W\|_F^2/\|W\|_2^2$		
Training data	Clean	Asym. 20%	White 20%	Clean	Asym. 20%	White 20%
SGD	2.64	2.57	2.59	7.61	7.71	7.91
Muon	10.56	10.53	10.50	31.87	32.57	32.40

Table 5: The update-geometry and weight-rank separation is invariant to label corruption. Mean update spread and mean final stable rank over the 16 hidden matrices. Single seed.

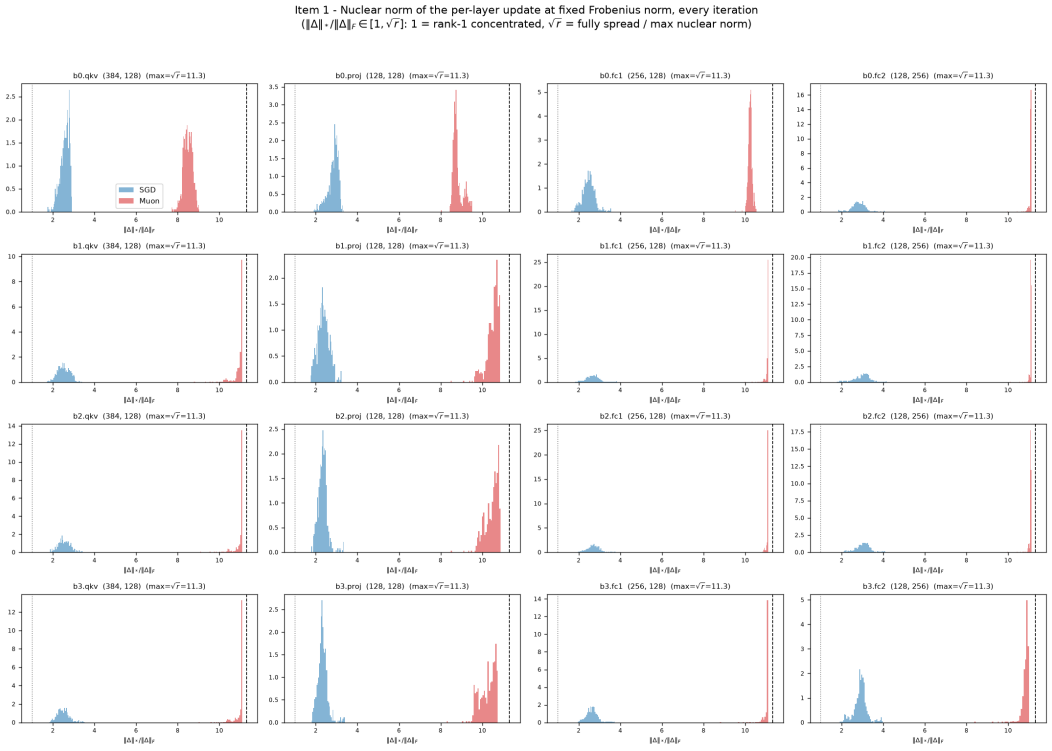

Figure 1: Per-layer distribution of the update's nuclear norm at fixed Frobenius norm, over all 624 training iterations of the compact ViT of section B.3. Each panel is one hidden weight matrix; the dashed line marks the upper bound $\sqrt{r}$ of theorem 3.1 with $r = \min(m, d)$ and the dotted line the rank-one lower bound 1. SGD's updates stay well below the bound, whereas Muon's approach it at every layer and throughout training.

B.2 Early-stopping experiment: construction and protocol

Construction. Section 4 establishes that coherent label noise produces a layer gradient with a dominant mode $\sigma_{\max}$ and a weak mode $\sigma_{\min}$ separated by a spectral gap. Here we impose that structure by construction, with $\alpha = \sigma_{\max}/\sigma_{\min} \in \{2, 4, 8\}$, and ask only whether the two modes

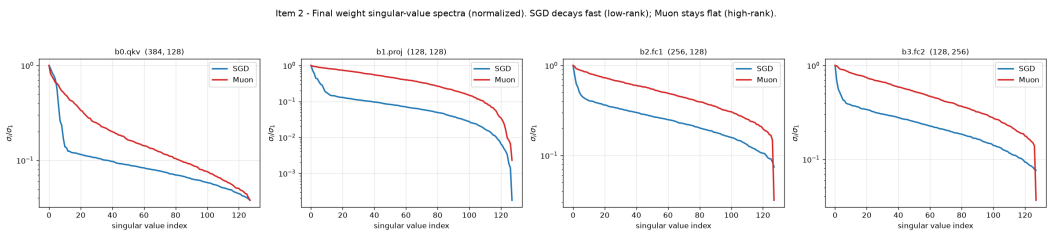

Figure 2: Final weight singular-value spectra, normalized by σ_1, for representative layers. SGD's spectra decay rapidly, indicating low-rank structure; Muon's remain comparatively flat.

are learned at separated rates. A network $f(x) = W_2\phi(W_1x)$ is trained with squared loss on the corresponding rank-two teacher, comparing SGD with *idealized* Muon, which replaces each layer gradient by $\mathrm{Ortho}(\cdot)$ via Newton–Schulz with ≥ 20 iterations, so that the update is orthogonal to numerical precision rather than the five iterations used in the rank experiment.

Reading off the modes. Modes are measured on the effective operator $\hat{A} = \mathbb{E}[f(x)x^\top]$, which reduces to W_2W_1 when $\phi = \mathrm{id}$, giving residuals $e_i(t) = (\sigma_i - u_i^\top \hat{A} v_i)/\sigma_i$ with $|e_{\max}(0)| \approx |e_{\min}(0)|$ as Lemma 3.3 assumes, and hitting times $\tau_i = \inf\{t : |e_i(t)| \leq \varepsilon\}$ at $\varepsilon = 0.3$. We report the exploitable window $W = \max(0, \tau_{\min} - \tau_{\max})$, the interval in which the dominant mode has been fit but the weak one has not, and the threshold-free separation $g = |e_{\min}| - |e_{\max}|$, which requires no choice of ε. Under SGD, $\tau_{\max}$ shrinks as α grows while $\tau_{\min}$ stays near 3.3k, which is what widens the window.

	SGD W	Muon W	SGD g	Muon g
Spectral gap sweep ($\phi = \mathrm{ReLU}$)				
$\alpha = 2$	1243 ± 399	1 ± 2	0.143 ± 0.058	-0.000 ± 0.030
$\alpha = 4$	2006 ± 316	0 ± 0	0.240 ± 0.044	-0.036 ± 0.028
$\alpha = 8$	2547 ± 289	0 ± 0	0.310 ± 0.038	-0.086 ± 0.028
Activation ablation ($\alpha = 4$)				
$\phi = \mathrm{id}$	419 ± 129	0 ± 0	0.083 ± 0.033	-0.012 ± 0.022
$\phi = \mathrm{ReLU}$	2027 ± 404	0 ± 0	0.233 ± 0.052	-0.019 ± 0.023
$\phi = \mathrm{GELU}$	2001 ± 400	0 ± 0	0.231 ± 0.052	-0.021 ± 0.025
$\phi = \tanh$	877 ± 206	0 ± 0	0.134 ± 0.043	-0.011 ± 0.023

Table 6: Exploitable early-stopping window W (steps) and separation g, mean $\pm$ standard deviation. SGD opens a window that widens with the spectral gap α; idealized Muon closes it for every gap and every activation, including GELU and $\tanh$, which fall outside the positively homogeneous setting of §4. The nonlinear activations widen the SGD window relative to the $\phi = \mathrm{id}$ baseline.

Learning dynamics, rate separation, and early stopping. A complementary line explains generalization through the *order* in which structure is learned: gradient descent tends to fit robust, simple, or high-magnitude components before weak or complex ones (Kalimeris et al., 2019; Saxe et al., 2014; Gidel et al., 2019), opening a window that early stopping exploits. In the infinite-width limit this connects to the distinction between the lazy/kernel regime and rich feature learning (Jacot et al., 2018; Chizat et al., 2019; Woodworth et al., 2020). Our Lemma 3.3 shows that spectral equalization forces strong and weak modes to a common learning rate, closing this early-stopping window, and our appendix analysis shows that the induced $O(\sqrt{r})$ parameter displacement pushes training out of the lazy regime.

Equalization closes the early-stopping window. Finally we test Lemma 3.3 in isolation, imposing the two-mode structure of Proposition 4.2 by construction with gap $\alpha = \sigma_{\max}/\sigma_{\min} \in \{2, 4, 8\}$ and training $f(x) = W_2\phi(W_1x)$ against the corresponding rank-two teacher. We measure the exploitable window $W = \max(0, \tau_{\min} - \tau_{\max})$ between the two hitting times. Under SGD the dominant mode reaches tolerance first and the window *widens* with the gap, $W = 1243 \to 2547$ for

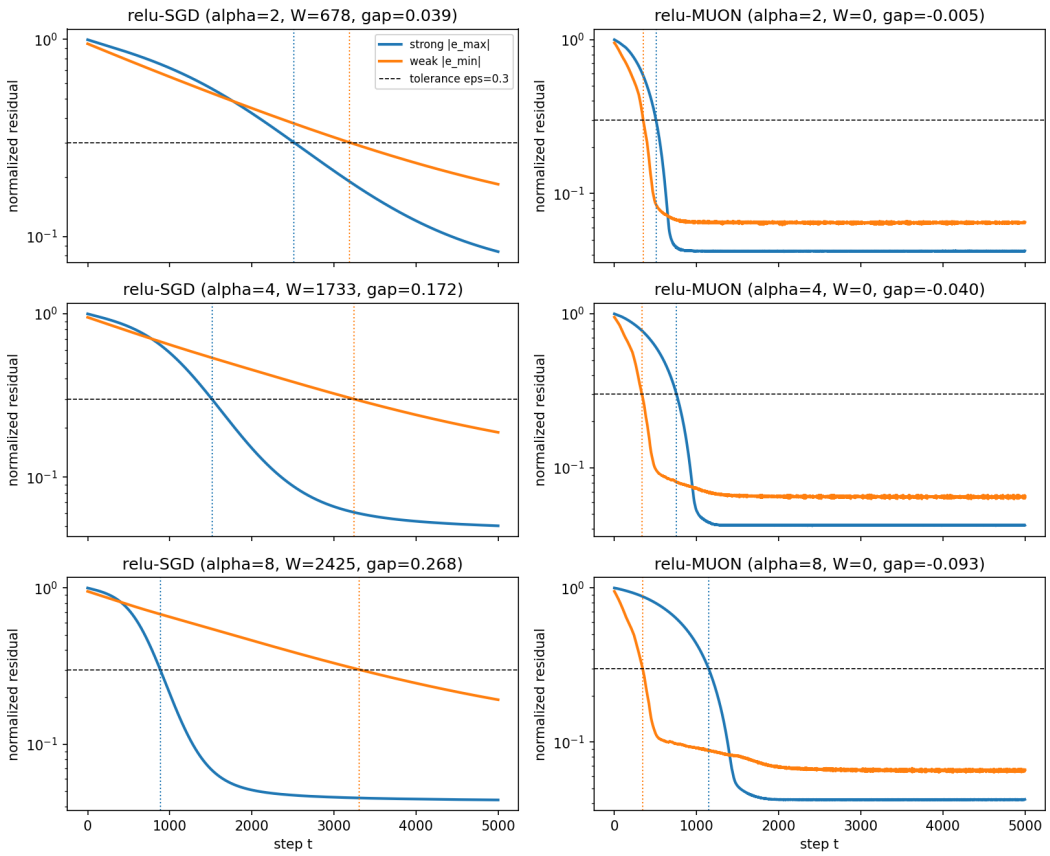

Figure 3: Residuals $|e_{\max}|$ (dominant mode, solid) and $|e_{\min}|$ (weak mode, dashed) over training (log scale), $\phi = \mathrm{ReLU}$, $\alpha \in \{2, 4, 8\}$. Dotted line marks ε. **SGD (left):** the dominant mode crosses ε before the weak mode, and the interval between hitting times widens with α. **Muon (right):** the two modes cross ε together, closing the window, and then plateau at a nonzero floor (Theorem 3.4). Panel titles report a single representative run, so the W and gap values shown differ from the seed-averaged values in Table 6.

$\alpha = 2 \to 8$: this is the $\tau^{\mathrm{amp}}_{\max} < \tau^{\mathrm{amp}}_{\min}$ ordering of Lemma 3.3, and the interval early stopping exploits. Idealized Muon imposes a common rate, giving $W \approx 0$ at every gap, and the separation is in fact mildly *negative* at large gaps: equalization grants both modes equal absolute progress per step while residuals are normalized by σ_i, so the weak mode reaches at least parity with the dominant one. The collapse is not an artifact of linearity, persisting for GELU and $\tanh$, which lie outside the homogeneous setting of §4 (Table 6). Muon's residuals also plateau at a nonzero floor rather than converging (Figure 3), as Theorem 3.4 predicts: the equalized step has magnitude independent of the residual singular values, so it cannot settle.

B.3 Rank experiment: measures and protocol

Rank measures. For a matrix W with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ we report the *stable rank* $\|W\|_F^2/\|W\|_2^2 = \sum_i \sigma_i^2/\sigma_1^2$ and the *effective (entropy) rank* $\exp\big(-\sum_i p_i \log p_i\big)$ with $p_i = \sigma_i/\sum_j \sigma_j$ (Roy & Vetterli, 2007). Both lie in $[1, \mathrm{rank}(W)]$, equal 1 for a rank-one matrix, and equal k when k singular values are equal and the remainder vanish. Unlike the numerical rank, which counts singular values above a tolerance and is near-maximal for all configurations we consider, these measures are continuous and sensitive to how energy is *distributed* across singular directions, which is precisely what spectral equalization alters. The update statistic $\|\Delta\|_*/\|\Delta\|_F$ is closely related: its square $(\sum_i \sigma_i)^2/\sum_i \sigma_i^2$ is the participation ratio of the update spectrum, so it is itself a soft rank, expressed in the units in which Theorem 3.1 is stated.

Layer	Shape	SGD	Muon
b0.qkv	384×128	5.13	9.20
b0.proj	128×128	7.11	21.64
b0.fc1	256×128	6.28	19.73
b0.fc2	128×256	5.13	37.27
b1.qkv	384×128	6.35	31.50
b1.proj	128×128	3.38	30.94
b1.fc1	256×128	9.69	40.00
b1.fc2	128×256	11.52	34.78
b2.qkv	384×128	8.21	34.90
b2.proj	128×128	4.03	30.46
b2.fc1	256×128	10.89	37.04
b2.fc2	128×256	12.68	36.97
b3.qkv	384×128	7.75	40.17
b3.proj	128×128	5.32	30.28
b3.fc1	256×128	8.90	38.83
b3.fc2	128×256	9.33	36.28
Mean		**7.61**	**31.87**

Table 7: Per-layer stable rank $\|W\|_F^2/\|W\|_2^2$ at the end of training, where $\mathrm{b}\ell$ denotes transformer block ℓ. Initialization stable rank averages 44.29 (range 32–54). On average SGD retains 17% of the initial stable rank and Muon 72%.

Architecture and data. We train a compact Vision Transformer on CIFAR-100: patch size 4 on 32×32 inputs, embedding dimension 128, depth 4, 4 attention heads, and MLP ratio 2. This yields 16 two-dimensional hidden weight matrices, namely the attention $W_{qkv} \in \mathbb{R}^{384\times128}$ and $W_{\mathrm{proj}} \in \mathbb{R}^{128\times128}$ and the MLP $W_{fc_1} \in \mathbb{R}^{256\times128}$ and $W_{fc_2} \in \mathbb{R}^{128\times256}$ in each of the four blocks. All linear layers are initialized with a truncated normal of standard deviation 0.02 and zero biases. We minimize the cross-entropy loss on a fixed random subset of $10{,}000$ training images for 8 epochs at batch size 128 (624 iterations), without data augmentation, and evaluate on a fixed random subset of $2{,}000$ test images. Labels are uncorrupted.

Optimizers. Both runs use heavy-ball momentum with coefficient 0.9, and neither uses weight decay or a learning-rate schedule. Muon applies five Newton–Schulz iterations to the momentum matrix of each of the 16 hidden matrices, with the standard $\sqrt{\max(1, d_{\mathrm{out}}/d_{\mathrm{in}})}$ scaling, at learning rate 0.03; SGD applies the momentum matrix directly at learning rate 0.2. All remaining parameters, namely layer normalization, biases, patch embedding, class token, positional embedding, and classification head, are updated by SGD with momentum at learning rate 0.1 in *both* runs, so that the two configurations differ only in the update applied to the hidden weight matrices. Both runs share the same seed and therefore have identical initialization and data order.

Measurements. At every iteration and for each hidden matrix we compute the singular values of the update direction actually applied, namely the momentum matrix for SGD and its Newton–Schulz orthogonalization for Muon, and record its nuclear, Frobenius, and spectral norms. At the end of training we compute the full singular value decomposition of every hidden weight matrix. Initialization statistics are computed from the same seeded initialization before any update is applied. All results are from a single seed.

B.4 Noise robustness experiment: protocol and tuning

Dataset and architecture. We train a ViT-B/16 on CIFAR-100 ($50{,}000$ training images, $10{,}000$ test images, 100 classes). Images are resized to 224×224 and normalised with ImageNet statistics. We use a standard pre-trained ViT-B/16 checkpoint and fine-tune all parameters.

Label-noise injection. We evaluate two noise types at rates $r \in \{0, 0.1, 0.2, 0.3, 0.4\}$:

- **Asymmetric noise:** each label is flipped to a fixed neighbouring class with probability r, following the class-transition matrix of Patrini et al. (2017) adapted to CIFAR-100.

- **White (uniform) noise:** each label is replaced uniformly at random with probability r.

Noise is injected once before training and held fixed across seeds.

Optimiser hyper-parameters and tuning. We compare AdamW, Muon, and SUMO under identical training budgets: 50 epochs with a cosine learning-rate schedule and a 5-epoch linear warm-up. The peak learning rate and weight-decay coefficient are tuned independently for each optimiser via a coarse-to-fine grid search evaluated on a 10% held-out subset of the training data at zero noise (r=0); the selected values are then used for all noise rates. All other hyper-parameters (batch size 256, gradient-clip norm 1.0) are shared. Results in Table 1 are averaged over 3 random seeds; standard deviations are reported in the table.

B.5 MNIST MEMORIZATION EXPERIMENT: PROTOCOL

Data and label noise. We train on the 60,000 MNIST training images and evaluate on the 10,000 *clean* test images. Pixels are scaled to $[0, 1]$ with no further normalization and no augmentation. Independently for each training example, the observed label is corrupted with probability $r \in \{0.3, 0.4, 0.5\}$. Under *symmetric (white)* noise, a corrupted label is drawn uniformly from the nine other classes. Under *coherent (pair)* noise, it is replaced by a fixed target class through the map T:

$$0\mapsto 6, \quad 1\mapsto 7, \quad 2\mapsto 8, \quad 3\mapsto 9, \quad 4\mapsto 0, \quad 5\mapsto 3, \quad 6\mapsto 0, \quad 7\mapsto 1, \quad 8\mapsto 3, \quad 9\mapsto 4.$$

T has no fixed points and is not a permutation: classes 4 and 6 share the target 0, and classes 5 and 8 share the target 3. The corruption depends only on the seed. For a given seed, all optimizers therefore see identical noisy labels, and the reported standard deviations include variation in the noise draw.

Model and training. All optimizers train the same MLP, $784 \to 1024 \to 1024 \to 10$ with GELU activations and PyTorch's default initialization. It has three weight matrices, of shapes 1024×784, 1024×1024 and 10×1024. We minimize cross-entropy for 20 epochs at batch size 128 with reshuffling every epoch. The learning rate is constant, with no warm-up, schedule or early stopping. We report means $\pm$ standard deviations over three seeds $(0, 1, 2)$. Each seed fixes the initialization, the data order and the noise draw.

Optimizers. All methods use weight decay $\lambda = 10^{-4}$.

- **SGD:** applied to all parameters, with heavy-ball momentum 0.9, learning rate 10^{-2} and coupled (L_2) weight decay.
- **Adam:** applied to all parameters, with learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and coupled weight decay.
- **Muon ($k = \text{full}$):** applied to the three weight matrices, with Nesterov momentum 0.95. The Nesterov momentum is Frobenius-normalized and approximately orthogonalized by five quintic Newton–Schulz iterations with coefficients $(3.4445, -4.7750, 2.0315)$. The result is scaled by $\max(1, \sqrt{d_{\text{in}}})$, which is 28 for the first matrix and 32 for the other two. The learning rate is 3×10^{-4}, so the step per unit singular value is 8.4×10^{-3} for the first matrix and 9.6×10^{-3} for the other two. Weight decay is decoupled, $W \leftarrow (1 - \text{lr}\,\lambda)W$.
- **SUMO ($k = 16$):** applied to the three weight matrices. Let $G \in \mathbb{R}^{m\times n}$ be the gradient. An orthonormal basis Q of its top-k singular subspace is recomputed every 200 steps. It is taken on the right when $m \geq n$ and on the left otherwise, with $k = 16$. Nesterov momentum (0.95) is accumulated only on the projected gradient $\hat{G}$. It is orthogonalized by an exact SVD, $O = U_k V_k^\top$, and a norm-growth limiter caps $\|O\|_F$ at $\gamma = 1.1$ times its value at the previous step. The update is

$$W \;\leftarrow\; W - \alpha\,\text{lr}\left[(G - \mathcal{P}_Q\hat{G}) + 0.2\sqrt{\max(m,n)}\;\mathcal{P}_Q O\right],$$

  where $\mathcal{P}_Q$ maps back to the full space, and $\alpha = 4$, $\text{lr} = 10^{-3}$. The out-of-subspace part of the current gradient, without momentum, is applied *unmodified* and keeps its singular-value magnitudes. The step per unit singular value inside the subspace is 2.56×10^{-2} for all three matrices, about $3\times$ Muon's. Weight decay is decoupled, $W \leftarrow W - \text{lr}\,\lambda W$.

Under Muon and SUMO, bias vectors are updated by Adam at learning rate 10^{-3} (minor implementation differences in its weight-decay form and ϵ have negligible effect at $\lambda = 10^{-4}$).

SUMO is thus not a partial equalization of the whole spectrum. It *restricts equalization to the dominant k-dimensional subspace*, and modes outside the top k keep their gradient amplitudes. The last matrix has only 10 rows, so the rank-16 projection is the identity there, and SUMO orthogonalizes its full momentum just as Muon does. The spectral treatment of the two methods therefore differs only in the first two matrices. Muon is the $k = \text{full}$ endpoint of this family, up to the differences listed next. The two methods also differ in step size, in the limiter, in exact versus Newton–Schulz orthogonalization, and in SUMO's momentum-free out-of-subspace component.

Hyperparameter selection. No shared tuning protocol was used; each optimizer's settings come from a separate preliminary run.

- **Adam:** the default learning rate, untuned.
- **Muon:** the best learning rate in $\{3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$ on clean labels, from one seed, a width-512 MLP and 3 epochs. The selected value is the lower edge of that grid.
- **SUMO:** its configuration (width, k, lr) was the best of $\{(1024, 16, 10^{-3}), (2048, 32, 10^{-3}), (2048, 64, 2 \times 10^{-3})\}$ under symmetric noise at $r = 0.5$, from one seed and 5 epochs. This selection also fixed the width of 1024 used for all optimizers.
- **SGD:** the best learning rate in $\{0.003, 0.01, 0.03, 0.1, 0.3\}$ at $r = 0.3$ from seed 0, by mean accuracy over the two noise types. The optimum for symmetric noise alone was 0.03.

Every selection used clean test accuracy; no validation split was held out. SUMO and SGD were thus tuned under label noise, whereas Muon was tuned on clean data and at the edge of its grid, and Adam was not tuned. This asymmetry favours SUMO and SGD. It bears directly on the SUMO–Muon comparison below.

Measurements. Per epoch we record the training loss and the training accuracy on the observed (noisy) labels, both averaged over the epoch's minibatches. We also record the clean test loss and clean test accuracy. table 3 reports clean test accuracy after the final epoch. table 4 reports the final-epoch training accuracy a on the observed labels together with the peak clean test accuracy over the run. A model that fits every clean label and no corrupted one scores $a = 1 - r$ (70% at $r = 0.3$). Since at most the $1 - r$ clean fraction can be fit correctly, the fraction of corrupted labels memorized is at least $(a - (1 - r))/r$.

Results. At $r = 0.5$ under pair noise, each class's labels are split evenly between the true class and its target, so the observed labels no longer identify the class. All four optimizers land at 48–50%, within one standard deviation of each other, and we exclude this setting from the comparisons that follow. In the remaining five settings, SGD with momentum is the most robust optimizer in these settings. This ordering is consistent with theorem 3.2. Under that result, equalizing weak modes amplifies the noise they carry: SGD equalizes no modes, SUMO equalizes only the top k, and Muon equalizes all of them. We do not, however, measure which modes carry the corrupted signal. The SUMO–Muon gap also cannot be attributed to the spectral treatment alone. The two differ in step size (about $3\times$ inside the subspace), in the norm-growth limiter, in their momentum outside the subspace, and in their tuning protocol.