AUGMENTING MATRIX ORTHOGONALIZATION WITH SAMPLE-SPACE GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Orthogonalized optimizers such as Muon exploit the matrix structure of parameter updates to improve neural network training. Their normalization acts on the row and column directions of aggregated updates, leaving the co-update directions shared across individual training examples without explicit spectral control. We propose augmenting matrix-based orthogonalization with this complementary sample-space geometry. By organizing per-sample updates into a matrix, we identify shared update directions through a sample-space Gram matrix, enabling spectral shaping across vector, matrix, and heterogeneous parameter blocks. To incorporate these directions without disproportionately amplifying weak components, we introduce Selective Orthogonalization: a smooth transformation that normalizes strong singular directions while preserving the scale of vanishingly weak ones. We prove that, for a fixed positive transition scale, its relative amplification approaches one as a singular value vanishes, and derive an equivalent parameter-side spectral formulation. The resulting approach supports a hybrid optimizer that alternates Selective and Muon updates across iterations, bringing sample-space and parameter-matrix geometry into the same training process. Standalone Selective experiments establish the behavior of the sample-space component: controlled studies validate weak-direction scale preservation across five parameter structures, while image-classification experiments demonstrate substantially reduced weak-direction amplification and competitive accuracy across datasets, architectures, and batch sizes. Together, these results motivate per-sample update geometry as a complementary source of structure for orthogonalized optimization.

1 INTRODUCTION

Orthogonalized optimizers such as Muon exploit the matrix structure of parameter updates to improve neural network training (Jordan et al., 2024). Specifically, Muon applies Newton–Schulz iterations to matrix-shaped momentum updates to approximately orthogonalize them (Liu et al., 2025; Kim & Oh, 2026; Shulgin et al., 2026). For an update with singular value decomposition $P = U\Sigma V^\top$, this operation approximates the hard polar transform $UV^\top$, preserving the active singular directions while flattening the nonzero singular values (Higham, 2008; Golub & Van Loan, 2013). By removing scale disparities across the active spectrum, this transformation is closely connected to steepest descent under the spectral norm (Bernstein & Newhouse, 2024), and recent studies have demonstrated the effectiveness of such orthogonalized updates in large-scale training (Refael et al., 2025; Khaled et al., 2026; Amsel et al., 2026; Wu et al., 2026). Muon therefore uses spectral normalization to act on the row and column directions of matrix-shaped aggregated momentum updates.

However, aggregation does not preserve an explicit representation of how individual training examples jointly contribute to the update. Before aggregation, per-sample updates contain a sample-space geometry: their pairwise inner products describe how examples align and co-update the parameters (Jacot et al., 2018; Sankararaman et al., 2020). We therefore make this sample-space geometry explicit and use it as an additional spectral view of the update. By stacking these updates into a matrix S_t, this structure is captured by the sample Gram matrix $K_t = S_t S_t^\top$. Crucially, K_t and the corresponding parameter-side matrix $C_t = S_t^\top S_t$ share the same nonzero spectrum, while K_t is only $B \times B$ (Golub & Van Loan, 2013; Horn & Johnson, 2012). This makes the active spectrum

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.

directly accessible from sample space and enables explicit spectral shaping of the co-update directions. At the same time, the shared nonzero spectrum links the sample-space and parameter-side views of the update, making sample-space geometry complementary to the matrix geometry used by orthogonalized optimizers.

Meanwhile, whether orthogonalization is applied to matrix-shaped aggregated updates or to the per-sample update matrix, it reduces spectral imbalance by flattening active singular values. Hard polar normalization, however, removes magnitude distinctions across the active spectrum: every nonzero singular value is mapped to the same target, regardless of whether the corresponding direction is strong or vanishingly weak. For a singular value $\sigma > 0$, the resulting relative amplification is $1/\sigma$. Thus, as an active singular value approaches zero, its transformed value remains at the fixed hard target rather than decaying with the original update, and its relative amplification diverges. Prior work has shown that modifying the singular structure of gradient updates can materially affect optimization dynamics (Carlson et al., 2015a;b; Tuddenham et al., 2022; Wang et al., 2026; Chakravarthy et al., 2025). Our controlled and real-data experiments also confirm that weak singular directions arise in practice and that hard normalization increasingly amplifies them as their scale decreases.

These observations motivate a spectral rule that preserves the useful normalization of strong directions while allowing vanishingly weak directions to retain their original scale. We introduce Selective Orthogonalization, a smooth transformation of the per-sample update spectrum. Strong singular directions are treated similarly to hard orthogonalization, whereas weak directions are left progressively closer to their original scale. For a fixed positive transition scale, we show that the relative amplification of Selective approaches one as a singular value vanishes, in contrast to the divergent amplification of hard normalization. In practice, an adaptive transition scale determines where the transform moves between these two regimes. This provides a smooth alternative to hard sample-space orthogonalization while retaining its normalization effect on strong directions.

The resulting approach also supports a hybrid optimizer that alternates Selective and Muon updates across iterations. Since both transformations are tied to a shared momentum state, the sample-space and matrix-space updates can alternate within the same optimization trajectory. We instantiate this construction as Selective–Muon, bringing sample-space and parameter-matrix geometry into the same training process.

Our analysis formalizes this behavior through the weak-direction limit and an equivalent parameter-side spectral formulation. Controlled experiments across five parameter structures verify that Selective preserves the scale of weak directions while maintaining strong optimization performance, supporting its advantage over hard sample-space orthogonalization. Real-data experiments further show substantially reduced weak-direction amplification together with strong test accuracy with test-time augmentation (TTA) (Shanmugam et al., 2021) across datasets, architectures, and batch sizes. The Selective–Muon hybrid also achieves higher mean TTA test accuracy than the corresponding Muon baseline in all nine tested settings, supporting sample-space geometry as a complementary source of structure for matrix orthogonalization.

Our contributions are summarized as follows:

- **Complementary sample-space geometry.** We formulate per-sample co-update geometry as a complementary view of matrix-based orthogonalization and show that its active spectrum can be accessed through the sample Gram matrix.
- **Selective Orthogonalization.** We introduce a spectral transformation that normalizes strong directions while preserving the scale of vanishingly weak directions. For a fixed positive transition scale, its relative amplification approaches one as the singular value vanishes, and we derive an equivalent parameter-side spectral formulation.
- **Shared-momentum combination with Muon.** We introduce Selective–Muon, which alternates sample-space Selective updates and matrix-space Muon updates while maintaining a shared momentum state, bringing the two geometries into a single optimization trajectory.
- **Empirical validation.** Controlled and real-data experiments validate weak-direction scale preservation and strong optimization performance across parameter structures, datasets, architectures, and batch sizes. Selective–Muon further achieves higher mean TTA test accuracy than the corresponding Muon baseline in all nine tested settings.

2 SPECTRAL SETUP

2.1 MATRIX AND SAMPLE-SPACE VIEWS

Matrix orthogonalization and sample-space spectral shaping act on different views of the update. Muon operates on matrix-shaped aggregated momentum updates, using their row–column structure for spectral normalization. To expose the complementary geometry across individual training examples, we instead retain the update at sample resolution. Consider one parameter block $\theta_t \in \mathbb{R}^p$ at iteration t. Let $S_t \in \mathbb{R}^{B\times p}$ denote its per-sample update matrix, where B is the batch size and p is the flattened dimension of the parameter block. The b-th row of S_t contains the update contributed by the b-th sample. Only a flattened per-sample representation is required, so the construction applies equally to vector, matrix, and heterogeneous parameter blocks. From S_t, we define

$$K_t := S_t S_t^\top \in \mathbb{R}^{B\times B}, \qquad C_t := S_t^\top S_t \in \mathbb{R}^{p\times p}.$$

The matrix K_t is the sample Gram matrix, while C_t is the parameter-side unnormalized second-moment matrix. The entries of K_t encode pairwise inner products between per-sample updates and therefore describe how examples align and co-update the parameters. In contrast, C_t describes how the same minibatch update is distributed across directions in parameter space. For any direction $a \in \mathbb{R}^p$, $a^\top C_t a = \|S_t a\|_2^2$, so the quadratic form of C_t measures the total minibatch update energy along a. Let $S_t = U_t \Sigma_t V_t^\top$ be the compact singular value decomposition, where $r_t = \operatorname{rank}(S_t)$, $U_t \in \mathbb{R}^{B\times r_t}$ and $V_t \in \mathbb{R}^{p\times r_t}$ contain the active left and right singular vectors, and $\Sigma_t = \operatorname{diag}(\sigma_{t,1}, \ldots, \sigma_{t,r_t}) \in \mathbb{R}^{r_t\times r_t}$, with $\sigma_{t,1} \geq \cdots \geq \sigma_{t,r_t} > 0$. Then

$$K_t = U_t \Sigma_t^2 U_t^\top, \qquad C_t = V_t \Sigma_t^2 V_t^\top. \tag{1}$$

Thus, the left singular vectors describe active directions in sample space, while the right singular vectors describe the corresponding active directions in parameter space. Both views are governed by the same nonzero singular values of S_t (Golub & Van Loan, 2013; Horn & Johnson, 2012). This shared spectrum links sample-space co-update geometry to the corresponding parameter-side update geometry and provides the basis for the spectral transformation developed below.

2.2 HARD SAMPLE-SPACE ORTHOGONALIZATION

We use hard sample-space orthogonalization as the reference transform on the per-sample update matrix S_t. Given the compact SVD $S_t = U_t \Sigma_t V_t^\top$, hard sample-space orthogonalization replaces S_t by its polar factor

$$Q_t = U_t V_t^\top.$$

Equivalently, every nonzero singular value is mapped to the same target value of one. The transformation therefore preserves the active singular directions while removing magnitude differences across the active spectrum. On the parameter side, $C_t = V_t \Sigma_t^2 V_t^\top$ is transformed into $Q_t^\top Q_t = V_t V_t^\top$, so the active second-moment spectrum is completely flattened. This is the standard spectral effect of the polar factor (Higham, 2008). This flattening is useful when the goal is to reduce large disparities in scale across active directions. However, it treats strong and weak singular directions identically. A weak but nonzero singular value is mapped to the same target as a strong one, even though the two directions carry very different update scales before normalization. Accordingly, a spectral rule for sample-space shaping should retain hard-like normalization on strong directions while allowing sufficiently weak directions to remain closer to their original scale. Section 3 introduces a smooth transformation that realizes this behavior.

2.3 SAMPLE-SPACE ACCESS TO THE ACTIVE SPECTRUM

The shared spectrum of K_t and C_t also provides a direct computational route to the active singular values of S_t. Using the factorizations in Equation 1, let

$$\lambda_{t,i} := \lambda_i(K_t) = \lambda_i(C_t) = \sigma_{t,i}^2, \qquad i = 1, \ldots, r_t,$$

denote the shared active eigenvalues of K_t and C_t. Hence the active singular values of S_t can be recovered as $\sigma_{t,i} = \sqrt{\lambda_{t,i}}$ from the eigendecomposition of the $B \times B$ sample Gram matrix. When $B \ll p$, this avoids spectrally decomposing the $p \times p$ parameter-side matrix C_t. More importantly, the sample-space eigendecomposition exposes the spectral quantities associated with the co-update directions across examples, making them available for explicit spectral shaping. Section 3 uses these quantities to define the Selective transform and its adaptive transition scale.

3 SELECTIVE ORTHOGONALIZATION

Section 2 establishes the sample-space geometry on which our method operates. Hard sample-space orthogonalization provides the reference spectral transform, while the Gram matrix exposes the active singular spectrum required to modify it. We now define Selective Orthogonalization, which retains hard-like normalization on strong directions while preserving the scale of sufficiently weak ones.

(1) **Selective spectral map.** The main design requirement is to preserve the benefit of hard normalization on strong directions without forcing weak directions to the same unit scale. For a singular value $\sigma \geq 0$ and a positive transition scale ϵ, we define

$$g_\epsilon(\sigma) = \sigma^{\epsilon/(\sigma^2+\epsilon)}, \qquad g_\epsilon(0) = 0.$$

The behavior of this map is controlled by the ratio between σ^2 and ϵ. When $\sigma^2 \gg \epsilon$, the exponent approaches zero and $g_\epsilon(\sigma)$ moves toward the hard-orthogonalization target 1. When $\sigma^2 \ll \epsilon$, the exponent approaches one and $g_\epsilon(\sigma)$ remains close to the original value σ. Thus, the map smoothly connects hard-like normalization on strong directions with scale preservation on weak directions. Given the compact SVD $S_t = U_t \Sigma_t V_t^\top$, we apply g_{ϵ_t} to the active singular values while keeping the singular directions unchanged:

$$T_{\epsilon_t}(S_t) = U_t \operatorname{diag}\big(g_{\epsilon_t}(\sigma_{t,1}), \ldots, g_{\epsilon_t}(\sigma_{t,r_t})\big) V_t^\top.$$

The remaining task is to choose the transition scale ϵ_t from the current minibatch spectrum.

(2) **Adaptive transition scale.** A fixed transition scale would not adapt to changes in the magnitude and concentration of the update spectrum during training. We therefore define the spectrum-dependent level

$$L_t := \frac{\operatorname{tr}(K_t^2)}{\operatorname{tr}(K_t)} = \frac{\sum_i \lambda_{t,i}^2}{\sum_i \lambda_{t,i}},$$

where $\lambda_{t,i} = \sigma_{t,i}^2$ are the active eigenvalues of K_t. The quantity L_t is an eigenvalue-weighted level of the current spectrum, so directions carrying more update energy contribute more strongly to the transition scale. We then set

$$\epsilon_t = \tau L_t, \qquad \tau > 0.$$

The scalar τ controls the overall softness of the transform. Larger values of τ keep a larger portion of the spectrum closer to the original update, whereas smaller values move more active directions toward hard orthogonalization. The same spectrum-dependent level can be written as $L_t = \operatorname{tr}(K_t)/r_{\text{eff}}(K_t)$, where $r_{\text{eff}}(K_t) = \operatorname{tr}(K_t)^2/\operatorname{tr}(K_t^2)$. This form gives an equivalent effective-rank interpretation of the spectrum-dependent level L_t. In Section 5, we examine the sensitivity to τ and show that the method remains stable over a broad range of values.

(3) **Sample-space implementation.** The transform above is written in SVD form for clarity, but its implementation does not require the parameter-side singular vectors V_t. From Section 2, the eigendecomposition $K_t = U_t \Lambda_t U_t^\top, \Lambda_t = \operatorname{diag}(\lambda_{t,1}, \ldots, \lambda_{t,r_t})$, provides the left singular vectors and the active singular values $\sigma_{t,i} = \sqrt{\lambda_{t,i}}$. For each active direction, define the spectral multiplier $\gamma_{t,i} = g_{\epsilon_t}(\sigma_{t,i})/\sigma_{t,i}$, and let $\gamma_t = (\gamma_{t,1}, \ldots, \gamma_{t,r_t})^\top$. Then the selective transform can be written directly from the sample side as

$$T_{\epsilon_t}(S_t) = U_t \operatorname{diag}(\gamma_t) U_t^\top S_t.$$

Therefore, the required spectral transformation uses only the eigendecomposition of the $B \times B$ Gram matrix K_t. No $p \times p$ parameter-side decomposition is needed. The transformed minibatch update can also be aggregated without explicitly forming the full transformed matrix. Let $\mathbf{1} \in \mathbb{R}^B$ denote the all-ones vector and define the sample-side aggregation coefficients $\xi_t^\top = B^{-1}\mathbf{1}^\top U_t \operatorname{diag}(\gamma_t) U_t^\top$. The transformed minibatch aggregate is then

$$u_t = \frac{1}{B}\mathbf{1}^\top T_{\epsilon_t}(S_t) = \xi_t^\top S_t.$$

Once the sample-side coefficients ξ_t are available, the transformed parameter update is obtained by a single weighted aggregation of the per-sample updates. The corresponding identity is derived in Appendix C.

Algorithm 1 Sample-Space Selective Orthogonalization for One Parameter Block

Require: Per-sample update matrix $S_t \in \mathbb{R}^{B\times p}$, parameters θ_t, learning rate η_t, transition parameter $\tau > 0$, target RMS $\rho_0 > 0$, minimum RMS $\rho_{\min} > 0$

1: Form $K_t = S_t S_t^\top$ and compute its active eigendecomposition $K_t = U_t \Lambda_t U_t^\top$
2: Set $\sigma_{t,i} = \sqrt{\lambda_{t,i}}$ and $L_t = \operatorname{tr}(K_t^2)/\operatorname{tr}(K_t)$
3: Set $\epsilon_t = \tau L_t$ and $\gamma_{t,i} = g_{\epsilon_t}(\sigma_{t,i})/\sigma_{t,i}$
4: Compute $\xi_t^\top = B^{-1}\mathbf{1}^\top U_t \operatorname{diag}(\gamma_t) U_t^\top$ and $u_t = \xi_t^\top S_t$
5: Compute $\bar{s}_t = B^{-1}\mathbf{1}^\top S_t$ and $\omega_t = \epsilon_t/(\lambda_{\max}(K_t) + \epsilon_t)$
6: Set $\rho_t = (1-\omega_t)\rho_0 + \omega_t \max\{\rho_{\min}, \mathrm{RMS}(\bar{s}_t)\}$
7: Update $\theta_{t+1} = \theta_t - \eta_t \frac{\rho_t}{\mathrm{RMS}(u_t)} u_t$

(4) **Aggregation and update scaling.** The selective map determines the relative scaling across singular directions, while the optimizer still needs a global step magnitude. These two roles are handled separately. Let $\bar{s}_t = B^{-1}\mathbf{1}^\top S_t$ denote the ordinary minibatch aggregate. For any $v \in \mathbb{R}^p$, define $\mathrm{RMS}(v) := \frac{\|v\|_2}{\sqrt{p}}$. Structured matrix optimizers such as K-FAC, Shampoo, and Muon use non-coordinatewise transformations of gradient or momentum information (Martens & Grosse, 2015; Gupta et al., 2018; Jordan et al., 2024). Muon-style optimizers commonly rescale orthogonalized updates to match the magnitude conventions used by Adam or AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Standard stochastic first-order methods, in contrast, retain the scale carried by the gradient or momentum update (Polyak, 1964; Sutskever et al., 2013; Bottou et al., 2018). We therefore let the target RMS move between these two regimes using the same transition scale that controls the selective transform. Specifically, we define the transition weight ω_t and adaptive target RMS ρ_t as

$$\omega_t = \frac{\epsilon_t}{\lambda_{\max}(K_t) + \epsilon_t}, \qquad \rho_t = (1-\omega_t)\rho_0 + \omega_t \max\{\rho_{\min}, \mathrm{RMS}(\bar{s}_t)\}\,.$$

Here $\rho_0 > 0$ is the target RMS in the hard-like regime and $\rho_{\min} > 0$ prevents the reference scale from becoming exactly zero. When $\lambda_{\max}(K_t) \gg \epsilon_t$, the weight ω_t is small and the update remains close to the fixed target scale ρ_0. As the spectrum moves into the softer regime, ω_t increases and the target scale moves toward the RMS of the original minibatch update.

The final parameter update is

$$\theta_{t+1} = \theta_t - \eta_t \frac{\rho_t}{\mathrm{RMS}(u_t)} u_t,$$

where $\eta_t > 0$ is the outer learning rate. Thus, the selective map controls the relative strength of the spectral directions, while ρ_t controls the overall update magnitude. Algorithm 1 summarizes the complete update for one parameter block.

(5) **Shared-momentum combination with Muon.** The Selective transform can also be combined with Muon through a common momentum state. Because the same momentum history is retained across iterations, Selective and Muon can alternate as the transformation applied to the current update without introducing separate optimizer states. We refer to this shared-momentum construction as Selective–Muon. It allows parameter-matrix and sample-space geometry to influence the same training process while preserving a single momentum trajectory.

4 Theoretical Analysis

We now analyze the Selective component introduced in Section 3. We first compare Selective with hard orthogonalization on weak singular directions. We then derive an equivalent parameter-side form of the transform, clarifying its relation to hard inverse-square-root normalization and the identity update. Finally, we show that the Selective update retains the standard first-order stationarity rate under smoothness and positive alignment. All proofs and auxiliary results are deferred to Appendix B.

4.1 Behavior on Weak Singular Directions

The main difference between hard and selective orthogonalization appears when a singular direction becomes weak. For a singular value $\sigma > 0$, define the relative amplification as the transformed

singular value divided by its original value. Hard orthogonalization therefore has amplification $A_{\mathrm{hard}}(\sigma) = 1/\sigma$, whereas Selective Orthogonalization has $A_\epsilon(\sigma) = g_\epsilon(\sigma)/\sigma$.

Theorem 4.1. *For any fixed $\epsilon > 0$,*

$$\lim_{\sigma \to 0^+} A_{\mathrm{hard}}(\sigma) = \infty, \qquad \lim_{\sigma \to 0^+} A_\epsilon(\sigma) = 1.$$

Moreover, $g_\epsilon(\sigma)$ lies between σ and 1 for every $\sigma > 0$. Thus, the selective map moves each active singular value toward the hard-orthogonalization target without crossing it.

The theorem makes the weak-spectrum behavior explicit. Hard orthogonalization keeps every active direction at unit scale, so its relative amplification grows without bound as the original singular value vanishes. In contrast, the selective transform approaches the identity on the same weak directions. This is the theoretical counterpart of the training behavior examined later in Section 5. The transform also preserves the singular directions of the original update. If $S = U\Sigma V^\top$, then $T_\epsilon(S) = U g_\epsilon(\Sigma) V^\top$; only the singular values change. In particular, the sample-side and parameter-side active subspaces are unchanged. Additional contraction properties of g_ϵ, including its logarithmic action on the spectrum, are given in Appendix B.

4.2 Equivalent Parameter-Side Spectral Form

Although the method is implemented from the sample Gram matrix, it also has a direct parameter-side interpretation. Let $C = S^\top S$ denote the parameter-side unnormalized second-moment matrix and define, on its positive eigenspace, $h_\epsilon(\lambda) = \lambda^{-\lambda/[2(\lambda+\epsilon)]}$.

Theorem 4.2. *For any S and $\epsilon > 0$, the selective transform satisfies*

$$T_\epsilon(S) = S\, h_\epsilon(C), \qquad C = S^\top S,$$

where $h_\epsilon(C)$ is defined by spectral functional calculus on the active eigenspace of C.

This identity gives two equivalent views of the same operation. The sample-space form reshapes the singular values of the per-sample update matrix, while the parameter-side form applies a spectrum-dependent multiplier to the corresponding second-moment directions. The limiting cases make this interpretation especially clear. As $\epsilon \to 0$, $h_\epsilon(\lambda) \to \lambda^{-1/2}$ on every positive eigenvalue, so the transform approaches hard inverse-square-root normalization on the active subspace. As $\epsilon \to \infty$, $h_\epsilon(\lambda) \to 1$, and the transform approaches the identity. Selective Orthogonalization therefore moves continuously between hard spectral normalization and the untransformed update, with the interpolation determined separately along each active direction.

4.3 First-Order Stationarity

The previous results describe how the Selective transform reshapes the update. We next verify that the resulting Selective update remains compatible with a standard first-order descent analysis. Let $u_t = B^{-1}\mathbf{1}^\top T_{\epsilon_t}(S_t)$ be the transformed minibatch update from Section 3, and let $d_t = (\rho_t / \mathrm{RMS}(u_t))u_t$ denote the direction used in the final parameter step. Since the spectral transform changes the minibatch direction, the analysis requires this direction to remain positively aligned with the objective gradient.

Assumption 4.3. *The objective $f : \mathbb{R}^p \to \mathbb{R}$ is L_f-smooth and bounded below by $f_\star$. There exist constants $c_{\min} > 0$ and $\rho_{\max} > 0$ such that $\rho_t \le \rho_{\max}$ for every iteration and, whenever $\nabla f(\theta_t) \neq 0$, we have $u_t \neq 0$ and*

$$\frac{\langle \nabla f(\theta_t), u_t \rangle}{\|\nabla f(\theta_t)\|\, \|u_t\|} \ge c_{\min}.$$

The RMS construction in Section 3 also gives $\rho_t \ge \underline{\rho}$, where $\underline{\rho} := \min\{\rho_0, \rho_{\min}\} > 0$. Together, these conditions provide a uniform lower bound on descent alignment and a bounded update scale.

Theorem 4.4. *Under Assumption 4.3, for a fixed horizon N and some fixed $\eta > 0$, set $\eta_t = \eta/\sqrt{N}$ for $t = 0, \dots, N-1$. Then the iterates generated by Selective Orthogonalization satisfy*

$$\min_{0 \le t < N} \|\nabla f(\theta_t)\| = \mathcal{O}\Big(N^{-1/2}\Big).$$

Thus, the selective spectral transformation and RMS scaling preserve the standard first-order stationarity rate for smooth nonconvex optimization under positive alignment (Ghadimi & Lan, 2013). The explicit finite-horizon bound and the one-step descent argument are given in Appendix B.

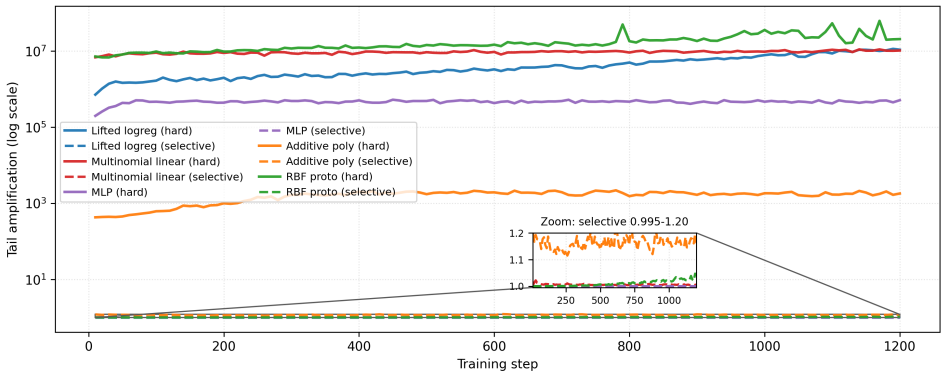


Figure 1: Weak-direction amplification over training on the synthetic weak-spectrum benchmark. Hard orthogonalization produces persistent orders-of-magnitude amplification, whereas Selective remains close to the original weak-direction scale across all five parameter structures.

5 Experiments

We evaluate the proposed approach from three complementary perspectives. First, controlled synthetic experiments isolate the weak-direction mechanism and test the same sample-space construction across different parameter structures. Second, real-data experiments examine whether weak singular directions arise during deep-network training and whether Selective reduces their amplification while maintaining strong optimization performance. Third, we evaluate Selective–Muon to test whether adding sample-space geometry improves the corresponding matrix-orthogonalized Muon baseline. We consider CIFAR-10/Airbench94, CIFAR-10/ResNet-18, and CIFAR-100/Airbench-style classification across multiple physical batch sizes.

5.1 Synthetic Experiments

Experimental setup. We use controlled classification problems to isolate the spectral behavior of the optimizer from the complexity of a particular deep architecture. For the binary setting, the two classes are generated from separated two-dimensional Gaussian distributions. For the multinomial linear classifier, we use the corresponding multi-class construction with class means placed evenly on a circle in the two-dimensional plane.

We evaluate five model families with different parameter structures: a lifted logistic regression model with vector parameters, a multinomial linear classifier with a matrix parameter, a small multilayer perceptron (MLP) with standard matrix-valued weights, an additive polynomial model with multiple vector parameter blocks, and a radial basis function (RBF) prototype classifier with heterogeneous center, scale, coefficient, and bias parameters. The same sample-space spectral rule is used in all five cases. For each model, we consider a default benchmark and a weak-spectrum benchmark. The default benchmark uses the original low-dimensional features. The weak-spectrum benchmark augments the representation as $\phi(x, y) = [\phi_{\mathrm{sig}}(x), \phi_{\mathrm{red}}(x), \phi_{\mathrm{tail}}(x, y)]$. Here, ϕ_{sig} contains the main predictive features, ϕ_{red} contains redundant features that increase spectral anisotropy, and ϕ_{tail} contains low-magnitude, high-dimensional features that create weak singular directions.

We run the default benchmark for 400 optimization steps and the weak-spectrum benchmark for 1200 steps. All reported results are averaged over seeds $0, \ldots, 4$. Hyperparameters are selected by validation loss using the same tuning protocol for Hard and Selective. We report validation loss as the optimization metric and weak-direction amplification as the spectral diagnostic. Following Theorem 4.1, weak-direction amplification is measured as the mean relative amplification over the weakest active singular directions of the per-sample update matrix.

Weak-spectrum behavior. Figure 1 shows a consistent separation between Hard and Selective across all five parameter structures. Hard produces persistent weak-direction amplification ranging from approximately 10^3 to 10^7, whereas Selective remains close to unit amplification for most parameter blocks. The separation persists throughout training and directly matches the contrasting weak-direction limits in Theorem 4.1.

Table 1: Synthetic results across five parameter structures. The default benchmark reports tuned validation loss. The weak-spectrum benchmark reports validation loss and weak-direction amplification. Lower loss is better, while amplification closer to 1 indicates stronger preservation of the original weak-direction scale. For the RBF prototype classifier, α, μ, and $\log s$ denote the coefficient, center, and log-scale parameter blocks, respectively.

Parameter structure	Model	Default loss		Weak-spectrum loss		Weak-direction amplification	
		Hard	Selective	Hard	Selective	Hard	Selective
Vector	Lifted logistic	4.196×10^{-4}	$\mathbf{1.035\times10^{-4}}$	0.3013	**0.2532**	1.079×10^{7}	**1.000**
Matrix, non-deep	Multinomial linear	2.126×10^{-3}	$\mathbf{1.869\times10^{-3}}$	0.4151	**0.3663**	1.026×10^{7}	**1.006**
Matrix NN	MLP	1.534×10^{-4}	$\mathbf{1.447\times10^{-4}}$	0.2652	**0.2639**	5.083×10^{5}	**1.000**
Multi-vector	Additive polynomial	2.638×10^{-4}	$\mathbf{2.336\times10^{-4}}$	0.2375	**0.2368**	1.801×10^{3}	**1.161**
Prototype/RBF	RBF prototype	2.556×10^{-4}	$\mathbf{1.899\times10^{-4}}$	0.2427	**0.2387**	α: 6.791×10^{7} μ: 7.581×10^{2} $\log s$: 1.097×10^{3}	α: **1.000** μ: **1.112** $\log s$: **1.041**

Optimization across parameter structures. Table 1 shows that Selective attains lower tuned validation loss than Hard across all five models in both the default and weak-spectrum benchmarks. The difference in optimization loss is modest in several cases, whereas the reduction in weak-direction amplification is substantially larger. Together, these results show that Selective preserves the intended weak-direction behavior without sacrificing optimization performance across vector, matrix, multi-vector, and heterogeneous parameter structures.

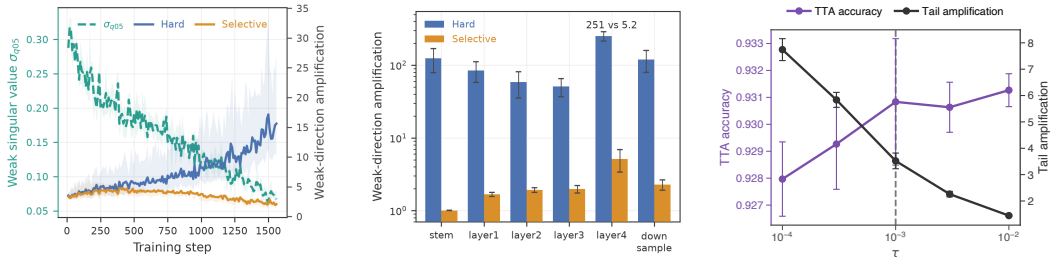
Ablation study. We further examine whether the improvement of Selective can be explained by global update scaling alone. We vary the target RMS ρ_0 for Hard on the weak-spectrum benchmark and compare the resulting validation loss with Selective. Across the tested range, retuning the global scale does not recover the selective result. We also track the adaptive transition scale ϵ_t, transition weight ω_t, target RMS ρ_t, and transformed update scale $\mathrm{RMS}(u_t)$. These diagnostics confirm that the selective transition remains active throughout training rather than collapsing to either the hard or identity limit. Together, the ablation supports that the observed difference comes from the singular-value-dependent transformation rather than from global rescaling alone.

5.2 Real-World Experiments

Experimental setup. We evaluate the methods on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), using Airbench94 (Jordan, 2024), ResNet-18 (He et al., 2016), and an Airbench-style CIFAR-100 setting. Physical batch sizes are 64, 128, and 256 under a fixed eight-epoch training horizon. We compare AdamW, Hard, Selective, Muon, and Selective–Muon. Here, Hard denotes hard polar orthogonalization applied to the per-sample update matrix S_t, while Selective applies the proposed sample-space transform to the same object. Muon provides the matrix-orthogonalized baseline, and Selective–Muon uses the shared-momentum alternating construction from Section 3. TTA test accuracy is the main performance metric (Shanmugam et al., 2021). The main table reports mean accuracy over four seeds. For the spectral diagnostics, we additionally track weak singular values of the per-sample update matrix and the corresponding relative amplification under Hard and Selective. Unless stated otherwise, Selective uses $\tau = 10^{-3}$.

Weak-spectrum behavior in real training. Figure 2(a) shows the weak-spectrum dynamics on CIFAR-10/Airbench94 at batch size 256. As the 5th-percentile active singular value $\sigma_{0.05}$ decreases during training, the corresponding Hard amplification increases, whereas Selective remains substantially lower. This directly matches the contrasting weak-direction behavior predicted by Theorem 4.1. Figure 2(b) shows the layerwise behavior on CIFAR-10/ResNet-18 at batch size 128. Hard produces substantially larger weak-direction amplification across network stages. In layer4, for example, amplification reaches 251.3 ± 36.7 under Hard and 5.15 ± 1.76 under Selective. The same effect therefore persists after transferring from Airbench94 to a substantially different architecture.

Optimization performance. Table 2 reports TTA test accuracy across all three real-data settings and physical batch sizes. Selective consistently improves over Hard across the nine tested settings, showing that substantially reducing weak-direction amplification can improve downstream optimization performance. The hybrid optimizer Selective–Muon further improves over the corresponding Muon baseline in all nine settings. This result supports augmenting matrix orthogonalization with sample-space geometry as an effective optimization strategy.



(a) CIFAR-10 / Airbench94 (b) CIFAR-10 / ResNet-18 (c) Sensitivity to τ

Figure 2: Weak-spectrum diagnostics and sensitivity on CIFAR-10: (a) training dynamics on Airbench94, (b) layerwise weak-direction amplification in ResNet-18, and (c) sensitivity to τ on Airbench94.

Table 2: TTA test accuracy across datasets, architectures, and physical batch sizes. The final two columns isolate the effect of augmenting Muon with sample-space geometry. An upward arrow indicates improvement over the corresponding Muon baseline.

		Reference	Sample-space optimizer		Matrix optimizer	Hybrid optimizer
Setting	Batch size	AdamW	Hard	Selective	Muon	Selective–Muon ↑
CIFAR-10 Airbench94	64	0.9107	0.9182	**0.9217**	0.9081	0.9144↑
	128	0.8988	0.9237	**0.9272**	0.9196	0.9225↑
	256	0.9185	0.9291	**0.9311**	0.9277	0.9291↑
CIFAR-10 ResNet-18	64	0.9032	0.9075	0.9086	0.9069	**0.9094**↑
	128	0.8965	0.9123	0.9144	0.9184	**0.9194**↑
	256	0.8852	0.9152	0.9181	0.9159	**0.9258**↑
CIFAR-100 Airbench-style	64	0.5536	0.6522	**0.6530**	0.5712	0.6106↑
	128	0.5225	0.6725	**0.6741**	0.6208	0.6475↑
	256	0.4744	0.6802	**0.6804**	0.6565	0.6752↑

Sensitivity to the transition parameter τ. Figure 2(c) examines sensitivity to τ on CIFAR-10/Airbench94 at batch size 256 over three seeds. TTA test accuracy remains stable across the tested range, while weak-direction amplification decreases as τ increases. The default value $\tau = 10^{-3}$ lies within this stable region and already provides a substantial reduction in weak-direction amplification.

6 Discussion

Our results highlight sample-space geometry as a complementary view of orthogonalized optimization. Matrix orthogonalization acts on the row and column structure of aggregated updates, whereas per-sample updates expose co-update directions across training examples. Selective Orthogonalization provides a controlled way to use this sample-space spectrum. Rather than mapping every active singular value to the same target, it retains hard-like normalization on strong directions while allowing sufficiently weak directions to remain close to their original scale. Theory and experiments consistently show substantially reduced weak-direction amplification together with strong optimization performance across parameter structures, datasets, architectures, and batch sizes. Selective–Muon further shows the value of combining the two geometric views. Augmenting matrix orthogonalization with sample-space geometry improves the corresponding Muon baseline in all nine tested settings, indicating that the two forms of spectral structure can contribute within the same optimization trajectory. This complements the standalone Selective results: the former demonstrates the value of adding sample-space geometry to matrix orthogonalization, while the latter isolates the weak-direction behavior that makes such spectral shaping effective. The spectral structure of the update is a property of the data and model, not of training scale. Therefore, the mechanism—that hard orthogonalization amplifies vanishing directions while Selective preserves them—is testable at CIFAR scale, and the same structure is expected to appear at larger scales. The current exact implementation requires per-sample updates and additional sample-space spectral computation, and thus more efficient implementations and larger-scale training remain directions for future work. More broadly, the results suggest that orthogonalized optimization can benefit from combining complementary update geometries while adapting spectral normalization to directional scale.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and discovery, drafting and polishing parts of the manuscript, and the development and checking of mathematical arguments and proofs. The core research idea, methodological development, experimental design, and interpretation of the results were carried out by the authors. All references suggested by AI tools were manually verified against the original sources, and all mathematical claims, derivations, and proofs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and experimental results produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the full specification of the proposed method in the main paper, complete proofs in the appendix, and detailed experimental settings and additional diagnostics in the supplementary sections. Anonymous source code and experiment configurations are included in the supplementary material to support reproduction of the reported results.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 138323–138360, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/df88b275bef31ac96c85f0c4013734fc-Paper-Conference.pdf.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. In *NeurIPS 2024 Workshops: OPT*, 2024. URL https://mlanthology.org/neuripsw/2024/bernstein2024neuripsw-old/.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic Spectral Descent for Restricted Boltzmann Machines. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 111–119, San Diego, California, USA, 09–12 May 2015a. PMLR. URL https://proceedings.mlr.press/v38/carlson15.html.

David Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

Anirudh Chakravarthy, Shuai Zheng, Xin Huang, Sachithra Hemachandra, Xiao Zhang, Yuning Chai, and Zhao Chen. Profit: A specialized optimizer for deep fine tuning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 12815–12844. Curran Associates, Inc., 2025. doi: 10.52202/085713-0432. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/130534d1b9f0a54bb59b529589123cb4-Paper-Conference.pdf.

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. URL https://doi.org/10.1137/120880811.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Nicholas J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008. doi: 10.1137/1.9780898717778. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717778.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Keller Jordan. 94% on cifar-10 in 3.29 seconds on a single gpu. *arXiv preprint arXiv:2404.00498*, 2024.

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

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 17389–17416, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1d485d337d5622bdb1462333f4247522-Paper-Conference.pdf.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4c0129f1f7a036af5ed2409f1f9f8131-Paper-Conference.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://mlanthology.org/iclr/2015/kingma2015iclr-adam/.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://mlanthology.org/iclr/2019/loshchilov2019iclr-decoupled/.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

Yehonathan Refael, Guy Smorodinsky, Tom Tirer, and Ofir Lindenbaum. Sumo: Subspace-aware moment-orthogonalization for accelerating memory-efficient llm training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 147250–147281. Curran Associates, Inc., 2025. doi: 10.52202/085713-4926. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d85a66edadd443ac2350e93c0287f4f9-Paper-Conference.pdf.

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8469–8479. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sankararaman20a.html.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1214–1223, October 2021.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact muon update. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5167–5175. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/shulgin26a.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation, 2022. URL https://arxiv.org/abs/2202.07052.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 37382–37419, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3eec5006051d9544e717067de3220198-Paper-Conference.pdf.

Huaijin Wu, Bingrui Li, Yebin Yang, Yi Tu, Zhanpeng Zhou, Jianfei Chen, and Junchi Yan. Achieving low-bit muon through subspace preservation and grid quantization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 139395–139414, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e1af74bc0a8ab547b27e507d71be506a-Paper-Conference.pdf.

A Additional Derivations for the Spectral Setup

This section gives the linear-algebra details behind the sample-space and parameter-side constructions in Section 2. Throughout this section, let

$$S \in \mathbb{R}^{B\times p}, \qquad K := SS^\top, \qquad C := S^\top S,$$

where K is the sample Gram matrix and C is the parameter-side unnormalized second-moment matrix.

A.1 Parameter-side energy identity

For any $a \in \mathbb{R}^p$,

$$a^\top Ca = a^\top S^\top Sa = (Sa)^\top (Sa) = \|Sa\|_2^2.$$

If the b-th row of S is $s_b^\top$, then the b-th entry of Sa is $s_b^\top a$. Hence

$$\|Sa\|_2^2 = \sum_{b=1}^{B} (s_b^\top a)^2.$$

Therefore,

$$a^\top Ca = \sum_{b=1}^{B} (s_b^\top a)^2,$$

so the quadratic form of C measures the total squared minibatch update energy along the parameter-space direction a.

Now let

$$S = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r), \qquad \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

be the compact singular value decomposition of S. Then

$$C = S^\top S = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top.$$

Thus each right singular vector v_k is an eigenvector of C with eigenvalue σ_k^2, and

$$v_k^\top Cv_k = \sigma_k^2.$$

This gives the parameter-side energy interpretation used in Section 2.

A.2 Shared nonzero spectrum

Using the same compact SVD,

$$K = SS^\top = U\Sigma V^\top V\Sigma U^\top = U\Sigma^2 U^\top,$$

whereas

$$C = S^\top S = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top.$$

Therefore, the nonzero eigenvalues of both K and C are

$$\sigma_1^2, \dots, \sigma_r^2.$$

Equivalently,

$$\lambda_i(K) = \lambda_i(C) = \sigma_i^2, \qquad i = 1, \dots, r.$$

The two matrices differ in their eigenvectors and ambient dimensions, but they encode exactly the same active spectrum. This identity is what allows the method to obtain the active parameter-side spectrum from the $B \times B$ sample Gram matrix.

B Proofs and Auxiliary Results for Section 4

Throughout this section, let

$$g_\epsilon(\sigma) = \sigma^{\epsilon/(\sigma^2+\epsilon)}, \qquad \epsilon > 0,$$

with $g_\epsilon(0) = 0$.

B.1 Proof of Theorem 4.1

Proof. Hard orthogonalization maps every nonzero singular value to 1, so

$$A_{\mathrm{hard}}(\sigma) = \frac{1}{\sigma}.$$

Therefore,

$$\lim_{\sigma \to 0^+} A_{\mathrm{hard}}(\sigma) = +\infty.$$

For Selective Orthogonalization,

$$A_\epsilon(\sigma) = \frac{g_\epsilon(\sigma)}{\sigma} = \sigma^{\epsilon/(\sigma^2+\epsilon)-1} = \sigma^{-\sigma^2/(\sigma^2+\epsilon)}.$$

Taking logarithms gives

$$\log A_\epsilon(\sigma) = -\frac{\sigma^2}{\sigma^2+\epsilon} \log \sigma.$$

Since

$$0 \le \frac{\sigma^2}{\sigma^2+\epsilon} |\log \sigma| \le \frac{\sigma^2}{\epsilon} |\log \sigma|$$

and $\sigma^2 |\log \sigma| \to 0$ as $\sigma \to 0^+$, it follows that

$$\log A_\epsilon(\sigma) \to 0.$$

Hence

$$\lim_{\sigma \to 0^+} A_\epsilon(\sigma) = 1.$$

It remains to locate $g_\epsilon(\sigma)$ relative to σ and 1. Write

$$g_\epsilon(\sigma) = \sigma^{q_\epsilon(\sigma)}, \qquad q_\epsilon(\sigma) := \frac{\epsilon}{\sigma^2+\epsilon}.$$

For every $\sigma > 0$, $q_\epsilon(\sigma) \in (0,1)$. If $0 < \sigma < 1$, raising σ to an exponent in $(0,1)$ increases it without crossing 1, so

$$\sigma < g_\epsilon(\sigma) < 1.$$

If $\sigma > 1$, the same exponent decreases it without crossing 1, so

$$1 < g_\epsilon(\sigma) < \sigma.$$

Finally, $g_\epsilon(1) = 1$. Combining these cases gives

$$g_\epsilon(\sigma) \in [\min\{\sigma, 1\}, \max\{\sigma, 1\}],$$

which proves the theorem. □

B.2 Spectral action of the selective map

The following proposition records properties that support the interpretation in Section 4 but are not needed for the main weak-direction theorem.

Proposition B.1 (Spectral action)**.** *Let*

$$S = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r), \qquad \sigma_i > 0,$$

be the compact SVD of S. Then

$$T_\epsilon(S) = U\,\mathrm{diag}\big(g_\epsilon(\sigma_1), \dots, g_\epsilon(\sigma_r)\big)V^\top$$

has the same rank and the same active left and right singular subspaces as S. Its active singular values are $g_\epsilon(\sigma_i)$. Moreover,

$$g_\epsilon(\sigma) \in [\min\{\sigma, 1\}, \max\{\sigma, 1\}]$$

for every $\sigma > 0$, and

$$|\log g_\epsilon(\sigma)| \le |\log \sigma|,$$

with strict inequality whenever $\sigma \neq 1$.

Proof. For every active singular value $\sigma_i > 0$,

$$g_\epsilon(\sigma_i) = \sigma_i^{\epsilon/(\sigma_i^2+\epsilon)} > 0.$$

Therefore the transform does not remove any active singular direction and does not introduce a new one. It has rank r, the same left singular vectors U, and the same right singular vectors V.

The interval relation was established in the proof of Theorem 4.1. For the logarithmic relation,

$$\log g_\epsilon(\sigma) = \frac{\epsilon}{\sigma^2+\epsilon} \log \sigma.$$

Hence

$$|\log g_\epsilon(\sigma)| = \frac{\epsilon}{\sigma^2+\epsilon} |\log \sigma| \le |\log \sigma|.$$

If $\sigma \neq 1$, then $|\log \sigma| > 0$ and $0 < \epsilon/(\sigma^2+\epsilon) < 1$, so the inequality is strict. □

The final statement gives a precise sense in which every active singular value moves toward the hard target 1 in log scale without being forced directly to it. In particular,

$$g_\epsilon(\sigma) = 1$$

for finite $\epsilon > 0$ if and only if $\sigma = 1$.

B.3 Log-distance contraction toward the hard target

The coordinatewise logarithmic contraction above extends directly to aggregate measures of distance from the unit singular spectrum.

Corollary B.2 (Log-distance contraction)**.** *Let*

$$\tilde{\sigma}_i := g_\epsilon(\sigma_i), \qquad i = 1, \dots, r,$$

and for $q \in [1, \infty]$ define

$$\mathcal{D}_q(\sigma_1, \dots, \sigma_r) := \begin{cases} \left(\sum_{i=1}^r |\log \sigma_i|^q\right)^{1/q}, & 1 \le q < \infty, \\ \max_{1 \le i \le r} |\log \sigma_i|, & q = \infty. \end{cases}$$

Then

$$\mathcal{D}_q(\tilde{\sigma}_1, \dots, \tilde{\sigma}_r) \le \mathcal{D}_q(\sigma_1, \dots, \sigma_r).$$

For every finite q, the inequality is strict whenever at least one $\sigma_i \neq 1$.

Proof. Proposition B.1 gives

$$|\log \tilde{\sigma}_i| \le |\log \sigma_i|$$

for every i. Therefore the vector

$$\big(|\log \tilde{\sigma}_1|, \dots, |\log \tilde{\sigma}_r|\big)$$

is coordinatewise bounded by

$$\big(|\log \sigma_1|, \dots, |\log \sigma_r|\big).$$

Monotonicity of the ℓ_q norm on nonnegative vectors yields the stated inequality for all $q \in [1, \infty]$. If at least one $\sigma_i \neq 1$, the corresponding coordinate inequality is strict, which makes the finite-q sum strictly smaller. □

B.4 Amplification-factor interpretation

For $\sigma > 0$, define the relative amplification of the selective map by

$$A_\epsilon(\sigma) := \frac{g_\epsilon(\sigma)}{\sigma}.$$

Using the definition of g_ϵ,

$$A_\epsilon(\sigma) = \sigma^{-\sigma^2/(\sigma^2+\epsilon)}.$$

Hard orthogonalization instead has

$$A_{\mathrm{hard}}(\sigma) = \frac{1}{\sigma}.$$

For $0 < \sigma < 1$, the exponent $\sigma^2/(\sigma^2 + \epsilon)$ lies in $(0, 1)$, so

$$1 < A_\epsilon(\sigma) < A_{\mathrm{hard}}(\sigma).$$

Thus both transforms amplify a weak but nonzero singular value relative to its original scale, but Selective does so less strongly. More importantly, their limiting behavior differs:

$$\lim_{\sigma \to 0^+} A_\epsilon(\sigma) = 1, \qquad \lim_{\sigma \to 0^+} A_{\mathrm{hard}}(\sigma) = +\infty.$$

This distinction is the mechanism measured by the weak-direction amplification diagnostics in Section 5.

For $\sigma > 1$, both transforms reduce the singular value relative to its original scale. In that regime,

$$\frac{1}{\sigma} < A_\epsilon(\sigma) < 1,$$

so the selective map again lies between the identity transform and hard orthogonalization.

B.5 Proof of Theorem 4.2

Proof. Let

$$S = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \qquad \sigma_i > 0.$$

Then

$$C = S^\top S = V\Sigma^2 V^\top.$$

On the positive eigenspace of C, define

$$h_\epsilon(\lambda) = \lambda^{-\lambda/[2(\lambda+\epsilon)]}.$$

By spectral functional calculus,

$$h_\epsilon(C) = V\,\mathrm{diag}\big(h_\epsilon(\sigma_1^2), \ldots, h_\epsilon(\sigma_r^2)\big)V^\top$$

on the active subspace. Therefore,

$$\begin{aligned} Sh_\epsilon(C) &= U\Sigma\,\mathrm{diag}\big(h_\epsilon(\sigma_1^2), \ldots, h_\epsilon(\sigma_r^2)\big)V^\top \\ &= U\,\mathrm{diag}\left(\sigma_i(\sigma_i^2)^{-\sigma_i^2/[2(\sigma_i^2+\epsilon)]}\right)_{i=1}^r V^\top \\ &= U\,\mathrm{diag}\left(\sigma_i^{1-\sigma_i^2/(\sigma_i^2+\epsilon)}\right)_{i=1}^r V^\top \\ &= U\,\mathrm{diag}\left(\sigma_i^{\epsilon/(\sigma_i^2+\epsilon)}\right)_{i=1}^r V^\top \\ &= T_\epsilon(S). \end{aligned}$$

This proves

$$T_\epsilon(S) = Sh_\epsilon(C).$$

For every positive eigenvalue λ,

$$\lim_{\epsilon \to 0^+} h_\epsilon(\lambda) = \lambda^{-1/2}, \qquad \lim_{\epsilon \to \infty} h_\epsilon(\lambda) = 1.$$

Hence, on the active subspace,

$$\lim_{\epsilon \to 0^+} T_\epsilon(S) = S(C^\dagger)^{1/2} = UV^\top,$$

while

$$\lim_{\epsilon \to \infty} T_\epsilon(S) = S.$$

Thus the parameter-side form has the same hard and identity limits as the singular-value description in the main text. □

B.6 ADAPTIVE TRANSITION SCALE

This subsection gives a fuller interpretation of the spectrum-dependent transition scale in Section 3. We consider the nondegenerate case

$$\mathrm{tr}(K_t) > 0.$$

If $K_t = 0$, then

$$K_t = S_t S_t^\top = 0$$

implies $S_t = 0$, so all per-sample updates for the block vanish and the corresponding block update can be set to zero.

Let

$$\lambda_{t,1}, \dots, \lambda_{t,r_t} > 0$$

be the active eigenvalues of K_t. Since

$$\lambda_{t,i} = \sigma_{t,i}^2,$$

the transition scale is

$$\epsilon_t = \tau \frac{\mathrm{tr}(K_t^2)}{\mathrm{tr}(K_t)} = \tau \frac{\sum_{i=1}^{r_t} \lambda_{t,i}^2}{\sum_{i=1}^{r_t} \lambda_{t,i}}.$$

Define

$$\pi_{t,i} := \frac{\lambda_{t,i}}{\sum_{j=1}^{r_t} \lambda_{t,j}}.$$

Then

$$\pi_{t,i} \geq 0, \qquad \sum_{i=1}^{r_t} \pi_{t,i} = 1,$$

and

$$\frac{\epsilon_t}{\tau} = \sum_{i=1}^{r_t} \pi_{t,i} \lambda_{t,i}.$$

Thus ϵ_t/τ is an eigenvalue-weighted average of the active spectrum. Larger eigenvalues receive larger weights.

Because this is a convex combination of the active eigenvalues,

$$\tau \min_i \lambda_{t,i} \leq \epsilon_t \leq \tau \max_i \lambda_{t,i}.$$

For fixed total spectral mass $\sum_i \lambda_{t,i}$, the numerator $\sum_i \lambda_{t,i}^2$ increases as the spectrum becomes more concentrated. Hence the transition scale becomes larger when more update energy is concentrated in a small number of strong spectral directions.

The same quantity has an effective-rank form. Define

$$r_{\mathrm{eff}}(K_t) = \frac{\mathrm{tr}(K_t)^2}{\mathrm{tr}(K_t^2)} = \frac{\left(\sum_i \lambda_{t,i}\right)^2}{\sum_i \lambda_{t,i}^2}.$$

Then

$$\frac{\mathrm{tr}(K_t)}{r_{\mathrm{eff}}(K_t)} = \frac{\mathrm{tr}(K_t^2)}{\mathrm{tr}(K_t)},$$

and therefore

$$\epsilon_t = \tau \frac{\mathrm{tr}(K_t)}{r_{\mathrm{eff}}(K_t)}.$$

If all r_t active eigenvalues are equal, then

$$r_{\mathrm{eff}}(K_t) = r_t.$$

If a small number of eigenvalues dominate, $r_{\mathrm{eff}}(K_t)$ decreases. For comparable total spectral mass, this increases the transition scale.

Finally, the ratio between σ^2 and ϵ_t determines how the selective map acts on an individual direction. If

$$\sigma^2 \ll \epsilon_t,$$

then

$$\frac{\epsilon_t}{\sigma^2 + \epsilon_t} \approx 1$$

and

$$g_{\epsilon_t}(\sigma) \approx \sigma.$$

The weak singular value is therefore largely preserved. If

$$\sigma^2 \gg \epsilon_t,$$

then

$$\frac{\epsilon_t}{\sigma^2 + \epsilon_t} \approx 0$$

and

$$g_{\epsilon_t}(\sigma) \approx 1.$$

The strong singular value is therefore treated more like hard orthogonalization. This gives the weak-to-strong transition described in Section 3.

B.7 Proof of Theorem 4.4

If $\nabla f(\theta_t) = 0$ for some $t < N$, then

$$\min_{0 \le s < N} \|\nabla f(\theta_s)\| = 0,$$

and the conclusion of Theorem 4.4 is immediate. We therefore consider the remaining case in which $\nabla f(\theta_t) \neq 0$ for every $t = 0, \dots, N-1$. Under Assumption 4.3, $u_t \neq 0$ in this case, so define

$$c_t := \frac{\langle \nabla f(\theta_t), u_t \rangle}{\|\nabla f(\theta_t)\| \, \|u_t\|}.$$

We first state the two elementary facts used in the proof.

Lemma B.3 (One-step descent)**.** *Suppose f is L_f-smooth. Let*

$$d_t = \frac{\rho_t}{\mathrm{RMS}(u_t)} u_t, \qquad \theta_{t+1} = \theta_t - \eta_t d_t.$$

Then

$$f(\theta_{t+1}) \le f(\theta_t) - \eta_t \rho_t \sqrt{p}\, c_t \|\nabla f(\theta_t)\| + \frac{L_f}{2} \eta_t^2 \rho_t^2 p.$$

Proof. By L_f-smoothness,

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2} \|y - x\|^2$$

for all $x, y \in \mathbb{R}^p$. Set

$$x = \theta_t, \qquad y = \theta_t - \eta_t d_t.$$

Then

$$f(\theta_{t+1}) \le f(\theta_t) - \eta_t \langle \nabla f(\theta_t), d_t \rangle + \frac{L_f}{2} \eta_t^2 \|d_t\|^2.$$

Since

$$\mathrm{RMS}(u_t) = \frac{\|u_t\|}{\sqrt{p}},$$

we have

$$d_t = \rho_t \sqrt{p}\, \frac{u_t}{\|u_t\|}$$

and therefore

$$\|d_t\| = \rho_t \sqrt{p}.$$

Moreover,

$$\begin{aligned} \langle \nabla f(\theta_t), d_t \rangle &= \rho_t \sqrt{p}\, \frac{\langle \nabla f(\theta_t), u_t \rangle}{\|u_t\|} \\ &= \rho_t \sqrt{p}\, c_t \|\nabla f(\theta_t)\|. \end{aligned}$$

Substitution gives the result. □

Lemma B.4 (Lower bound on the RMS target)**.** *For every iteration,*

$$\rho_t \geq \underline{\rho} := \min\{\rho_0, \rho_{\min}\} > 0.$$

Proof. Section 3 defines

$$\rho_t = (1 - \omega_t)\rho_0 + \omega_t \max\{\rho_{\min}, \mathrm{RMS}(\bar{s}_t)\},$$

where

$$\omega_t = \frac{\epsilon_t}{\lambda_{\max}(K_t) + \epsilon_t} \in [0, 1].$$

Because

$$\max\{\rho_{\min}, \mathrm{RMS}(\bar{s}_t)\} \geq \rho_{\min},$$

we obtain

$$\rho_t \geq (1 - \omega_t)\rho_0 + \omega_t \rho_{\min}.$$

A convex combination of ρ_0 and $\rho_{\min}$ is at least their minimum, so

$$\rho_t \geq \min\{\rho_0, \rho_{\min}\}.$$

□

Proof of Theorem 4.4. Lemma B.3 gives

$$f(\theta_{t+1}) \leq f(\theta_t) - \eta_t \rho_t \sqrt{p}\, c_t \|\nabla f(\theta_t)\| + \frac{L_f}{2} \eta_t^2 \rho_t^2 p.$$

By Assumption 4.3 and Lemma B.4,

$$c_t \geq c_{\min}, \qquad \rho_t \geq \underline{\rho}, \qquad \rho_t \leq \rho_{\max}.$$

Hence

$$f(\theta_{t+1}) \leq f(\theta_t) - \eta_t \underline{\rho} \sqrt{p}\, c_{\min} \|\nabla f(\theta_t)\| + \frac{L_f}{2} \eta_t^2 \rho_{\max}^2 p.$$

Rearranging,

$$\eta_t \underline{\rho} \sqrt{p}\, c_{\min} \|\nabla f(\theta_t)\| \leq f(\theta_t) - f(\theta_{t+1}) + \frac{L_f}{2} \eta_t^2 \rho_{\max}^2 p.$$

Summing from $t = 0$ to $N - 1$ yields

$$\underline{\rho} \sqrt{p}\, c_{\min} \sum_{t=0}^{N-1} \eta_t \|\nabla f(\theta_t)\| \leq f(\theta_0) - f(\theta_N) + \frac{L_f}{2} \rho_{\max}^2 p \sum_{t=0}^{N-1} \eta_t^2.$$

Because f is bounded below by $f_\star$,

$$f(\theta_N) \geq f_\star,$$

so

$$\underline{\rho} \sqrt{p}\, c_{\min} \sum_{t=0}^{N-1} \eta_t \|\nabla f(\theta_t)\| \leq f(\theta_0) - f_\star + \frac{L_f}{2} \rho_{\max}^2 p \sum_{t=0}^{N-1} \eta_t^2.$$

Now set

$$\eta_t = \frac{\eta}{\sqrt{N}}.$$

Then

$$\sum_{t=0}^{N-1} \eta_t = \eta \sqrt{N}, \qquad \sum_{t=0}^{N-1} \eta_t^2 = \eta^2.$$

Also,

$$\sum_{t=0}^{N-1} \eta_t \|\nabla f(\theta_t)\| \geq \left(\sum_{t=0}^{N-1} \eta_t \right) \min_{0 \leq t < N} \|\nabla f(\theta_t)\|.$$

Combining these relations gives

$$\underline{\rho} \sqrt{p}\, c_{\min} \eta \sqrt{N} \min_{0 \leq t < N} \|\nabla f(\theta_t)\| \leq f(\theta_0) - f_\star + \frac{L_f}{2} \eta^2 \rho_{\max}^2 p.$$

Dividing by $\underline{\rho}\sqrt{p}\, c_{\min}\eta\sqrt{N}$ gives the explicit finite-horizon bound

$$\min_{0\le t<N} \|\nabla f(\theta_t)\| \le \frac{f(\theta_0)-f_\star}{\eta c_{\min}\underline{\rho}\sqrt{p}}\frac{1}{\sqrt{N}} + \frac{L_f \eta \rho_{\max}^2\sqrt{p}}{2c_{\min}\underline{\rho}}\frac{1}{\sqrt{N}}.$$

Therefore,

$$\min_{0\le t<N} \|\nabla f(\theta_t)\| = \mathcal{O}(N^{-1/2}).$$

□

B.8 Interpretation of the alignment condition

For iterations with $\nabla f(\theta_t) \neq 0$, the coefficient

$$c_t = \frac{\langle \nabla f(\theta_t), u_t\rangle}{\|\nabla f(\theta_t)\|\,\|u_t\|}$$

is the cosine similarity between the transformed update u_t and the objective gradient. The condition

$$c_t \ge c_{\min} > 0$$

requires the transformed update to remain uniformly descent-aligned. Under this condition, the proof above follows the standard smoothness-based descent argument for a normalized first-order direction. If the objective gradient is exactly zero at any iterate, the stationarity claim is already satisfied and no cosine similarity needs to be defined there.

C Sample-Side Implementation Identities

This section derives the sample-side implementation used in Section 3. Although the selective transform is defined through the singular value decomposition of S_t, its active spectral quantities and final aggregate can be computed from the eigendecomposition of

$$K_t = S_t S_t^\top.$$

C.1 Rewriting the selective transform from the sample side

Let

$$S_t = U_t \Sigma_t V_t^\top, \qquad \Sigma_t = \operatorname{diag}(\sigma_{t,1},\dots,\sigma_{t,r_t}), \qquad \sigma_{t,i} > 0,$$

be the compact SVD of S_t. By definition,

$$T_{\epsilon_t}(S_t) = U_t \operatorname{diag}\big(g_{\epsilon_t}(\sigma_{t,1}),\dots,g_{\epsilon_t}(\sigma_{t,r_t})\big)V_t^\top.$$

On the active singular subspace,

$$V_t^\top = \Sigma_t^{-1} U_t^\top S_t.$$

Substituting gives

$$\begin{aligned} T_{\epsilon_t}(S_t) &= U_t \operatorname{diag}\big(g_{\epsilon_t}(\sigma_{t,1}),\dots,g_{\epsilon_t}(\sigma_{t,r_t})\big)\Sigma_t^{-1}U_t^\top S_t \\ &= U_t \operatorname{diag}(\gamma_t) U_t^\top S_t, \end{aligned}$$

where

$$\gamma_t := \left(\frac{g_{\epsilon_t}(\sigma_{t,1})}{\sigma_{t,1}},\dots,\frac{g_{\epsilon_t}(\sigma_{t,r_t})}{\sigma_{t,r_t}}\right).$$

This is the sample-side transform used in the algorithm.

C.2 Derivation of the aggregate update

The transformed minibatch aggregate is

$$u_t = \frac{1}{B}\mathbf{1}^\top T_{\epsilon_t}(S_t).$$

Using the sample-side form,

$$u_t = \frac{1}{B}\mathbf{1}^\top U_t \operatorname{diag}(\gamma_t) U_t^\top S_t.$$

Define

$$\xi_t^\top := \frac{1}{B}\mathbf{1}^\top U_t \operatorname{diag}(\gamma_t) U_t^\top \in \mathbb{R}^{1\times B}.$$

Then

$$u_t = \xi_t^\top S_t.$$

Thus all spectral coefficients are formed in sample space, and the final parameter-space direction is obtained by a weighted aggregation of the original per-sample updates.

C.3 Connection to the Gram matrix

The required sample-side quantities follow from

$$K_t = S_t S_t^\top.$$

Using the compact SVD of S_t,

$$K_t = U_t \Sigma_t^2 U_t^\top.$$

Therefore, the active eigenvectors of K_t give U_t, while its positive eigenvalues give

$$\lambda_{t,i} = \sigma_{t,i}^2.$$

The singular values needed by the selective map are therefore

$$\sigma_{t,i} = \sqrt{\lambda_{t,i}}.$$

Consequently, both γ_t and ξ_t can be formed from the eigendecomposition of K_t without recovering the parameter-side singular vectors V_t or decomposing $C_t = S_t^\top S_t$.

C.4 RMS and computational form

After the transformed aggregate has been computed,

$$u_t = \xi_t^\top S_t,$$

the root-mean-square quantity in the final update is

$$\mathrm{RMS}(u_t) = \frac{\|u_t\|_2}{\sqrt{p}}.$$

The spectral computation therefore acts on the $B \times B$ matrix K_t, while the remaining parameter-space operation is the weighted aggregation

$$\xi_t^\top S_t.$$

When $B \ll p$, this avoids forming or spectrally decomposing the $p \times p$ parameter-side unnormalized second-moment matrix.

This dimensional reduction concerns the spectral decomposition itself. The exact implementation still requires access to per-sample updates S_t, whose extraction can be a substantial practical cost, as noted in Section 6.

D Additional Experimental Results

This section provides additional diagnostics for the ablation study in Section 5.

The trajectories complement the global-scale control reported in the main text. In particular, they show that the adaptive quantities continue to evolve during training and that the selective transition remains active rather than collapsing to either the hard or identity limit.

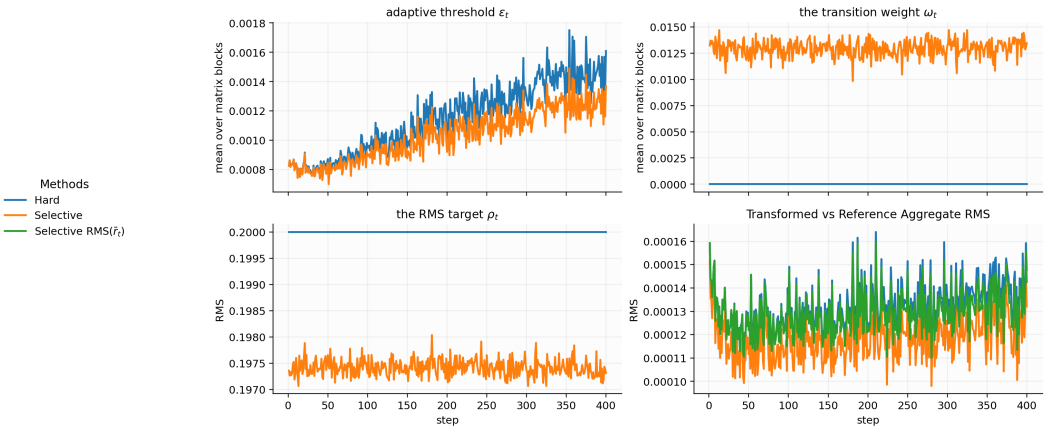


Figure 3: Optimizer diagnostics in a representative weak-spectrum experiment. The panels track the adaptive transition scale ϵ_t, the transition weight ω_t, the RMS target ρ_t, and the transformed aggregate scale $\mathrm{RMS}(u_t)$.

E Additional Experimental Details

This section records additional implementation and runtime details for the experiments in Section 5. The purpose is reproducibility; the runtime measurements below are not intended as optimized systems benchmarks.

E.1 Synthetic experiments

The synthetic experiments use the five model families described in the main text: lifted logistic regression, a multinomial linear classifier, a small MLP, an additive polynomial model, and an RBF prototype classifier. These models cover vector, matrix, multi-vector, and heterogeneous parameter blocks, while using the same sample-space selective rule.

The default benchmark is run for 400 optimization steps. The weak-spectrum benchmark is run for 1200 steps. All reported synthetic results are averaged over seeds

$$0, \ldots, 4.$$

Hard and Selective use the same validation-loss tuning protocol. The weak-spectrum benchmark follows the feature construction in Section 5, with predictive features, redundant features that increase spectral anisotropy, and low-magnitude high-dimensional features that create weak singular directions.

The synthetic experiments are run on CPU. The machine used for these runs is a MacBook Air with an Apple M2 chip, 8 CPU cores (4 performance and 4 efficiency cores), 8 logical CPUs, and 24 GB of unified memory. Table 3 reports wall-clock runtime in seconds as mean $\pm$ population standard deviation over the five seeds.

The synthetic runtime results show that Hard and Selective have similar wall-clock cost in these controlled CPU experiments. This comparison isolates the small-scale experimental implementation and should not be interpreted as a general systems-level efficiency result.

E.2 Real-world experiments

The real-world evaluation covers CIFAR-10/Airbench94, CIFAR-10/ResNet-18, and CIFAR-100/Airbench-style classification. All settings use a fixed eight-epoch training horizon and physical batch sizes 64, 128, and 256. We compare AdamW, Hard, Selective, Muon, and the hybrid optimizer Selective–Muon. Hard and Selective operate on the per-sample update matrix, Muon applies matrix orthogonalization, and Selective–Muon alternates the two transformations while maintaining the shared momentum state described in Section 3.

Table 3: Runtime summary for the synthetic experiments. Default runs use 400 steps and weak-spectrum runs use 1200 steps. Runtime is reported in seconds as mean $\pm$ population standard deviation over seeds $0, \ldots, 4$.

Setting	Model	HARD	SELECTIVE
Default	Lifted logistic	0.783 ± 0.041	0.786 ± 0.054
Default	Multinomial linear	2.158 ± 0.181	2.230 ± 0.104
Default	MLP	1.195 ± 0.122	1.194 ± 0.094
Default	Additive polynomial	1.764 ± 0.012	1.877 ± 0.121
Default	RBF prototype	0.876 ± 0.019	0.899 ± 0.013
Weak-spectrum	Lifted logistic	1.664 ± 0.032	1.794 ± 0.284
Weak-spectrum	Multinomial linear	4.731 ± 0.053	4.799 ± 0.077
Weak-spectrum	MLP	2.163 ± 0.064	2.198 ± 0.031
Weak-spectrum	Additive polynomial	2.431 ± 0.094	2.454 ± 0.020
Weak-spectrum	RBF prototype	2.567 ± 0.077	2.739 ± 0.191

TTA test accuracy is averaged over four seeds. Selective uses the transition parameter $\tau = 10^{-3}$. The sensitivity experiment in Figure 2(c) uses CIFAR-10/Airbench94 at batch size 256 over three seeds. The spectral diagnostics use the active singular values of the per-sample update matrix and the corresponding relative amplification defined in Section 4.

The real-world experiments are implemented in PyTorch (Paszke et al., 2019). Exact per-sample updates are obtained using automatic differentiation. The experiments were run on single NVIDIA GPUs using either an RTX 5090 with 32 GB of memory or an A100 with 80 GB of memory. Because the current exact implementation requires per-sample update extraction and additional sample-space spectral computation, its runtime depends strongly on the extraction path. We therefore do not treat the present implementation as an optimized systems benchmark; this computational limitation is discussed in Section 6.