RETHINKING SPECTRAL CONSTRAINTS IN MUON: BEYOND NORM PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large neural networks requires optimizers that make effective progress at a practical computational cost. Muon controls the scale of its updates, but does not directly constrain the weights themselves. Spectral constraints address this gap by controlling the scale of weight matrices. However, how these constraints are enforced matters: keeping the norm unchanged does not necessarily preserve the model's ability to learn. We revisit spectral constraints in Muon through their effects on parameter updates and computational cost. We establish conditions for norm preservation and show how certain constraint designs can unnecessarily restrict the changes available to the model. Guided by this analysis, we propose two Muon-based optimizers. Spectral Block Correction (SBC) selectively corrects update blocks while preserving the spectral norm. Regularized Spectral Projection (RSP) shrinks trial weights and caps only singular values exceeding a prescribed bound. Under the reported single-seed configurations, SBC reduces final BERT-Large validation loss on Wikipedia by 14.3% versus AdamW. On Imagenette, SBC and RSP reduce final ViT-L/16 validation cross-entropy by 16.2% and 13.2%, respectively, versus Muon. SBC also reduces the mean cross-width learning-rate transfer penalty by 69.6% relative to Muon. These results highlight that spectral control must account for learning and per-step cost, not norm preservation alone.

1 INTRODUCTION

Machine learning has become a central tool for extracting structure from data and addressing complex scientific and engineering problems (Jordan & Mitchell, 2015). Deep neural networks have driven much of this progress, from perception and language processing to protein structure prediction (LeCun et al., 2015; Jumper et al., 2021). Realizing their potential at scale depends not only on model capacity and training data, but also on the effective use of computational resources (Hoffmann et al., 2022). Optimizer design is central to this effort: it can reduce the computation needed to reach a target performance (Bottou et al., 2018; Liu et al., 2025) and improve the quality and generalization of the learned solutions (Loshchilov & Hutter, 2019). Muon (Jordan et al., 2024) takes a matrix-based approach to this optimization problem, constructing updates by approximately orthogonalizing momentum. Its application to large-scale language model training (Liu et al., 2025) motivates closer examination of how to scale and constrain matrix updates during training.

Maximal-update parameterization (μP) provides a foundation for feature learning at large width (Yang & Hu, 2021) and cross-width hyperparameter transfer through μTransfer (Yang et al., 2021). A spectral formulation connects these ideas to the norms of both weights and their updates (Yang et al., 2023). Muon++ (Zhao, 2026) and the Spectral Sphere Optimizer (SSO) (Xie et al., 2026) translate this perspective into concrete update rules that explicitly control weight spectral norms. Yet norm control alone does not determine which update components remain available or how they change layer outputs. How can spectral scale be controlled without unnecessarily restricting the updates needed for learning?

We analyze norm-preserving rules that exclude update components involving selected leading left and right singular directions, quantifying the tradeoff between admissible step sizes and changes in layer outputs. Under exact norm preservation, repeating these exclusions over all current dominant directions can also lock the initially dominant directions across iterations. These findings motivate two Muon-based optimizers (Figure 1(a)). Spectral Block Correction (SBC) jointly corrects update

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.

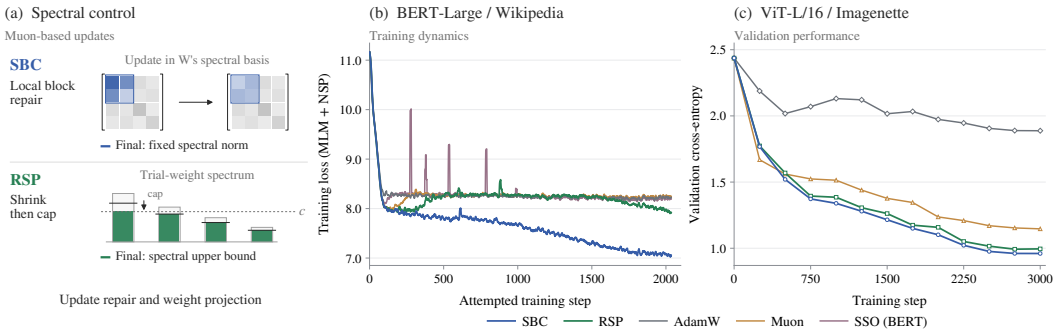

Figure 1: Spectral control in SBC and RSP (a), BERT-Large training loss on Wikipedia (10-step trailing mean; b), and ViT-L/16 validation cross-entropy on Imagenette (c).

components associated with near-maximal weight singular values while maintaining a fixed spectral norm. Regularized Spectral Projection (RSP) shrinks trial weights and caps only singular values above a prescribed bound, allowing norm contraction.

We evaluate SBC and RSP on BERT-Large pretraining on Wikipedia and ViT-L/16 classification on Imagenette (Figure 1(b,c)). Under the reported single-seed configurations, linear interpolation between validation checkpoints estimates that SBC and RSP reach Muon's final ViT validation loss with approximately 41.1% and 32.4% fewer updates, respectively (Table 2). We also evaluate cross-width learning-rate transfer, measuring the transfer penalty as excess validation MLM loss over each optimizer's target-width grid-search minimum (Section 4.3).

Our main contributions are as follows:

- We establish sharp step-size bounds, quantify layer-response tradeoffs, and prove that repeated exact two-sided exclusion of dominant singular directions can cause locking despite norm preservation.
- We introduce two new Muon-based optimizers: SBC selectively corrects update blocks, while RSP combines trial-weight shrinkage with spectral projection, each with explicit constraint guarantees.
- Under reported configurations, SBC reduces BERT validation loss by 14.3% versus AdamW; SBC and RSP reduce ViT validation loss by 16.2% and 13.2% versus Muon.
- From width 256 to 512–2048, SBC achieves the lowest mean learning-rate transfer penalty among six optimizers, reducing Muon's and SSO's by 69.6% and 72.4%, respectively.

Organization. Sections 2, 3, and 4 present related work, methods and analysis, and experiments, respectively. Section 5 concludes with limitations and future directions. Proofs, implementation details, and additional results are provided in the appendices.

2 Related Works

Matrix-Structured Optimization and Muon. Adaptive optimizers such as Adam (Kingma & Ba, 2015) and AdamW (Loshchilov & Hutter, 2019) rescale gradients coordinatewise, whereas Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2025) exploit matrix structure through preconditioning. Muon (Jordan et al., 2024) belongs to a broader line of spectral descent and gradient orthogonalization (Carlson et al., 2015; Tuddenham et al., 2022). It approximately orthogonalizes momentum using Newton–Schulz iterations. Its exact-polar counterpart admits a spectral-norm steepest-descent interpretation (Bernstein & Newhouse, 2024; Kovalev, 2025), while PolarGrad links polar-based updates to a broader preconditioning framework (Lau et al., 2025). Subsequent developments address two complementary aspects of Muon: large-scale training recipes through update-scale calibration and weight decay (Liu et al., 2025), and finite-step orthogonalization through improved polynomial approximations and convergence analysis (Amsel et al., 2026; Kim & Oh, 2026).

μP, Spectral Scaling, and Hyperparameter Transfer. Maximal-update parameterization (μP) (Yang & Hu, 2021) prescribes width-dependent scaling for feature learning and cross-width hyperparameter transfer (Yang et al., 2021). Its scaling rules depend on the optimizer and the resulting feature changes (Chizat & Netrapalli, 2024; Everett et al., 2024; Ishikawa & Karakida, 2024), while weight decay also influences learning-rate transfer in practice (Kosson et al., 2026). The spectral perspective (Yang et al., 2023) motivates controlling the norms of both weights and updates during training. Muon++ (Zhao, 2026) pursues this goal through constrained updates, but its main two-sided projection excludes all components involving the leading singular directions, and its norm-preservation guarantee requires a step size bounded by the gap between the two largest singular values. These limitations motivate our central question: how can we control spectral scale while allowing weights greater freedom to change during training?

Spectrally Constrained Optimizers. Scion (Pethick et al., 2025) and Gluon (Riabinin et al., 2025) study optimization through linear minimization over norm balls. Other geometries impose semi-orthogonality (Bernstein, 2025) or preserve the full singular-value spectrum through orthogonal transformations (Shi et al., 2026). Spectral-sphere formulations instead constrain only the largest singular value (Su, 2025; Xie et al., 2026). When this value is simple, SSO uses a Lagrange-multiplier search to construct a first-order norm-preserving direction, with radial rescaling restoring the target norm. Our analysis shows that repeated exact two-sided protection can lock dominant singular directions across iterations. SBC instead corrects an already formed Muon direction within a step-dependent group of near-boundary singular directions, accounting for their distance to the spectral bound. Unlike SSO's multiplier search, this correction does not require recomputing the matrix-sign direction for each trial multiplier. The step-size budget is checked against the actual parameter displacement after radial rescaling, not just the proposed direction.

Spectral Projection and Regularization. Spectral normalization rescales weights by their spectral norms (Miyato et al., 2018), whereas spectral-norm regularization penalizes large operator norms (Yoshida & Miyato, 2017). Related approaches control input sensitivity through layerwise Lipschitz constraints (Gouk et al., 2021) or approximate Parseval structure (Cisse et al., 2017). Projection and shrinkage have classical foundations in proximal optimization and unitarily invariant matrix functions (Parikh & Boyd, 2014; Lewis, 1995). Recent methods act on different objects: SPECTRA clips optimizer updates and optionally gradients (Jiang et al., 2026), while Musec clips momentum spectra (Liu et al., 2026). Oikonomidis et al. (2026) develop anisotropic proximal spectral methods whose forward preconditioning and backward correction share a reference function. RSP applies a Frobenius-metric correction after forming the Muon proposal: it shrinks the trial weights and then caps only singular values exceeding the spectral bound. Unlike SSO's fixed-radius retraction (Xie et al., 2026), the projection leaves points already inside the ball unchanged and avoids rescaling nonviolating singular values. Shrinkage acts on the full trial matrix, rather than only on the old weights as in decoupled weight decay (Loshchilov & Hutter, 2019). This makes the spectral upper bound and the strength of quadratic regularization separate design choices.

3 Methods and Analysis

We first introduce the common Muon proposal and analyze the costs of hard spectral protection (Sections 3.1–3.2). This analysis motivates SBC and RSP, which apply spectral corrections to update directions and trial weights, respectively (Sections 3.3 and 3.4). Finally, we summarize their scale-control properties and introduce diagnostics for the resulting changes in layer outputs (Section 3.5).

3.1 Spectral Constraints and Muon Updates

We consider $W_t \in \mathbb{R}^{m\times n}$, with $d = \min(m, n)$ and $t = 0, 1, \ldots$. We use $\|\cdot\|_2$ for the vector Euclidean or matrix spectral norm, $\|\cdot\|_F$ for the Frobenius norm, and I_p for the $p \times p$ identity. For $A \in \mathbb{R}^{m\times n}$, its singular values are ordered as $\sigma_1(A) \geq \cdots \geq \sigma_d(A) \geq 0$. Fix $S = \|W_0\|_2 > 0$. SBC uses the sphere $\mathcal{S}_S = \{W : \|W\|_2 = S\}$, whereas RSP uses the ball $\mathcal{B}_c = \{W : \|W\|_2 \leq c\}$, with fixed $c \geq S$; both sets consist of $m \times n$ matrices. These are alternative constraint geometries, not consecutive stages of a single optimizer.

Given the minibatch gradient G_t, learning rate $\eta_t \geq 0$, and momentum coefficient $\beta \in [0, 1)$, both methods start from a Muon-based proposal (Jordan et al., 2024):

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta)G_t, \\ D_t &= \mathcal{N}_5(M_t), \qquad \alpha_t = \eta_t S, \qquad Y_t = W_t - \alpha_t D_t, \end{aligned} \tag{1}$$

where $M_{-1} = 0$ and the direction operator is defined by

$$\begin{aligned} \Psi(X) &= 3.4445X - 4.775XX^\top X + 2.0315(XX^\top)^2 X, \\ \mathcal{N}_5(M) &= \frac{1}{1.203}\Psi^{\circ 5}\left(\frac{M}{\|M\|_F}\right), \qquad M \neq 0, \end{aligned} \tag{2}$$

with $\mathcal{N}_5(0) = 0$ and $\Psi^{\circ 5}$ denoting five compositions. This gives $\|D_t\|_2 \leq 1$ in exact arithmetic (Appendix D.1), but generally differs from the exact polar direction. The requested step α_t bounds the raw proposal; the executed change $\Delta_t = W_{t+1} - W_t$ also includes spectral correction.

3.2 Spectral Protection and Its Costs

We extend Muon++'s leading-direction projection (Zhao, 2026) to examine the cost of protecting several directions. Write $W = W_t$ and $M = M_t$, and take the thin SVD

$$W = U\,\mathrm{diag}(\sigma_i)_{i=1}^d V^\top, \qquad U \in \mathbb{R}^{m\times d}, \quad V \in \mathbb{R}^{n\times d},$$

where $\sigma_i = \sigma_i(W)$ and U, V have orthonormal columns. For $1 \leq k \leq d$, let U_k, V_k contain their first k columns. For any thin SVD of A, define the exact polar direction by

$$\begin{aligned} A &= U_A\,\mathrm{diag}(s_i)V_A^\top, \\ \mathrm{msign}(A) &= U_A\,\mathrm{diag}(\mathbf{1}_{\{s_i>0\}})V_A^\top, \end{aligned}$$

where $\mathbf{1}$ denotes the indicator and $\mathrm{msign}(0) = 0$. For $1 \leq k < d$, hard protection uses

$$B_k = (I_m - U_kU_k^\top)M(I_n - V_kV_k^\top), \qquad R_k = \mathrm{msign}(B_k). \tag{3}$$

The resulting direction satisfies $U_k^\top R_k = 0$, $R_kV_k = 0$, and $\|R_k\|_2 \leq 1$, removing updates involving either protected subspace.

Proposition 1 (Spectral safety of hard protection)**.** *For $1 \leq k < d$, let $g_k = \sigma_1(W) - \sigma_{k+1}(W)$. For every $X \in \mathbb{R}^{m\times n}$ satisfying $U_k^\top X = 0$, $XV_k = 0$, and $\|X\|_2 \leq 1$,*

$$0 \leq a \leq g_k \quad \Longrightarrow \quad \|W - aX\|_2 = \|W\|_2. \tag{4}$$

For fixed W, k and $a \geq 0$, $a \leq g_k$ is also necessary for $\|W - aX\|_2 \leq \|W\|_2$ to hold uniformly over this class.

The protected block stays fixed; the remaining block has norm at most $\sigma_{k+1}(W) + a$. Increasing k can enlarge the guaranteed range by moving the gap beyond the first two singular values, without eliminating gap dependence (Appendix A.1). Uniform necessity concerns the entire direction class; a particular direction can admit a larger safe step.

This gain competes with the loss of update components. For the linear map $h \mapsto Wh$, a fixed parameter change $A \in \mathbb{R}^{m\times n}$ changes its response by Ah at the same input. For $h \in \mathbb{R}^n$ with finite second moment $\Sigma_h = \mathbb{E}[hh^\top]$, define

$$\mathcal{E}(A) := \frac{1}{m}\mathbb{E}\|Ah\|_2^2 = \frac{1}{m}\|A\Sigma_h^{1/2}\|_F^2. \tag{5}$$

This input-response energy is the mean squared linear output change per coordinate under a fixed input distribution, not the change in the full network output. For $\Sigma_h = qI_n$, $q > 0$, requested step $\alpha \geq 0$, and $r_k = \mathrm{rank}(B_k)$,

$$a_k = \min\{\alpha, g_k\}, \qquad \mathcal{E}(-a_kR_k) = q\,\frac{a_k^2 r_k}{m}. \tag{6}$$

A larger k can increase a_k while reducing r_k. This rank identity uses exact polar directions; Equation (5) remains the appropriate measure for general inputs and finite Newton–Schulz updates (Appendix A.2).

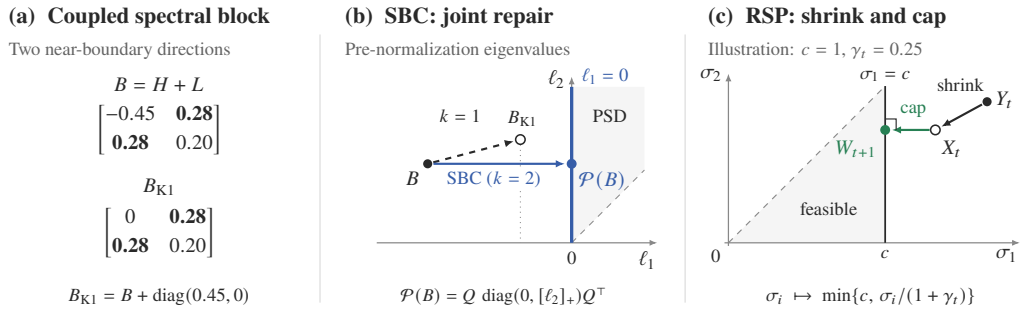

Figure 2: SBC jointly corrects a coupled spectral block (a,b), whereas RSP shrinks and caps the trial-weight spectrum (c).

Proposition 2 (Singular-direction locking). *Let $U_t^\star, V_t^\star$ be orthonormal bases of the full left and right singular subspaces for $\sigma_1(W_t)$. Suppose, for all t, that*

$$\|W_t\|_2 = S, \qquad (U_t^\star)^\top \Delta_t = 0, \qquad \Delta_t V_t^\star = 0.$$

Then every initial unit singular pair (u_0, v_0) associated with S satisfies

$$W_t v_0 = S u_0, \qquad W_t^\top u_0 = S v_0 \qquad \text{for all } t. \tag{7}$$

An initially dominant pair remains dominant and is protected again at the next iteration. Recomputing the protected subspaces therefore need not release it. Appendix A.3 proves this invariant and gives targets with nonzero residual lower bounds. The conclusion concerns this two-sided restriction, not spectral constraints in general. For approximate subspaces, Appendix A.4 bounds spectral error by step-size excess and subspace leakage. Together, these results motivate replacing hard exclusion with selective correction and reconsidering exact norm preservation as the constraint itself.

3.3 Spectral Block Correction (SBC)

At one iteration, write $W = W_t \in \mathcal{S}_S$, $D = D_t$, and $\alpha = \alpha_t$. SBC returns W if $\alpha = 0$ or $M_t = 0$. Otherwise, start with $a = \alpha$ and select

$$k(a) = \max\left\{1, \#\{i \in \{1, \dots, d\} : \sigma_i(W) \geq S - 2a\}\right\}, \tag{8}$$

where $\#$ denotes cardinality. The boundary neighborhood has width $2a$, so its size tracks the trial step rather than a fixed direction count. Clustered leading singular values can therefore be considered together, without requiring the associated update components to vanish.

Using the SVD notation above and $k = k(a)$, form

$$\begin{aligned} H &= \operatorname{sym}(U_k^\top D V_k), \quad & L &= \operatorname{diag}\left(\frac{S - \sigma_i(W)}{a}\right)_{i=1}^{k}, \\ B &= H + L, & K &= \mathcal{P}(B) - B, \\ \widehat{D}(a) &= D + U_k K V_k^\top, \end{aligned} \tag{9}$$

where $\operatorname{sym}(A) = (A + A^\top)/2$ for square A, and L is the spectral slack divided by the trial step. It accounts for how far each selected singular value already lies below the boundary. Choose an orthogonal eigendecomposition

$$B = Q \operatorname{diag}(\ell_1, \dots, \ell_k) Q^\top, \qquad \ell_1 \leq \cdots \leq \ell_k,$$

and define

$$\mathcal{P}(B) = Q \operatorname{diag}(\widetilde{\ell}_i)_{i=1}^k Q^\top, \qquad \widetilde{\ell}_i = \begin{cases} 0, & i = 1, \\ \max\{\ell_i, 0\}, & 2 \leq i \leq k. \end{cases} \tag{10}$$

This selects a Frobenius-nearest positive-semidefinite singular matrix; the chosen eigenbasis specifies the output when the nearest point is nonunique. Unlike hard protection, this operation corrects

the selected block without zeroing whole rows or columns in spectral coordinates. Figure 2(a,b) compares the ordered block eigenvalues before direction normalization and radial retraction. In this example, the single-coordinate comparator $B_{\mathrm{K1}} = B + \mathrm{diag}(0.45, 0)$ leaves a negative eigenvalue, whereas joint correction gives the positive-semidefinite singular block $B + K = \mathcal{P}(B)$. The figure uses $[x]_+ = \max\{x, 0\}$. The local symmetric-block interpretation, including its limits as a spectral-norm proxy, is derived in Appendices B.1 and B.2.

The normalized direction and radial candidate are

$$R(a) = \frac{\widehat{D}(a)}{\max\{1, \|\widehat{D}(a)\|_2\}}, \qquad Z(a) = S\frac{W - aR(a)}{\|W - aR(a)\|_2}. \tag{11}$$

For a nonzero trial matrix, accept $W_{t+1} = Z(a)$ only if $\|Z(a) - W\|_2 \le \alpha$. The check follows rescaling because direction normalization alone does not guarantee that the final displacement remains within the requested budget α. Otherwise reduce a and recompute k, L, K, keeping D and α fixed. After twelve unsuccessful trials, retain W. Zero or nonfinite trial matrices are treated as numerical failures, not accepted updates. The reduction rule, numerical failures, and checks on stored weights are specified in Appendix B.4.

3.4 Regularized Spectral Projection (RSP)

Set $c = \kappa S$, with $\kappa \ge 1$, and $F_0 = \|W_0\|_F$. RSP uses $\gamma_t = \lambda\alpha_t\sqrt{d}/F_0$, where $\lambda \ge 0$ controls shrinkage; the default values are $\lambda = 0.05$ and $\kappa = 1.1$. Given the common proposal Y_t, RSP solves

$$W_{t+1} = \arg\min_{Z \in \mathcal{B}_c} \left\{ \frac{1}{2}\|Z - Y_t\|_F^2 + \frac{\gamma_t}{2}\|Z\|_F^2 \right\} \tag{12}$$

$$= \Pi_c(X_t), \qquad X_t = \frac{Y_t}{1 + \gamma_t}. \tag{13}$$

For a thin SVD of X, write

$$X = U_X\,\mathrm{diag}(s_i)V_X^\top,$$
$$\Pi_c(X) = U_X\,\mathrm{diag}(\min\{s_i, c\})V_X^\top.$$

Here $\Pi_c(X)$ is the Frobenius-nearest projection onto $\mathcal{B}_c$ (Parikh & Boyd, 2014; Lewis, 1995). Figure 2(c) illustrates this update in singular-value coordinates: shrinkage moves Y_t to X_t, after which projection caps the first singular value at c while leaving the second unchanged in this example. The distance term keeps the result close to the proposal, while the quadratic penalty controls overall shrinkage. This is a local correction, not exact minimization of the regularized training loss. The projection neither expands interior points to the boundary nor rescales nonviolating singular values; shrinkage itself can still change those values. With $M_t = 0$, RSP still shrinks the weights when $\gamma_t > 0$. The local proximal derivation and initialization-based calibration are given in Appendices C.1 and C.3.

To isolate projection from shrinkage, compare it at the same X_t and c with radial capping: $T_c(X) = X\min\{1, c/\|X\|_2\}$ for $X \neq 0$, and $T_c(0) = 0$. Componentwise comparison of the singular-value corrections gives

$$\|(\Pi_c(X_t) - X_t)h\|_2 \le \|(T_c(X_t) - X_t)h\|_2, \qquad h \in \mathbb{R}^n. \tag{14}$$

This compares the additional repair at a matched trial point, not total updates or training losses (Appendix C.4).

3.5 Spectral Scaling and Update Response

Scale control. For feasible weights in exact arithmetic,

$$\begin{aligned} \text{SBC}: &\quad \frac{\|W_{t+1}\|_2}{S} = 1, \quad \frac{\|\Delta_t\|_2}{S} \le \eta_t, \\ \text{RSP}: &\quad \frac{\|W_{t+1}\|_2}{S} \le \kappa, \quad \frac{\|\Delta_t\|_2}{S} \le 2\eta_t + \frac{\kappa\gamma_t}{1 + \gamma_t}. \end{aligned} \tag{15}$$

Across widths, S and F_0 are computed from each initialization under the protocol in Appendix E.3. Writing the initial stable rank explicitly gives

$$\mathrm{sr}(W_0) = \frac{F_0^2}{S^2}, \qquad \gamma_t = \lambda\eta_t\sqrt{\frac{d}{\mathrm{sr}(W_0)}}.$$

This introduces no extra width factor when $\mathrm{sr}(W_0)/d$ stays bounded away from zero. These bounds are relative to the chosen initialization, rather than a separate initialization prescription. They do not guarantee nonvanishing responses or optimal learning-rate transfer (Appendix D.2).

Input response. For $N \geq 1$ fixed inputs, collect the input vectors and write the empirical counterpart of Equation (5) as

$$\mathsf{H} = [h_1, \ldots, h_N] \in \mathbb{R}^{n\times N}, \qquad \mathcal{E}_{\mathsf{H}}(A) = \frac{\|A\mathsf{H}\|_F^2}{mN}.$$

We compare the raw change $-\alpha_t D_t$, the intermediate change $-aR(a)$ for an accepted SBC step or $X_t - W_t$ for RSP, and the executed change Δ_t. Local comparisons fix weights, momentum, and inputs; isolating SBC's block correction also matches a. Trajectory-level statistics are analyzed separately, and greater response energy is not identified with better optimization. Relative response error, cosine similarity, and degenerate cases are defined in Appendix D.3.

4 Experiments

4.1 Experimental Setup

We train BERT-Large (336M parameters) (Devlin et al., 2019) from scratch on English Wikipedia, and ViT-L/16 (303M) (Dosovitskiy et al., 2021) on an Imagenette subset (Howard, 2019) with 8,000 training and 2,000 validation images. BERT uses 2,035 attempted batches of 192 length-512 sequences with masked language modeling (MLM) and next-sentence prediction (NSP); ViT uses 3,000 successful updates with batch size 64. We compare against AdamW, Muon, and SSO in single-seed runs. Muon uses controlled NS5 on BERT and native Muon on ViT; SSO uses the common-initialization-radius adaptation.[1] Additional settings and evaluation protocols are provided in Appendix E.

4.2 Training Performance

Table 1 reports final validation results for the matrix-based optimizers. On BERT, SBC and RSP reduce total loss by 14.3% and 3.7% relative to AdamW's 8.2401; SBC's MLM loss alone falls by 10.4% from AdamW's 7.5804. On ViT, they reduce cross-entropy (CE) by 16.2% and 13.2% relative to Muon, with top-1 gains of 5.60 and 4.00 percentage points. SBC has the lowest observed loss in both tasks; SSO's final ViT metrics lie between SBC and RSP. Figure 3 complements Figure 1 with ViT training and BERT validation trajectories.

Table 1: Final validation results for matrix-based optimizers under the reported single-seed configurations.

	BERT-Large		ViT-L/16	
Method	MLM + NSP ↓	MLM ↓	CE ↓	Top-1 (%) ↑
Muon	9.1269	8.0909	1.1466	68.05
SSO	9.3736	8.3065	0.9739	73.35
RSP	7.9386	7.4010	0.9953	72.05
SBC	**7.0622**	**6.7915**	**0.9604**	**73.65**

[1] BERT Muon has 644 overflow skips; SBC's BERT initialization hash differs, so identical initial weights are not assumed. ViT uses shared initialization without overflow; its AdamW learning rate is fixed rather than exhaustively tuned.

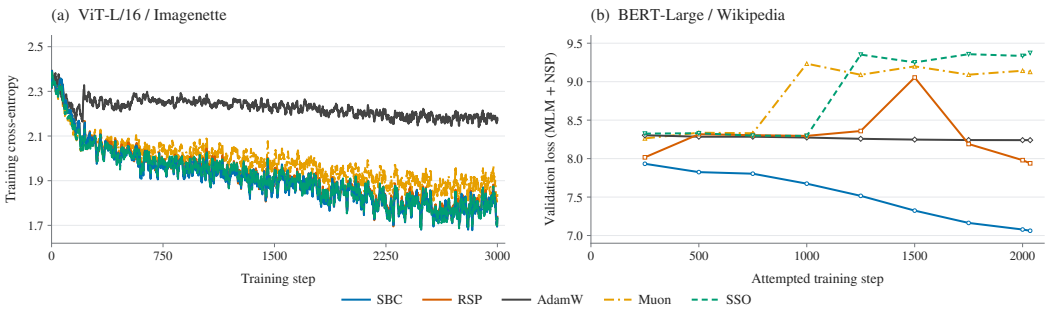

Figure 3: ViT-L/16 training cross-entropy on Imagenette (10-step trailing mean; a) and BERT-Large validation loss on Wikipedia (MLM + NSP; b).

Interpolating between recorded validation checkpoints, SBC and RSP reach Muon's final ViT loss with approximately 41.1% and 32.4% fewer updates (Table 2). These are post-hoc crossing estimates, not evaluations at those exact steps.

Table 2: Estimated ViT updates to reach validation CE targets, using linear interpolation.

	Target validation CE				
Method	1.30	1.20	$L_\star$	1.00	Fewer updates at $L_\star$ (%)
Muon	1855	2308	3000	–	0.0
SSO	1252	1633	1906	2479	36.5
RSP	1287	1677	2027	2672	32.4
SBC	**1165**	**1559**	**1768**	**2370**	**41.1**

$L_\star \approx 1.146578$ is Muon's final CE at 3,000 updates. Crossings are rounded; reductions use unrounded estimates. "–" denotes an unreached target; AdamW reaches none of these targets within 3,000 updates.

Among the directly measured single-V100 acceptance runs, RSP's per-step cost is close to SSO's. The SBC value in Table 3 is an estimate, not a direct V100 measurement.[2]

Table 3: BERT training-step times. Muon, SSO, and RSP are measured averages over 200 acceptance steps on one V100 GPU. SBC is an estimate, not a direct V100 measurement.

Method	Muon	SSO	RSP	SBC (est.)
Time (s/step) ↓	**25.19**	67.79	67.01	119.60

4.3 Learning-Rate Transfer Across Widths

For 24-layer BERT at widths 256, 512, 1024, and 2048, we select each optimizer's learning rate by minimum validation MLM loss at width 256 and reuse it unchanged at larger widths. Each included run completes 200 successful updates. The transfer penalty subtracts the same optimizer's target-width grid-search minimum from its transferred validation MLM loss; the mean equally weights the three targets. NR adds radial restoration and displacement checking to the common proposal without block correction.

Table 4 summarizes mean penalties for SBC, RSP, and four baselines. SBC's mean penalty of 0.0311 is 69.6% below Muon-NS5 and 72.4% below SSO. RSP's mean penalty is 0.0703, also

[2]For fixed matrix sizes and bounded correction trials, more tokens per optimizer update amortize the relative arithmetic overhead (Jordan et al., 2024); see Appendix D.4. This is not a measured time saving.

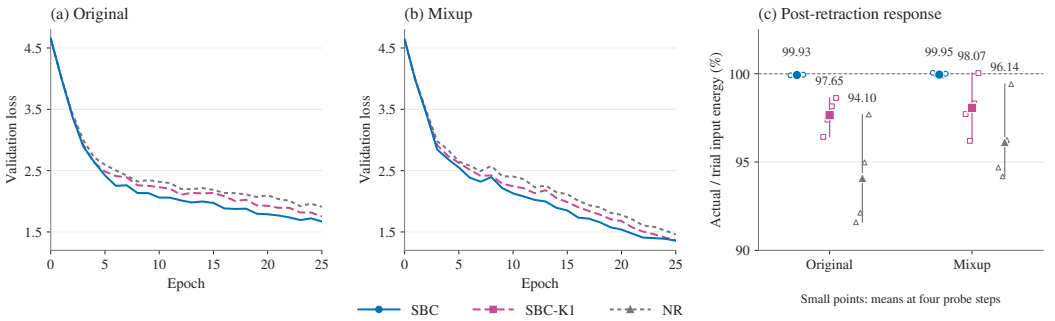

Figure 4: SBC ablations over the first 25 epochs without Mixup (a) and with Mixup (b), and post-retraction input-response energy relative to the corrected trial update (c).

below both baselines. Search ranges and grid sizes differ by method; sweeps, absolute losses, and source-selection sensitivity are reported in Appendix F.1.

Table 4: Mean validation MLM penalty for learning-rate transfer from width 256 to widths 512, 1024, and 2048.

Method	AdamW	Muon-NS5	SSO	NR	RSP	SBC
Mean penalty ↓	0.2970	0.1024	0.1127	0.0659	0.0703	**0.0311**

4.4 Mechanism Analysis and Ablations

On a six-layer, width-192 ViT for CIFAR-100 (Krizhevsky, 2009), we compare SBC with SBC-K1 (fixed block size one) and NR under matched settings within each condition. Figure 4(a,b) shows the recorded first 25 epochs in each condition. The conditions and display window were selected exploratorily. SBC has lower validation loss than both ablations at epoch 25; this comparison concerns early training rather than terminal performance (Appendix F.2).

Panel (c) reports mean post-retraction to corrected-trial input-response energy ratios (Section 3.5). SBC retains 99.93% without Mixup and 99.95% with Mixup, showing less attenuation during retraction than either ablation. These are trajectory-level measurements, not matched-snapshot comparisons or retention relative to the raw proposal; probe details are in Appendix F.3.

5 Conclusion and Future Work

We revisit spectral control in Muon beyond norm preservation, characterizing the step-size–response tradeoff and singular-direction locking under hard two-sided update restrictions. These findings motivate SBC's selective update correction and RSP's regularized projection of trial weights. Under the reported single-seed configurations, both methods lower final validation loss relative to AdamW on BERT and Muon on ViT, and reach Muon's final ViT validation loss with fewer updates. SBC also reduces cross-width learning-rate transfer penalties. Together, these findings emphasize two distinct design choices: which spectral constraints to impose and how to enforce them without unnecessarily restricting learning.

The main limitation is the higher per-update computational cost than AdamW and Muon. Future work will explore lower-cost algorithmic variants and more efficient implementations of SBC and RSP, aiming to retain their validation gains and reliable spectral control. Directions include approximate decompositions, reuse of spectral information across steps, and adaptive correction frequency. A key question is how approximation errors and less frequent corrections affect constraint satisfaction and optimization progress. Longer, multi-seed evaluations on larger models and across more diverse architectures and tasks, with broader hyperparameter tuning and matched computational budgets, are needed to assess robustness and whether update-efficiency gains translate into lower time to target quality.

AI USE STATEMENT

We used generative AI tools to polish the language and improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits for accuracy and consistency with the intended meaning. The authors take full responsibility for the final content of this work, including its text, claims, and results.

ETHICS STATEMENT

This work focuses on general-purpose optimization methods for neural network training. We do not identify specific ethical concerns arising from the proposed methods beyond those generally associated with the development and use of machine learning systems. The ethical implications of downstream applications should be assessed in their respective deployment contexts.

REPRODUCIBILITY STATEMENT

The proposed methods and their update rules are described in the main text. The assumptions and proofs underlying the theoretical results are provided in the main text and appendix. The experimental section and appendix describe the datasets, model architectures, optimizer configurations, and training and evaluation protocols used in our experiments.

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 *International Conference on Learning Representations*, 2026.

Jeremy Bernstein. Modular manifolds. Thinking Machines Lab technical blog, September 2025. URL `https://thinkingmachines.ai/blog/modular-manifolds/`.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. arXiv preprint arXiv:2409.20325, 2024.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 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.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Katie E. Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12666–12700. PMLR, 2024.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Jeremy Howard. Imagenette. Dataset repository, 2019. URL `https://github.com/fastai/imagenette`.

Satoki Ishikawa and Ryo Karakida. On the parameterization of second-order optimization effective towards the infinite width. In *International Conference on Learning Representations*, 2024.

Xiaowen Jiang, Andrei Semenov, and Sebastian U. Stich. Enhancing LLM training via spectral clipping. In *International Conference on Machine Learning*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Technical blog post, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton-Schulz. In *International Conference on Learning Representations*, 2026.

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

Atli Kosson, Jeremy Welborn, Yang Liu, Martin Jaggi, and Xi Chen. Weight decay may matter more than μP for learning rate transfer in practice. In *International Conference on Learning Representations*, 2026.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. arXiv preprint arXiv:2505.21799, 2025.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1–2):173–183, 1995.

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

Zhuanghua Liu, Menglian Wang, and Luo Luo. Musec: MomentUm SpEctral Clipping for stable Muon-type training. arXiv preprint arXiv:2609.11655, 2026.

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

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Konstantinos Oikonomidis, Jan Quan, Kimon Antonakopoulos, Antonio Silveti-Falls, Volkan Cevher, and Panagiotis Patrinos. Constrained stochastic spectral preconditioning converges for nonconvex objectives. arXiv preprint arXiv:2605.11850, 2026.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, 2025.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs). arXiv preprint arXiv:2505.13416, 2025.

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

G. W. Stewart and J. G. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

Jianlin Su. Steepest descent on manifolds: 4. Muon + spectral sphere. Scientific Spaces technical blog, August 2025. URL `https://kexue.fm/archives/11241`. In Chinese; title translated into English.

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

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

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

Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, Chong Luo, and Baining Guo. Controlled LLM training on spectral sphere. In *International Conference on Machine Learning*, 2026.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning. arXiv preprint arXiv:2310.17813, 2023.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10941, 2017.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

John Zhao. Towards a principled Muon under μP: Ensuring spectral conditions throughout training. arXiv preprint arXiv:2601.01306, 2026.

NOTATION

The table collects notation from the main text and appendices. Local abbreviations suppress iteration indices where stated. Reused letters are distinguished by their section-specific meanings; in particular, H and H denote different matrices.

Table 5: Notation summary.

Symbol	Meaning
Basic notation and matrix operations	
m, n, d	Output and input dimensions of a matrix parameter $W \in \mathbb{R}^{m\times n}$, with $d = \min(m, n)$.
t; i, j	Optimizer-iteration index t; local component, summation, or iteration indices i, j, with ranges specified where used.
$\mathbb{R}^{m\times n}$	Space of real $m \times n$ matrices.
I_p; 0	The $p \times p$ identity matrix; a zero scalar, vector, or matrix of the appropriate size.
$A^\top$	Transpose of a matrix A.
$\|\cdot\|_2$	Euclidean norm for vectors and spectral norm (largest singular value) for matrices.
$\|\cdot\|_F$	Frobenius norm.
$\langle A, B\rangle_F$	Frobenius inner product, $\mathrm{tr}(A^\top B)$, for matrices of the same shape.
$\mathrm{tr}(A)$; $\mathrm{rank}(A)$	Trace of a square matrix; matrix rank.
$\mathrm{diag}(a_i)$	Diagonal matrix with the listed entries a_i.
$\mathrm{sym}(A)$	Symmetric part of a square matrix: $(A + A^\top)/2$.
$A \preceq B$; $A \succeq 0$	Loewner order: $B - A$ is positive semidefinite; A is positive semidefinite.
$\lambda_{\min}(A)$	Smallest eigenvalue of a real symmetric matrix.
$\mathbb{E}$; $[x]_+$	Expectation; positive part $[x]_+ = \max\{x, 0\}$.
$\mathbf{1}_{\{\cdot\}}$	Indicator of the stated condition.
$\#$; $\|\cdot\|$	Cardinality of a finite set; absolute value for a scalar or cardinality for a finite set, according to the argument.
Weights, constraint geometry, and the common proposal	
W_t; W_0; W	Matrix parameter at iteration t; its initialization; the current matrix with the time index suppressed.
G_t; M_t, M	Minibatch gradient; momentum matrix $M_t = \beta M_{t-1} + (1-\beta)G_t$, initialized with $M_{-1} = 0$. The local notation is $M = M_t$.
β; η_t, η	Momentum coefficient in $[0, 1)$; nonnegative learning rate, with η also used for a nominal matrix learning rate in experiments.
S; F_0	Fixed initialization scales $S = \|W_0\|_2 > 0$ and $F_0 = \|W_0\|_F$.
$\mathcal{S}_S$	Spectral sphere $\{W \in \mathbb{R}^{m\times n} : \|W\|_2 = S\}$ used by SBC.
$\mathcal{B}_c$	Spectral-norm ball $\{W \in \mathbb{R}^{m\times n} : \|W\|_2 \le c\}$ used by RSP.
c; κ	Fixed spectral ceiling $c = \kappa S$, with ceiling multiplier $\kappa \ge 1$ (default 1.1).
D_t, D	Common direction $D_t = \mathcal{N}_5(M_t)$, with $\|D_t\|_2 \le 1$ in exact arithmetic; $D = D_t$ locally.
α_t, α	Requested step scale $\alpha_t = \eta_t S$, bounding the raw proposal displacement; $\alpha = \alpha_t$ locally.
Y_t; Δ_t	Raw trial weights $Y_t = W_t - \alpha_t D_t$; executed parameter change $\Delta_t = W_{t+1} - W_t$ after all corrections.
W^+	Post-update weights in local one-step statements. In execution logs, W^+ denotes the stored post-update matrix.
Spectral decompositions and hard protection	
$\sigma_i(A)$; σ_i	Singular values in nonincreasing order; $\sigma_i = \sigma_i(W)$ when the matrix argument is omitted.
U, V; u_i, v_i	Paired left and right singular-vector matrices in a thin SVD of W, and their columns. The same SVD notation is used locally for other matrix arguments.
U_A, V_A; U_X, V_X; s_i	Thin singular-vector matrices for the indicated argument A or X; its singular values s_i. In the Newton–Schulz analysis, s_i also denote singular values of M.
U_k, V_k; Σ_k	First k columns of U, V; $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$.
$U_\perp, V_\perp$; $\overline{U}, \overline{V}$	Orthonormal complements of U_k, V_k; square orthogonal completions $\overline{U} = [U_k, U_\perp]$, $\overline{V} = [V_k, V_\perp]$.
$W_\perp, X_\perp$	Complementary coordinate blocks $U_\perp^\top W V_\perp$ and $U_\perp^\top X V_\perp$ in the hard-protection proof.
P_U, P_V	Projectors onto the selected singular subspaces: $P_U = U_k U_k^\top$ and $P_V = V_k V_k^\top$.

Continued on next page

Symbol	Meaning
k; $\mathcal{X}_k(W)$	Number of protected directions; the class of $X \in \mathbb{R}^{m\times n}$ satisfying $U_k^\top X = 0$, $XV_k = 0$, and $\|X\|_2 \le 1$. The spectral-gap results use $1 \le k < d$.
B_k	Two-sided protected momentum $(I_m - P_U)M(I_n - P_V)$.
$\widetilde{U}, \widetilde{V}, \widetilde{\Sigma}$	Compact SVD factors $B_k = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$, retaining only positive singular values.
$\mathrm{msign}(A)$	Exact polar direction $U_A \,\mathrm{diag}(\mathbf{1}_{\{s_i>0\}})V_A^\top$, with $\mathrm{msign}(0) = 0$; zero singular directions are not completed.
R_k; R_0	Protected exact polar direction $R_k = \mathrm{msign}(B_k)$; unprotected reference direction $R_0 = \mathrm{msign}(M)$.
r_k; g_k	Protected-momentum rank $r_k = \mathrm{rank}(B_k)$; spectral gap $g_k = \sigma_1(W) - \sigma_{k+1}(W)$.
a; a_k	Nonnegative step in the hard-protection analysis; safely clipped step $a_k = \min\{\alpha, g_k\}$.
$U_t^\star, V_t^\star$; u_0, v_0	Bases of the full current leading singular subspaces; an initial unit singular pair associated with S in the locking result.
T; $f(W)$	Fixed target matrix in the locking example; fitting objective $f(W) = \frac{1}{2}\|W - T\|_F^2$.
b_L, b_R	Locked-pair residuals $b_L = Su_0 - Tv_0$ and $b_R = Sv_0 - T^\top u_0$.
Approximate protection and spectral-error bounds	
$\widehat{U}_k, \widehat{V}_k$	Orthonormal estimates of the selected left and right singular subspaces.
$\widehat{P}_U, \widehat{P}_V$	Estimated projectors $\widehat{U}_k\widehat{U}_k^\top$ and $\widehat{V}_k\widehat{V}_k^\top$.
$\varepsilon_U, \varepsilon_V$	Two-sided leakage measures $\|U_k^\top X\|_2$ and $\|XV_k\|_2$ for a direction X with $\|X\|_2 \le 1$.
X_0; E	Exactly protected component $X_0 = (I_m - P_U)X(I_n - P_V)$ and leakage residual $E = X - X_0$. Elsewhere, E may denote a generic matrix perturbation.
s; Y, Z	Current norm $s = \|W\|_2 > 0$, trial matrix $Y = W - aX$, and radially restored matrix $Z = sY/\|Y\|_2$ in the approximate-protection analysis.
$b(a)$; ϵ	Spectral-error bound $b(a) = [a - g_k]_+ + a(\varepsilon_U + \varepsilon_V)$; relative tolerance $0 \le \epsilon < 1$ in $b(a) \le \epsilon s$.
Spectral Block Correction (SBC)	
a; $k(a)$	SBC trial step, initially $a = \alpha$; selected block size $k(a) = \max\{1, \#\{i : \sigma_i(W) \ge S - 2a\}\}$, with $1 \le k(a) \le d$.
H; L; B	Symmetric direction block $H = \mathrm{sym}(U_k^\top DV_k)$; scaled spectral slack $L = \mathrm{diag}((S - \sigma_i(W))/a)_{i=1}^k$; combined block $B = H + L$.
$\mathbb{S}^k$; $\mathcal{C}_0^k$	Real symmetric $k \times k$ matrices; the subset $\mathcal{C}_0^k = \{Z \in \mathbb{S}^k : Z \succeq 0, \lambda_{\min}(Z) = 0\}$ of positive-semidefinite singular matrices.
Q; ℓ_i	Orthogonal eigenbasis and ordered eigenvalues in $B = Q\,\mathrm{diag}(\ell_i)Q^\top$, where $\ell_1 \le \cdots \le \ell_k$.
$\widetilde{\ell}_i$; z_i	Corrected eigenvalues $\widetilde{\ell}_1 = 0$ and $\widetilde{\ell}_i = [\ell_i]_+$ for $i \ge 2$; z_i denote eigenvalues $0 = z_1 \le \cdots \le z_k$ of a feasible $Z \in \mathcal{C}_0^k$ in the projection proof.
$\mathcal{P}(B)$; K	A selected Frobenius-nearest point to B in $\mathcal{C}_0^k$; correction $K = \mathcal{P}(B) - B$. The selected eigenbasis specifies the output when the projection is nonunique.
$\widehat{D}(a)$	Block-corrected direction before normalization: $D + U_k K V_k^\top$.
$R(a)$; Y_a	Norm-bounded direction $R(a) = \widehat{D}(a)/\max\{1, \|\widehat{D}(a)\|_2\}$; corrected trial weights $Y_a = W - aR(a)$.
$Z(a)$	Radially restored candidate $SY_a/\|Y_a\|_2$, defined for $Y_a \neq 0$ and accepted only if $\|Z(a) - W\|_2 \le \alpha$ in exact arithmetic.
B_{K1}	Single-coordinate comparator $B + \mathrm{diag}(0.45, 0)$ in the two-dimensional block illustration.
b; R	Local example quantities: $0 < b \le 1$ is the coefficient of the skew-symmetric direction in the block-proxy counterexample; R is a generic normalized direction in the radial-restoration counterexample.
J; j, a_j	SBC candidate budget $J = 12$; candidate index $j = 0, \ldots, J-1$ and its trial step a_j, with $a_0 = \alpha$.
ρ_j	Candidate displacement ratio $\|Z(a_j) - W\|_2/\alpha$ used by the backtracking rule; evaluated numerically in the implementation.
Finite-precision quantities	
ϵ_g, ϵ_s	Machine epsilon values of the geometry and output-storage types.
τ_{step}	Displacement-check tolerance $\max\{10^{-8}, 8\epsilon_g\}$.

Continued on next page

Symbol	Meaning
τ_{rad}	Additional radial-restoration threshold $\max\{2\times10^{-12}, 8\epsilon_g\}$.
τ_{stored}	Stored-radius tolerance $\max\{\tau_{\mathrm{rad}}, 8\epsilon_s\}$.
$\widetilde{Z}$; $\widehat{\lVert\cdot\rVert}_2$	SBC candidate rounded to the output-storage type before acceptance; numerically evaluated spectral norm in the acceptance checks.
$\widetilde{W}^+$; E_{num}	Stored RSP output and its numerical error relative to the ideal update, $\widetilde{W}^+ = W^+ + E_{\mathrm{num}}$.
ϵ (AdamW)	Numerical stabilization parameter 10^{-8} in the reported AdamW configurations; distinct from the relative tolerance in the protection analysis.
Regularized Spectral Projection (RSP)	
λ; γ_t, γ	Nonnegative shrinkage strength λ (default 0.05); quadratic-penalty coefficient $\gamma_t = \lambda\alpha_t\sqrt{d}/F_0$, with γ its local version.
a; Y; X_t, X	In local RSP statements, $a = \alpha_t$ is the requested step, not a backtracking variable; $Y = W - aD$ and $X = Y/(1+\gamma)$, corresponding to $X_t = Y_t/(1+\gamma_t)$.
$\Pi_c(X)$	Frobenius-nearest projection onto $\mathcal{B}_c$: $U_X\,\mathrm{diag}(\min\{s_i, c\})V_X^\top$. Only singular values above c are capped.
$T_c(X)$	Radial-capping comparator $X\min\{1, c/\lVert X\rVert_2\}$ for $X \neq 0$, with $T_c(0) = 0$; interior points are unchanged.
z_i; V; P_i	Local projection-proof quantities: z_i are the nonincreasing singular values of a feasible matrix Z; $V \in \mathcal{B}_c$ is a test point; $P_i = \Pi_c(X_i)$.
P; θ; y	Auxiliary quantities in the displacement proof: $P = \Pi_c(Y)$, shrinkage factor $\theta = (1+\gamma)^{-1}$, and a generic nonnegative scalar y.
e_i; p_i, r_i	Projection residual magnitude $e_i = [s_i - c]_+$; in the matched comparison, $p_i = [s_i - c]_+$ and $r_i = (1 - c/s_1)s_i$ when $s_1 > c$.
E_P, E_T	Additional repairs at the same input matrix: $E_P = \Pi_c(X) - X$ and $E_T = T_c(X) - X$.
ω; $\widetilde{a}$	Shrinkage calibration $\omega = \lambda\sqrt{d}/F_0$, so $\gamma = a\omega$; equivalent rescaled direction step $\widetilde{a} = a/(1+a\omega)$.
τ	Uniform singular-value subtraction level $\tau = \lambda a$ used only in the shrinkage-calibration comparison.
Finite-step Newton–Schulz direction and proof quantities	
$\Psi(X)$	Matrix polynomial $3.4445X - 4.775XX^\top X + 2.0315(XX^\top)^2X$.
X_j; C	Newton–Schulz iterates $X_0 = M/\lVert M\rVert_F$, $X_{j+1} = \Psi(X_j)$ for $j = 0,\dots,4$; final scale constant $C = 1.203$.
$\mathcal{N}_5(M)$	Common five-step direction $X_5/C = \Psi^{\circ5}(M/\lVert M\rVert_F)/C$ for $M \neq 0$, and zero for $M = 0$. It generally differs from $\mathrm{msign}(M)$.
$p(s)$; $p^{\circ5}$	Scalar polynomial $p(s) = (6889s - 9550s^3 + 4063s^5)/2000$ with scalar argument s; its fivefold composition. The superscript $\circ5$ has the same meaning for Ψ.
$p'(s)$; r_-, r_+	Derivative of p; its two positive roots, with $r_\pm^2 = (28650 \pm \sqrt{261022360})/40630$.
l, u; z	Rational bounds $l = 554528/10^6 < r_- < u = 554530/10^6$; scalar argument of $4063z^2 - 9550z + 6889$ in the positivity proof.
f_i	Singular values of $\mathcal{N}_5(M)$, possibly reordered: $f_i = p^{\circ5}(s_i/\lVert M\rVert_F)/C$ for $M \neq 0$.
L_Ψ	Frobenius-Lipschitz bound $3.4445 + 3(4.775)C^2 + 5(2.0315)C^4$ for Ψ on the stated spectral-norm-bounded set.
N; δ	Comparison matrix in the local sensitivity bound; positive lower bound $\min\{\lVert M\rVert_F, \lVert N\rVert_F\} \geq \delta$. Here N is not an input count.
Population and empirical input response	
A; h, h_j	Fixed parameter change of shape $m \times n$ in response calculations; input vector and sampled input vectors in $\mathbb{R}^n$. Their linear output changes are Ah and Ah_j.
Σ_h; Σ	Uncentered second moment $\Sigma_h = \mathbb{E}[hh^\top]$; a generic positive-semidefinite second-moment matrix in the matched-repair comparison.
$\Sigma_h^{1/2}$; $\Sigma^{1/2}$	Symmetric positive-semidefinite square roots of the respective second-moment matrices.
$\mathcal{E}(A)$	Population input-response energy $\mathbb{E}\lVert Ah\rVert_2^2/m = \lVert A\Sigma_h^{1/2}\rVert_F^2/m$, for the same fixed input distribution.
q; q_-, q_+	Isotropic second-moment scale in $\Sigma_h = qI_n$; positive bounds satisfying $q_-I_n \preceq \Sigma_h \preceq q_+I_n$.

Continued on next page

Symbol	Meaning
ρ_k	Weighted protected-direction response ratio $\|R_k\Sigma_h^{1/2}\|_F^2/\|R_0\Sigma_h^{1/2}\|_F^2$, defined for a positive denominator; not necessarily at most one.
N; H	Number of fixed input vectors; input matrix $\mathsf{H} = [h_1, \dots, h_N] \in \mathbb{R}^{n\times N}$. The sans-serif H differs from SBC's block H.
$\widehat{\Sigma}_H$	Empirical uncentered second moment $\mathsf{H}\mathsf{H}^\top/N$.
$\mathcal{E}_{\mathsf{H}}(A)$	Empirical input-response energy $\|A\mathsf{H}\|_F^2/(mN)$.
A_0; r, r_0	Reference parameter change; response amplitudes $r = \|A\mathsf{H}\|_F$ and $r_0 = \|A_0\mathsf{H}\|_F$.
$b_H(A; A_0)$	Absolute response deviation $\|(A - A_0)\mathsf{H}\|_F$.
$\delta_{\mathsf{H}}(A; A_0)$	Relative response deviation $b_H(A; A_0)/r_0$, defined for $r_0 > 0$.
$\cos_{\mathsf{H}}(A, A_0)$	Response cosine similarity $\langle A\mathsf{H}, A_0\mathsf{H}\rangle_F/(rr_0)$, defined for $r, r_0 > 0$.
$\nu_{\mathsf{H}}(A; A_0)$	Response amplitude ratio r/r_0, defined for $r_0 > 0$; distinct from a response-energy ratio.
Update stages and constraint diagnostics	
A_{req}; A_a	Raw requested change $-\alpha D$; raw change $-aD$ at the same SBC trial step used for a matched block-correction comparison.
A_{block}; A_{norm}	SBC changes after block correction and after direction normalization: $-a\widehat{D}(a)$ and $-aR(a)$.
A_{shrink}	RSP intermediate change $X - W = (A_{\mathrm{req}} - \gamma W)/(1 + \gamma)$, including shrinkage of the current weights.
A_{exec}	Executed change: $Z(a) - W$ for an accepted SBC trial, $\Pi_c(X) - W$ for RSP, and zero when the old weights are retained.
$-\langle M, A\rangle_F$	Alignment with negative momentum. Its normalized version divides by $\|M\|_F\|A\|_F$ when both norms are positive.
$\|A_{\mathrm{exec}}\|_2/\alpha$	Executed displacement divided by the requested budget, for $\alpha > 0$.
$\|\|W\|_2 - S\|/S$	Relative spectral-radius error for SBC; evaluated at W^+ in post-update execution logs.
$[\|W\|_2 - c]_+/S$	Normalized spectral-ceiling violation for RSP; interior points have zero violation.
$\#\{i : s_i > c\}$	Number of singular values actively capped by projection, equivalently $\#\{i : \sigma_i(X) > c\}$; distinct from the selected SBC block size.
k/d; a/α	Selected SBC block fraction; trial-to-requested step ratio for a positive budget, as summarized in the execution diagnostics.
Width-dependent scaling and arithmetic cost	
μP; μTransfer	Maximal-update parameterization and its associated cross-width hyperparameter-transfer procedure; these are framework names, not scalar variables.
$\mathrm{sr}(W_0)$	Initialization stable rank $\|W_0\|_F^2/\|W_0\|_2^2 = F_0^2/S^2$.
w; m_w, n_w, d_w	Model-width index; width-dependent matrix dimensions with $d_w = \min(m_w, n_w)$.
$W_{0,w}$; S_w	Initialized matrix at width w; its spectral norm $S_w = \|W_{0,w}\|_2$.
$\gamma_{t,w}$; $\Delta_{t,w}^{\mathrm{RSP}}$	Width-indexed RSP shrinkage coefficient and executed update, respectively.
χ_-; r (scaling)	Positive width-independent lower bound on $\mathrm{sr}(W_{0,w})/d_w$; common positive rescaling factor in the local scaling identity.
b (cost)	Number of layer-input vectors processed between optimizer updates, including all accumulation microbatches; for tokenwise projections, a token-position count.
F_{fb}	Per-update layer forward/backward arithmetic cost, including both weight and input gradients.
$F_{\mathrm{opt}}^{\mathrm{RSP}}$, $F_{\mathrm{opt}}^{\mathrm{SBC}}$	Per-update arithmetic costs of RSP and SBC, including the common five-step direction.
$J_{\max}$	Upper bound on SBC candidate evaluations in the arithmetic-cost analysis; the specified implementation uses a budget of twelve.
r, q (timing)	Ratios to a baseline of required updates and average time per update under the same task and hardware protocol. The time-to-target ratio is rq.
$O(\cdot)$; $\Theta(\cdot)$	Asymptotic upper bound; matching upper and positive lower bounds. Constants are independent of the varying quantities specified in each statement.
Learning-rate transfer and validation crossings	
o; η, ξ	Optimizer index; candidate learning rates, with ξ used as a dummy grid-search variable.
$\mathcal{G}_{o,w}$	Qualified nominal matrix-learning-rate grid for optimizer o at hidden width w.
$\mathcal{V}_{o,w}(\eta)$	Recorded validation MLM loss after 200 successful updates at learning rate η.

Continued on next page

Symbol	Meaning
$\widehat{\eta}_o$	Source-width minimizer of $\mathcal{V}_{o,256}$ over $\mathcal{G}_{o,256}$, reused unchanged at widths 512, 1024, and 2048.
$P_{o,w}$	Target-width transfer penalty $\mathcal{V}_{o,w}(\widehat{\eta}_o) - \min_{\eta \in \mathcal{G}_{o,w}} \mathcal{V}_{o,w}(\eta)$.
$\overline{P}_o$	Equally weighted mean transfer penalty $\frac{1}{3}\sum_{w\in\{512,1024,2048\}} P_{o,w}$.
t_i; $\mathcal{L}_i$; N	Recorded validation step and loss. In the sequence $t_0 < \cdots < t_N$, N is the last checkpoint index, not the response-sample count.
$L_\star$	Validation-loss target; the main ViT comparison uses Muon's final CE, $L_\star = 1.14657785987854$.
$\widehat{t}(L_\star)$	First target-crossing estimate from piecewise-linear interpolation. On a downward-crossing segment it equals $t_{i-1} + (t_i - t_{i-1})(\mathcal{L}_{i-1} - L_\star)/(\mathcal{L}_{i-1} - \mathcal{L}_i)$.
$\widehat{t}_{\text{method}}, \widehat{t}_{\text{baseline}}$	Crossing estimates for a method and its baseline; the reported fractional update reduction is $1 - \widehat{t}_{\text{method}}/\widehat{t}_{\text{baseline}}$.
Trajectory-level input-response probes	
p; $\mathcal{I}$	Matrix-probe index; set of included probes for a method and condition. This index p is distinct from the scalar polynomial $p(s)$.
m_p, n_p, N_p	Output dimension, input dimension, and retained input-vector count for probe p.
W_p, W_p^+; H_p	Current and stored post-update probe weights; sampled input matrix $\mathsf{H}_p \in \mathbb{R}^{n_p \times N_p}$.
α_p, D_p; a_p, R_p	Probe-specific requested budget and raw direction; final attempted trial step and its corrected, norm-bounded direction.
$e_p(A)$	Probe input-response energy $\|A\mathsf{H}_p\|_F^2/(m_p N_p)$.
A_p^{trial}; A_p^{exec}	Corrected, normalized pre-retraction trial change $-a_p R_p$; actual stored change $W_p^+ - W_p$.
r_p	Post-retraction to corrected-trial energy ratio $e_p(A_p^{\text{exec}})/e_p(A_p^{\text{trial}})$, defined only for positive trial energy.
$\overline{r}$	Unweighted mean $\lvert\mathcal{I}\rvert^{-1}\sum_{p\in\mathcal{I}} r_p$. Reported retention percentages are $100\overline{r}$, not ratios of pooled energies.

A Analysis of Hard Spectral Protection

Throughout Appendices A–D, matrices are real, m, n are positive integers, and $d = \min(m, n)$. For matrices of the same size, write $\langle A, B\rangle_F = \operatorname{tr}(A^\top B)$. The notation $[x]_+ = \max\{x, 0\}$ denotes the positive part. For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite. All algebraic guarantees concern exact arithmetic unless finite precision is explicitly discussed. The analysis is local to a matrix parameter; it does not assume that all network parameters use the same update. The fixed initialization scale is $S = \|W_0\|_2 > 0$. In local statements about an arbitrary W, its current norm is written $\sigma_1(W)$ rather than implicitly identifying it with S.

A.1 Spectral Safety and Uniform Necessity

Take a thin singular value decomposition

$$W = U \operatorname{diag}(\sigma_1, \ldots, \sigma_d) V^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_d \geq 0,$$

where $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$ have orthonormal columns. Write $\sigma_i = \sigma_i(W)$, and let U_k, V_k contain the first k columns. At repeated singular values, fix any paired orthonormal singular bases. The following one-step statements hold for every such choice. For $1 \leq k < d$, define

$$P_U = U_k U_k^\top, \qquad P_V = V_k V_k^\top,$$

$$\mathcal{X}_k(W) = \left\{ X \in \mathbb{R}^{m\times n} : U_k^\top X = 0,\ X V_k = 0,\ \|X\|_2 \leq 1 \right\}.$$

The hard-protected proposal is

$$B_k = (I_m - P_U) M (I_n - P_V), \qquad R_k = \operatorname{msign}(B_k).$$

Here, for a compact SVD $B_k = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ containing only its positive singular values, $\operatorname{msign}(B_k) = \widetilde{U}\widetilde{V}^\top$; the value at zero is zero. This convention does not complete a rank-deficient matrix to a full-rank orthogonal factor.

Lemma A.1 (Admissibility of the protected polar direction)**.** *For every $M \in \mathbb{R}^{m\times n}$, $R_k \in \mathcal{X}_k(W)$. If $r_k = \operatorname{rank}(B_k)$, then*

$$\|R_k\|_F^2 = r_k, \qquad r_k \le \min\{\operatorname{rank}(M), m-k, n-k\}.$$

Proof. The column space of B_k is contained in the orthogonal complement of U_k, and its row space is contained in the orthogonal complement of V_k. Its compact singular vectors consequently satisfy $U_k^\top \widetilde{U} = 0$ and $V_k^\top \widetilde{V} = 0$. The same constraints hold for $R_k = \widetilde{U}\widetilde{V}^\top$. Its r_k nonzero singular values are one, proving the norm statements. The rank bound follows because multiplication by either projector cannot increase rank. If $B_k = 0$, all statements follow from $R_k = 0$. □

The next proposition proves and sharpens the uniform formulation of Proposition 1. It extends the leading-direction construction of Zhao (2026).

Proposition A.2 (Exact worst-case norm under hard protection)**.** *For $1 \le k < d$ and every $a \ge 0$,*

$$\sup_{X\in\mathcal{X}_k(W)} \|W - aX\|_2 = \max\{\sigma_1(W), \sigma_{k+1}(W) + a\}. \tag{A.1}$$

Consequently, every $X \in \mathcal{X}_k(W)$ preserves the norm whenever $a \le g_k := \sigma_1(W) - \sigma_{k+1}(W)$. Conversely, $a \le g_k$ is necessary for uniform non-increase over $\mathcal{X}_k(W)$.

Proof. Complete U_k, V_k to square orthogonal matrices $\overline{U} = [U_k, U_\perp]$ and $\overline{V} = [V_k, V_\perp]$. In these coordinates,

$$\overline{U}^\top W \overline{V} = \begin{pmatrix} \Sigma_k & 0 \\ 0 & W_\perp \end{pmatrix}, \qquad \overline{U}^\top X \overline{V} = \begin{pmatrix} 0 & 0 \\ 0 & X_\perp \end{pmatrix},$$

Here $\Sigma_k = \operatorname{diag}(\sigma_1, \dots, \sigma_k)$. The lower-right blocks have size $(m-k)\times(n-k)$ and satisfy

$$\|W_\perp\|_2 = \sigma_{k+1}(W), \qquad \|X_\perp\|_2 \le 1.$$

This also covers rectangular matrices, including their null directions. Thus

$$\|W - aX\|_2 = \max\{\sigma_1(W), \|W_\perp - aX_\perp\|_2\} \le \max\{\sigma_1(W), \sigma_{k+1}(W) + a\}.$$

The lower bound $\|W - aX\|_2 \ge \sigma_1(W)$ follows from the unchanged protected block. Choose $X = -u_{k+1}v_{k+1}^\top$, which belongs to $\mathcal{X}_k(W)$. Then $W - aX$ increases the $(k+1)$st diagonal entry in singular coordinates from σ_{k+1} to $\sigma_{k+1} + a$, attaining (A.1). The stated equivalence follows. □

The necessity is uniform over a direction class, not a necessary condition for every individual update. Increasing k moves the relevant gap beyond the first two singular values, but does not eliminate gap dependence. If $k = d$, the two-sided constraints force $X = 0$, because at least one of U_d, V_d is square orthogonal. That degenerate case requires no σ_{d+1} and is excluded from the gap formula. When $d = 1$, there is no nontrivial choice $1 \le k < d$.

A.2 Update Energy and Spectral Protection

Let $h \in \mathbb{R}^n$ have finite second moment $\Sigma_h = \mathbb{E}[hh^\top]$. This is an uncentered second-moment matrix, not necessarily a covariance matrix. For a fixed parameter change $A \in \mathbb{R}^{m\times n}$, the linear response Wh changes by Ah at the same input. Define

$$\mathcal{E}(A) = \frac{1}{m}\mathbb{E}\|Ah\|_2^2. \tag{A.2}$$

When A depends on a training state, the statement below is applied after fixing that state and specifying the input distribution. It does not replace A by its expectation or require independence between an unfixed update and the inputs.

Lemma A.3 (Second-moment identity)**.** *For every fixed A,*

$$\mathcal{E}(A) = \frac{1}{m}\operatorname{tr}(A^\top A \Sigma_h) = \frac{1}{m}\|A\Sigma_h^{1/2}\|_F^2, \tag{A.3}$$

where $\Sigma_h^{1/2}$ is the symmetric positive-semidefinite square root.

Proof. Since $\mathbb{E}\|h\|_2^2 < \infty$,

$$\mathbb{E}\|Ah\|_2^2 = \mathbb{E}\,\mathrm{tr}(A^\top Ahh^\top) = \mathrm{tr}(A^\top A\Sigma_h) = \mathrm{tr}(\Sigma_h^{1/2}A^\top A\Sigma_h^{1/2}).$$

The last expression is $\|A\Sigma_h^{1/2}\|_F^2$. □

Applying the identity to $A = -aR_k$ gives

$$\mathcal{E}(-aR_k) = \frac{a^2}{m}\|R_k\Sigma_h^{1/2}\|_F^2. \tag{A.4}$$

For isotropic inputs $\Sigma_h = qI_n$, $q > 0$, Lemma A.1 yields

$$\mathcal{E}(-aR_k) = q\frac{a^2 r_k}{m}. \tag{A.5}$$

More generally, if $0 < q_- \le q_+ < \infty$ and $q_- I_n \preceq \Sigma_h \preceq q_+ I_n$, then

$$q_-\frac{a^2 r_k}{m} \le \mathcal{E}(-aR_k) \le q_+\frac{a^2 r_k}{m}.$$

These inequalities follow by applying the Loewner bounds inside $\mathrm{tr}(R_k^\top R_k\Sigma_h)$.

For requested step $\alpha \ge 0$, hard-safe clipping sets

$$a_k = \min\{\alpha, g_k\}, \qquad \mathcal{E}(-a_kR_k) = q\frac{a_k^2 r_k}{m} \quad \text{when } \Sigma_h = qI_n.$$

Under a fixed nested choice of singular bases, g_k is nondecreasing and r_k is nonincreasing in k: indeed, $B_{k+1} = (I_m - U_{k+1}U_{k+1}^\top)B_k(I_n - V_{k+1}V_{k+1}^\top)$. Their product in the energy formula need not be monotone.

For comparison with the unprotected exact polar direction $R_0 = \mathrm{msign}(M)$, define, when the denominator is positive,

$$\rho_k = \frac{\|R_k\Sigma_h^{1/2}\|_F^2}{\|R_0\Sigma_h^{1/2}\|_F^2}.$$

For $\alpha > 0$ with $\mathcal{E}(-\alpha R_0) > 0$,

$$\frac{\mathcal{E}(-a_kR_k)}{\mathcal{E}(-\alpha R_0)} = \Big(\frac{a_k}{\alpha}\Big)^2\rho_k.$$

In the isotropic case with $M \neq 0$, $\rho_k = r_k/\,\mathrm{rank}(M)$. For general Σ_h, ρ_k measures a weighted response and is not asserted to lie below one: projecting M and then taking its polar direction can change its orientation relative to the inputs. The exact rank identities above do not apply to finite Newton–Schulz directions. Their response is still given by (A.3); Appendix D.1 gives the corresponding spectral expression.

A.3 Singular-Direction Locking

Let $W_{t+1} = W_t + \Delta_t$, and let $U_t^\star, V_t^\star$ span the full left and right singular subspaces associated with $\sigma_1(W_t)$. Suppose

$$\|W_t\|_2 = S, \qquad (U_t^\star)^\top\Delta_t = 0, \qquad \Delta_t V_t^\star = 0 \quad \text{for all } t.$$

These conditions apply to the final change Δ_t, after all update operations.

Proof of Proposition 2. Fix an initial unit singular pair (u_0, v_0) satisfying $W_0v_0 = Su_0$ and $W_0^\top u_0 = Sv_0$. Suppose these relations hold at time t. Since $\|W_t\|_2 = S$, u_0 and v_0 belong to the current full leading singular subspaces. Hence $\Delta_t v_0 = 0$ and $\Delta_t^\top u_0 = 0$. Therefore

$$W_{t+1}v_0 = W_tv_0 + \Delta_tv_0 = Su_0,$$

$$W_{t+1}^\top u_0 = W_t^\top u_0 + \Delta_t^\top u_0 = Sv_0.$$

The assertion follows by induction, including times when the multiplicity of the largest singular value increases. □

A target that requires changing a locked pair cannot be fitted exactly. Specifically, for any fixed target $T \in \mathbb{R}^{m\times n}$ and $f(W) = \frac{1}{2}\|W - T\|_F^2$, define

$$b_L = Su_0 - Tv_0, \qquad b_R = Sv_0 - T^\top u_0.$$

The invariant implies

$$f(W_t) \geq \frac{1}{2}\max\{\|b_L\|_2^2, \|b_R\|_2^2\}. \tag{A.6}$$

Indeed, $\|(W_t - T)v_0\|_2 \leq \|W_t - T\|_F$ and $\|(W_t - T)^\top u_0\|_2 \leq \|W_t - T\|_F$. As an explicit example, take

$$W_0 = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}, \qquad T = \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix}.$$

Both lie on the same spectral sphere. The initial pair $u_0 = v_0 = (1, 0)^\top$ gives $f(W_t) \geq S^2/2$, whereas $f(T) = 0$. This is an optimization obstruction caused by the specified update restriction, not by the sphere constraint alone.

The result does not state that every initially protected direction remains locked. It concerns initial maximizers that remain protected because the full current maximal subspaces are covered. A rescaling or decay operation that violates the displayed conditions can break the invariant. Likewise, a scalar tangent condition is not the same as the two-sided annihilation used here.

A.4 Approximate Subspaces and Robust Spectral Safety

Fix W, $1 \leq k < d$, and its exact projectors P_U, P_V. Let $X \in \mathbb{R}^{m\times n}$ satisfy $\|X\|_2 \leq 1$, and define

$$\varepsilon_U = \|U_k^\top X\|_2, \qquad \varepsilon_V = \|XV_k\|_2.$$

No assumption is made about how X was computed. Decompose it as

$$X_0 = (I_m - P_U)X(I_n - P_V), \qquad E = X - X_0.$$

Then $X_0 \in \mathcal{X}_k(W)$ and

$$E = P_U X + (I_m - P_U)XP_V, \qquad \|E\|_2 \leq \varepsilon_U + \varepsilon_V.$$

The last inequality follows from $\|P_U X\|_2 = \|U_k^\top X\|_2$, $\|XP_V\|_2 = \|XV_k\|_2$, and contractivity of orthogonal projectors.

Proposition A.4 (Robust spectral-safety bound)**.** *For every $a \geq 0$,*

$$|\|W - aX\|_2 - \|W\|_2| \leq [a - g_k]_+ + a(\varepsilon_U + \varepsilon_V), \tag{A.7}$$

where $g_k = \sigma_1(W) - \sigma_{k+1}(W)$.

Proof. The block argument in Appendix A.1 gives

$$\sigma_1(W) \leq \|W - aX_0\|_2 \leq \sigma_1(W) + [a - g_k]_+.$$

The reverse triangle inequality also gives

$$|\|W - aX\|_2 - \|W - aX_0\|_2| \leq a\|E\|_2.$$

Combining the two inequalities proves the result. □

For orthonormal estimates $\widehat{U}_k, \widehat{V}_k$, define $\widehat{P}_U = \widehat{U}_k\widehat{U}_k^\top$ and $\widehat{P}_V = \widehat{V}_k\widehat{V}_k^\top$. If $\widehat{P}_U X = 0$ and $X\widehat{P}_V = 0$, then

$$\varepsilon_U \leq \|P_U - \widehat{P}_U\|_2\|X\|_2, \qquad \varepsilon_V \leq \|P_V - \widehat{P}_V\|_2\|X\|_2.$$

For example, $P_U X = (P_U - \widehat{P}_U)X$. Subspace perturbation estimates can be used to bound these projector differences (Wedin, 1972; Stewart & Sun, 1990); (A.7) itself requires no additional estimate.

To relate the bound to radial restoration, let $s = \|W\|_2 > 0$, $Y = W - aX$, and

$$b(a) = [a - g_k]_+ + a(\varepsilon_U + \varepsilon_V).$$

If $b(a) < s$, then $\|Y\|_2 \geq s - b(a) > 0$ and $Z = sY/\|Y\|_2$ is well defined. Directly,

$$\|Z\|_2 = s, \qquad \|Z - Y\|_2 = |s - \|Y\|_2| \leq b(a), \qquad \|Z - W\|_2 \leq a + b(a).$$

Thus a condition $b(a) \leq \epsilon s$ with $0 \leq \epsilon < 1$ provides a relative spectral-error tolerance and a bound on the additional restoration. This analysis concerns approximate hard protection. SBC's different block correction has its own acceptance analysis below.

B Details and Analysis of SBC

This appendix studies the SBC update in Section 3.3. At a fixed iteration, write $W = W_t \in \mathcal{S}_S$, $D = D_t$, and $\alpha = \alpha_t$, with $\|D\|_2 \le 1$. The weights, momentum, direction, and requested budget are fixed throughout a backtracking loop. Only the trial step and its correction change.

B.1 Spectral Block Selection and Slack Construction

When $\alpha = 0$ or $M_t = 0$, SBC returns W. Otherwise let $0 < a \le \alpha$ and take the thin SVD of W used in Appendix A.1. Define

$$k(a) = \max\left\{1, \#\{i \in \{1, \dots, d\} : \sigma_i(W) \ge S - 2a\}\right\}. \tag{B.1}$$

The symbol $\#$ denotes cardinality. In exact arithmetic, the set is nonempty because $\sigma_1(W) = S$. The maximum with one preserves the stated implementation convention. The value $k = d$ is allowed here; the gap formula for hard protection is not used to define SBC.

The neighborhood width $2a$ supplies a step-dependent margin. Indeed, the singular-value perturbation inequality $\sigma_i(W+E) \le \sigma_i(W) + \|E\|_2$ implies that $\sigma_i(W) < S - 2a$ entails $\sigma_i(W+E) < S - a$ whenever $\|E\|_2 \le a$. This explains the scale of the neighborhood, but does not assert that its coefficient is uniquely optimal or that the selected block alone guarantees the final constraint. Radial restoration and the actual-displacement check remain necessary parts of the algorithm.

For $k = k(a)$ define

$$\begin{aligned} H &= \operatorname{sym}(U_k^\top D V_k), \\ L &= \operatorname{diag}\left(\frac{S - \sigma_1(W)}{a}, \dots, \frac{S - \sigma_k(W)}{a}\right), \\ B &= H + L. \end{aligned} \tag{B.2}$$

Here $\operatorname{sym}(A) = (A + A^\top)/2$ for square A. The matrices H, L, B are symmetric $k \times k$ matrices. Each diagonal entry of L is the distance to the upper boundary, divided by the current trial step, and lies in $[0, 2]$. The use of L therefore distinguishes directions that are exactly at the boundary from nearby directions with positive slack.

Let $\mathcal{P}(B)$ be the selected projection defined in the next subsection, and set

$$K = \mathcal{P}(B) - B, \qquad \widehat{D}(a) = D + U_k K V_k^\top.$$

With $\Sigma_k = \operatorname{diag}(\sigma_1(W), \dots, \sigma_k(W))$, $\Sigma_k + aL = SI_k$. It follows that

$$\begin{aligned} \operatorname{sym}\big(U_k^\top (W - a\widehat{D}(a)) V_k\big) &= \Sigma_k - a(H + K) \\ &= SI_k - a\mathcal{P}(B). \end{aligned} \tag{B.3}$$

Since $\mathcal{P}(B)$ is positive semidefinite and singular, the symmetric matrix on the right has largest eigenvalue S. This is a one-sided eigenvalue condition on a symmetric-block proxy, not a complete spectral-norm constraint on the trial weights. For example, with $W = SI_2$ and

$$D = b\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad 0 < b \le 1,$$

the selected block is the whole matrix, $H = L = B = K = 0$, and $\operatorname{sym}(W - aD) = SI_2$. Nevertheless, $\|W - aD\|_2 = \sqrt{S^2 + a^2 b^2} > S$ for $a > 0$. This family includes the common direction obtained from the displayed skew-symmetric matrix as momentum, by Appendix D.1. This example isolates what the proxy does and does not impose. Direction normalization can also change (B.3).

The unnormalized correction has

$$\|\widehat{D}(a) - D\|_F = \|K\|_F, \qquad \|\widehat{D}(a) - D\|_2 = \|K\|_2.$$

If $U_\perp, V_\perp$ complete the selected bases, its cross blocks $U_k^\top D V_\perp$ and $U_\perp^\top D V_k$, as well as its complementary block, are unchanged at this stage. Unlike hard protection, SBC does not set whole protected rows and columns to zero in singular coordinates.

B.2 Projection onto Positive-Semidefinite Singular Matrices

Let $\mathbb{S}^k$ denote the real symmetric $k \times k$ matrices and set

$$\mathcal{C}_0^k = \{Z \in \mathbb{S}^k : Z \succeq 0,\ \lambda_{\min}(Z) = 0\}.$$

The set is nonempty and closed, but generally nonconvex. For $B \in \mathbb{S}^k$, the continuous objective $\|Z - B\|_F^2$ attains its minimum on this set because its sublevel sets are bounded. Define $\mathcal{P}(B)$ to be a selected minimizer:

$$\mathcal{P}(B) \in \arg\min_{Z \in \mathcal{C}_0^k} \|Z - B\|_F^2. \tag{B.4}$$

Proposition B.1 (Spectral solution of the local projection)**.** *Choose an orthogonal eigendecomposition $B = Q\,\mathrm{diag}(\ell_1, \ldots, \ell_k)Q^\top$, with $\ell_1 \le \cdots \le \ell_k$. A minimizer in* (B.4) *is*

$$\mathcal{P}(B) = Q\,\mathrm{diag}(\widetilde{\ell}_1, \ldots, \widetilde{\ell}_k)Q^\top, \qquad \widetilde{\ell}_i = \begin{cases} 0, & i = 1, \\ [\ell_i]_+, & 2 \le i \le k. \end{cases} \tag{B.5}$$

The minimum squared distance is $\ell_1^2 + \sum_{i=2}^k [-\ell_i]_+^2$.

Proof. For $Z \in \mathcal{C}_0^k$, order its eigenvalues as $0 = z_1 \le z_2 \le \cdots \le z_k$. The trace inequality for symmetric matrices gives $\mathrm{tr}(BZ) \le \sum_i \ell_i z_i$; this is the spectral alignment underlying unitarily invariant matrix optimization (Lewis, 1995). Consequently,

$$\|B - Z\|_F^2 = \sum_i \ell_i^2 + \sum_i z_i^2 - 2\,\mathrm{tr}(BZ) \ge \sum_i (\ell_i - z_i)^2.$$

The first eigenvalue must be zero. Each remaining squared term is minimized over $z_i \ge 0$ by $z_i = [\ell_i]_+$. These values already satisfy the required ordering. Choosing the same eigenvectors as B attains the trace inequality with equality and gives the displayed minimum. □

If $\ell_1 \le 0$, ordinary projection onto the positive-semidefinite cone already has a zero eigenvalue, so it is also the unique nearest point in $\mathcal{C}_0^k$. If B is positive definite, an additional smallest-eigenvalue direction must be set to zero. When that eigenvalue is repeated, the nearest point is not unique. The eigendecomposition returned by the symmetric eigensolver specifies the selection used in the implementation. Every such selection satisfies the projection and acceptance statements here; no basis-independent unique output is claimed at a nonunique projection. For $k = 1$, the set is $\{0\}$.

B.3 Direction Normalization and Finite-Step Acceptance

For a finite corrected direction, define

$$R(a) = \frac{\widehat{D}(a)}{\max\{1, \|\widehat{D}(a)\|_2\}}, \qquad Y_a = W - aR(a).$$

Then $\|R(a)\|_2 \le 1$. For $Y_a \ne 0$, let $Z(a) = SY_a/\|Y_a\|_2$.

Lemma B.2 (Displacement after radial restoration)**.** *If $\|W\|_2 = S$, $\|R(a)\|_2 \le 1$, and $Y_a \ne 0$, then*

$$\|Z(a)\|_2 = S, \qquad \|Z(a) - Y_a\|_2 \le a, \qquad \|Z(a) - W\|_2 \le 2a. \tag{B.6}$$

In particular, direction normalization alone need not imply $\|Z(a) - W\|_2 \le \alpha$ when $a \le \alpha$.

Proof. Homogeneity of the norm gives $\|Z(a)\|_2 = S$ and

$$\|Z(a) - Y_a\|_2 = |S - \|Y_a\|_2| \le \|W - Y_a\|_2 = a\|R(a)\|_2 \le a.$$

The triangle inequality yields the $2a$ bound. For a concrete illustration that the original budget need not hold, consider a generic normalized direction with

$$W = \mathrm{diag}(1, 1/2), \quad R = \mathrm{diag}(1, -1), \quad a = \alpha = 1/10.$$

Then $Y_a = \mathrm{diag}(9/10, 3/5)$, $Z(a) = \mathrm{diag}(1, 2/3)$, and $\|Z(a) - W\|_2 = 1/6 > \alpha$. This example concerns the normalization-and-restoration operations; it is not an assertion that this R must be produced by SBC's block rule. □

SBC therefore accepts a candidate only if

$$\|Z(a) - W\|_2 \le \alpha. \tag{B.7}$$

Acceptance, or retention of an initially feasible W, gives

$$\|W_{t+1}\|_2 = S, \qquad \|W_{t+1} - W_t\|_2 \le \alpha_t.$$

No conclusion about descent of the training loss follows from this feasibility test alone.

If $a < S$, the candidate cannot be zero, since $\|Y_a\|_2 \ge S - a > 0$. Moreover, $a \le \alpha/2$ is sufficient for the displacement test whenever the candidate is defined. These facts give a sufficient small-step regime without requiring continuity of the selected spectral block or its projection.

B.4 Backtracking and Finite-Precision Details

For a nonzero request, the implementation updates the momentum once, forms D, and computes the SVD of the current W before the trial loop. There are at most $J = 12$ candidates, indexed by $j = 0, \dots, J-1$, with $a_0 = \alpha$. At each trial it forms $k(a_j), L, B, K, R(a_j), Z(a_j)$ and checks (B.7). For an unsuccessful finite candidate, write $\rho_j = \|Z(a_j) - W\|_2/\alpha$ in the ideal arithmetic version. The registered step reduction is

$$a_{j+1} = a_j \max\left\{0.1, \min\left\{0.9, \frac{0.97}{\max\{\rho_j, 10^{-30}\}}\right\}\right\}. \tag{B.8}$$

The original α and direction D remain fixed. The selected block and the slack are recomputed at the new a_j; this is not a rescaling of the final displacement of an old candidate. Quantities depending only on W, D and an unchanged k may be reused. The first accepted candidate is returned. If all twelve candidates are rejected, W is retained and rejection is recorded.

There is a useful exact-arithmetic consequence of Lemma B.2. If $0 < \alpha < S$, then every trial matrix is nonzero and $a_j \le 0.9^j \alpha$. Since $0.9^7 < 1/2$, a valid candidate must satisfy the displacement test no later than the eighth trial. This follows directly from the prescribed reduction and the $2a$ bound; it is not an empirical claim about trial counts. In floating point, approximation, storage, and spectral checks can prevent this ideal argument from describing the observed trial count.

For finite-precision evaluation, let ϵ_g and ϵ_s be the machine epsilon values of the geometry and storage types. Define

$$\begin{aligned} \tau_{\text{step}} &= \max\{10^{-8}, 8\epsilon_g\}, \\ \tau_{\text{rad}} &= \max\{2 \times 10^{-12}, 8\epsilon_g\}, \\ \tau_{\text{stored}} &= \max\{\tau_{\text{rad}}, 8\epsilon_s\}. \end{aligned} \tag{B.9}$$

These epsilon values are the quantities returned by the numerical library's machine-epsilon convention. Norm checks use temporary FP64 arithmetic, while the common Newton–Schulz direction is computed in FP32 in the registered implementation.

A candidate is checked for finiteness and nonzero norm, and receives an additional radial restoration if its measured relative norm error exceeds τ_{rad}. It is then rounded to the output storage type *before* acceptance. Let $\widetilde{Z}$ denote those stored values, and use hats for the numerically evaluated norms. The acceptance tests are

$$\frac{|\widehat{\|\widetilde{Z}\|_2} - S|}{S} \le \tau_{\text{stored}}, \qquad \frac{\widehat{\|\widetilde{Z} - W\|_2}}{\alpha} \le 1 + \tau_{\text{step}}. \tag{B.10}$$

The stored tensors are promoted before subtraction when evaluating the second norm. The measured displacement ratio replaces ρ_j in (B.8). These are reproducible numerical checks, not interval-certified bounds on the exact norms of floating-point matrices.

A zero or nonfinite trial matrix, or a failed decomposition, raises a numerical failure rather than silently producing an accepted step. A zero request, zero momentum, or exhausted trial budget retains the old weights only after checking that their measured relative radius error is at most 10^{-4}; otherwise the implementation raises a failure. This retention check is distinct from the tighter candidate-acceptance tolerance and does not repair the old weights. Consequently, exact sphere invariance applies to the ideal update, whereas the implementation records the accepted or retained numerical error. Algorithmic rejection and externally skipped training updates, such as mixed-precision overflow, must be counted separately.

C Details and Analysis of RSP

This appendix analyzes the RSP update in Section 3.4. The feasible set is the closed convex ball $\mathcal{B}_c = \{Z \in \mathbb{R}^{m\times n} : \|Z\|_2 \le c\}$, with $c = \kappa S > 0$ and $\kappa \ge 1$. When the time index is suppressed, $a = \alpha_t$ denotes the requested step, not a backtracking variable. Set $Y = W - aD$ with $\|D\|_2 \le 1$, and let $\gamma \ge 0$ be the shrinkage coefficient.

C.1 Regularized Spectral Projection

We first derive the projection used by RSP. This is a standard spectral-function construction (Lewis, 1995; Parikh & Boyd, 2014). Take a thin SVD

$$X = U_X \operatorname{diag}(s_1, \dots, s_d) V_X^\top, \qquad s_1 \ge \cdots \ge s_d \ge 0,$$

and define

$$\Pi_c(X) = U_X \operatorname{diag}(\min\{s_i, c\})_{i=1}^d V_X^\top. \tag{C.1}$$

Proposition C.1 (Projection and regularized correction)**.** *The matrix $\Pi_c(X)$ is the unique minimizer of $\frac{1}{2}\|Z - X\|_F^2$ over $Z \in \mathcal{B}_c$. For every Y and $\gamma \ge 0$, the unique minimizer of*

$$\min_{Z\in\mathcal{B}_c} \left\{ \frac{1}{2}\|Z - Y\|_F^2 + \frac{\gamma}{2}\|Z\|_F^2 \right\} \tag{C.2}$$

is

$$W^+ = \Pi_c\left(\frac{Y}{1+\gamma}\right). \tag{C.3}$$

Proof. Let $z_1 \ge \cdots \ge z_d \ge 0$ be the singular values of a feasible Z, so $z_i \le c$. The singular-value trace inequality gives $\langle X, Z\rangle_F \le \sum_i s_i z_i$, and hence

$$\|Z - X\|_F^2 \ge \sum_{i=1}^d (z_i - s_i)^2.$$

Each term is minimized over $0 \le z_i \le c$ by $z_i = \min\{s_i, c\}$. These values remain ordered, and choosing the singular vectors of X attains equality. Uniqueness follows from strict convexity of squared distance and convexity of $\mathcal{B}_c$, even when an SVD is not unique. Next, completing the square gives

$$\frac{1}{2}\|Z - Y\|_F^2 + \frac{\gamma}{2}\|Z\|_F^2 = \frac{1+\gamma}{2}\left\|Z - \frac{Y}{1+\gamma}\right\|_F^2 + \frac{\gamma}{2(1+\gamma)}\|Y\|_F^2.$$

The final term is independent of Z, yielding (C.3). □

For completeness, projection satisfies the variational inequality

$$\langle X - \Pi_c(X), V - \Pi_c(X)\rangle_F \le 0 \quad \text{for every } V \in \mathcal{B}_c. \tag{C.4}$$

It follows by taking the right derivative at zero of squared distance along the feasible segment from $\Pi_c(X)$ to V. For $P_i = \Pi_c(X_i)$, applying this inequality with $V = P_2$ and $V = P_1$, respectively, and adding gives

$$\|P_1 - P_2\|_F^2 \le \langle X_1 - X_2, P_1 - P_2\rangle_F.$$

Thus Π_c is nonexpansive in Frobenius norm. For fixed c, γ, the map $Y \mapsto \Pi_c(Y/(1+\gamma))$ is consequently $1/(1+\gamma)$-Lipschitz in that norm. This is a property of the correction map with fixed arguments; it does not establish contraction of the full training dynamics, which also change the gradient, momentum, and proposal. Likewise, (C.2) is not an exact minimization of the original training loss plus a penalty.

C.2 Feasibility and Local Displacement Bounds

For the singular values s_i of any X, let $e_i = [s_i - c]_+$. Equation (C.1) gives

$$\begin{aligned} \Pi_c(X) - X &= -U_X \operatorname{diag}(e_i) V_X^\top, \\ \|\Pi_c(X) - X\|_2 &= [\|X\|_2 - c]_+, \\ \|\Pi_c(X) - X\|_F^2 &= \sum_{i=1}^{d} e_i^2. \end{aligned} \tag{C.5}$$

The rank of this correction equals $\#\{i : s_i > c\}$. In particular, feasible X is unchanged by projection. This statement concerns projection alone, not the preceding shrinkage.

Proposition C.2 (RSP feasibility and displacement)**.** *Suppose $\|W\|_2 \le c$, $a \ge 0$, $\|D\|_2 \le 1$, and $\gamma \ge 0$. For $Y = W - aD$ and $W^+ = \Pi_c(Y/(1+\gamma))$,*

$$\|W^+\|_2 \le c, \qquad \|W^+ - W\|_2 \le 2a + \frac{c\gamma}{1+\gamma}. \tag{C.6}$$

Proof. Feasibility follows from the projection formula. Put $P = \Pi_c(Y)$ and $\theta = (1+\gamma)^{-1}$. Since $\|Y\|_2 \le c + a$, (C.5) gives

$$\|P - W\|_2 \le \|P - Y\|_2 + \|Y - W\|_2 \le 2a.$$

The matrices $\Pi_c(Y)$ and $\Pi_c(\theta Y)$ share singular vectors. For every $y \ge 0$,

$$0 \le \min\{y, c\} - \min\{\theta y, c\} \le c(1-\theta).$$

For $y \le c$ this follows from $(1-\theta)y \le (1-\theta)c$; for $y > c$, the difference is either zero or $c - \theta y \le c - \theta c$. Therefore

$$\|W^+ - P\|_2 \le c(1-\theta) = \frac{c\gamma}{1+\gamma}.$$

Combining the last two bounds proves (C.6). The argument uses the common singular vectors of these two matrices, not a general assertion of spectral-norm nonexpansiveness. □

Induction from $\|W_0\|_2 = S \le c$ establishes feasibility at every ideal RSP step. If $a = 0$, the prescribed calibration gives $\gamma = 0$ and feasible weights are unchanged. If $D = 0$ but $\gamma > 0$, then $W^+ = W/(1+\gamma)$ because this shrunken matrix is already feasible. Thus zero momentum does not suppress RSP's regularization.

The guarantees concern exact arithmetic. If the stored result is $\widetilde{W}^+ = W^+ + E_{\mathrm{num}}$, its exact norms obey

$$\begin{aligned} \|\widetilde{W}^+\|_2 &\le c + \|E_{\mathrm{num}}\|_2, \\ \|\widetilde{W}^+ - W\|_2 &\le 2a + \frac{c\gamma}{1+\gamma} + \|E_{\mathrm{num}}\|_2. \end{aligned}$$

A numerical residual measurement is not by itself a certified bound on E_{num}. RSP does not use SBC's final-displacement acceptance loop.

C.3 Initialization-Based Shrinkage Calibration

For each matrix, set $F_0 = \|W_0\|_F > 0$ and use

$$\gamma_t = \lambda \alpha_t \frac{\sqrt{d}}{F_0}, \qquad c = \kappa S, \qquad \lambda \ge 0, \quad \kappa \ge 1. \tag{C.7}$$

The registered default values are $\lambda = 0.05$ and $\kappa = 1.1$. The coefficient λ specifies shrinkage independently of the ceiling multiplier κ.

For $X = Y/(1+\gamma)$, the exact pre-projection shrinkage is

$$\|X - Y\|_F = \frac{\gamma}{1+\gamma}\|Y\|_F. \tag{C.8}$$

At $W = W_0$ and $Y = W_0 - aD$ with $\|D\|_2 \le 1$, $\|Y - W_0\|_F \le a\sqrt{d}$. For fixed W_0, m, n, λ and $a \to 0$, this implies

$$\|Y\|_F = F_0 + O(a), \qquad \|X - Y\|_F = \lambda a\sqrt{d} + O(a^2).$$

Here $O(a^2)$ has constants allowed to depend on the fixed quantities. A finite-step error bound, also allowing $W \neq W_0$, is

$$\left| \|X - Y\|_F - \lambda a\sqrt{d} \right| \le \frac{\gamma}{1+\gamma} \left(\|W - W_0\|_F + a\sqrt{d} + \gamma F_0 \right). \tag{C.9}$$

To obtain it, use $\lambda a\sqrt{d} = \gamma F_0$ in (C.8), then apply $|\|Y\|_F - F_0| \le \|Y - W_0\|_F$. The bound concerns shrinkage before projection. When the cap is inactive, it also describes the entire post-proposal correction.

The calibration also admits a budget interpretation. Compare it with subtracting $\tau = \lambda a$ from every singular value. For a full-rank trial matrix whose singular values all exceed τ, subtracting τ from each of its d singular values produces Frobenius correction $\tau\sqrt{d}$. The chosen γ matches this magnitude to first order at initialization. It does not match the corrections at every later iteration, when $\|Y\|_F$, rank, and active spectral caps can differ.

The location of shrinkage also has an exact algebraic interpretation. Write $\omega = \lambda\sqrt{d}/F_0$, so $\gamma = a\omega$, and let $\widetilde{a} = a/(1 + a\omega)$. Then, for a fixed direction D,

$$\frac{W - aD}{1 + a\omega} = (1 - \widetilde{a}\omega)W - \widetilde{a}D. \tag{C.10}$$

Thus trial-point shrinkage can be rewritten as decoupled decay with a correspondingly rescaled direction step. It is not the same as shrinking only the old weights while holding the nominal step a fixed, nor does it send a regularization gradient into the momentum. This relation identifies the role of step calibration; it does not introduce a new quadratic penalty (Loshchilov & Hutter, 2019).

C.4 Comparison with Matched Radial Rescaling

For a fixed $X \in \mathbb{R}^{m\times n}$ and $c > 0$, define

$$T_c(X) = \begin{cases} X \min\{1, c/\|X\|_2\}, & X \neq 0, \\ 0, & X = 0. \end{cases}$$

This is radial capping: interior points are left unchanged. It is different from restoration to a fixed sphere, which rescales every nonzero point to the target radius. Let $E_P = \Pi_c(X) - X$ and $E_T = T_c(X) - X$.

Proposition C.3 (Matched input-response comparison)**.** *For every X, $c > 0$, and $h \in \mathbb{R}^n$,*

$$E_P^\top E_P \preceq E_T^\top E_T, \qquad \|E_P h\|_2 \le \|E_T h\|_2. \tag{C.11}$$

Consequently, for every positive-semidefinite second-moment matrix $\Sigma \in \mathbb{R}^{n\times n}$,

$$\|E_P \Sigma^{1/2}\|_F^2 \le \|E_T \Sigma^{1/2}\|_F^2. \tag{C.12}$$

Proof. If $\|X\|_2 \le c$, both residuals vanish. Otherwise, let $s_1 = \|X\|_2 > c$ and write $X = U_X \operatorname{diag}(s_i) V_X^\top$. The residuals have the same left and right singular coordinates:

$$E_P = -U_X \operatorname{diag}(p_i) V_X^\top, \qquad E_T = -U_X \operatorname{diag}(r_i) V_X^\top,$$

where $p_i = [s_i - c]_+$ and $r_i = (1 - c/s_1)s_i$. If $s_i \le c$, then $p_i = 0 \le r_i$. If $s_i > c$, then

$$r_i - p_i = c\left(1 - \frac{s_i}{s_1}\right) \ge 0.$$

Hence $0 \le p_i \le r_i$ for every i, proving the Loewner inequality and the pointwise response bound. Finally,

$$\|E_T \Sigma^{1/2}\|_F^2 - \|E_P \Sigma^{1/2}\|_F^2 = \operatorname{tr}\left(\Sigma^{1/2} (E_T^\top E_T - E_P^\top E_P) \Sigma^{1/2} \right) \ge 0.$$

□

When $s_1 > c$, both residuals have spectral norm $s_1 - c$. The distinction lies in the smaller singular directions: $r_i > p_i$ whenever $0 < s_i < s_1$. Thus the response inequality is strict for inputs with a nonzero component in at least one such right-singular direction, and is an equality for inputs supported only on maximal or zero singular directions. For RSP the comparison uses the same $X = Y/(1+\gamma)$ on both sides, with the same shrinkage and cap. It proves a statement about the additional repair, not about $\|W^+ - W\|$, a comparison of different training trajectories, or the resulting loss. It also does not identify Frobenius projection as the minimizer of every input-weighted distance over the feasible ball.

D Common Update and Response Analysis

D.1 Finite-Step Newton–Schulz Updates

The common direction uses the five-step polynomial from the Muon construction (Jordan et al., 2024), with an explicit final scale factor. For $M \in \mathbb{R}^{m\times n}$, $M \neq 0$, define

$$\begin{aligned} X_0 &= \frac{M}{\|M\|_F}, \\ X_{j+1} &= \Psi(X_j), \qquad j = 0,1,2,3,4, \\ \Psi(X) &= 3.4445X - 4.775XX^\top X + 2.0315(XX^\top)^2X, \\ \mathcal{N}_5(M) &= \frac{X_5}{C}, \qquad C = 1.203. \end{aligned} \tag{D.1}$$

Set $\mathcal{N}_5(0) = 0$. Decimal coefficients in the following proof denote exact rationals. The two expressions $(XX^\top)^jX$ and $X(X^\top X)^j$ agree in exact arithmetic. Implementations may therefore use the smaller Gram matrix without changing the mathematical map. Their floating-point evaluation orders need not agree exactly.

For a thin SVD $X = U\,\mathrm{diag}(s_i)V^\top$, one has

$$\Psi(X) = U\,\mathrm{diag}(p(s_i))V^\top, \qquad p(s) = \frac{6889s - 9550s^3 + 4063s^5}{2000}.$$

We next prove the interval bound used in the main text.

Lemma D.1 (An invariant scalar interval)**.** *For every $s \in [0, C]$, $0 \le p(s) < C$. Moreover, $p(s) > 0$ whenever $s > 0$.*

Proof. Write

$$p(s) = \frac{s}{2000}\left(4063s^4 - 9550s^2 + 6889\right).$$

The quadratic $4063z^2 - 9550z + 6889$ has positive leading coefficient and discriminant $9550^2 - 4(4063)(6889) = -20757528 < 0$. It is strictly positive for every real z, proving positivity of p for $s > 0$. The derivative is

$$p'(s) = \frac{20315s^4 - 28650s^2 + 6889}{2000}.$$

Its two positive roots $r_- < r_+$ satisfy

$$r_\pm^2 = \frac{28650 \pm \sqrt{261022360}}{40630}.$$

They lie in $(0,1)$ and $(1,C)$, respectively, as also follows from $p'(0) > 0$, $p'(1) < 0$, and $p'(C) > 0$. Thus the maximum on $[0,C]$ is attained at 0, r_-, or C. Let $l = 554528/10^6$ and $u = 554530/10^6$. Direct rational evaluation of p' gives $l < r_- < u$, and therefore

$$p(r_-) \le \frac{6889u - 9550l^3 + 4063u^5}{2000} < \frac{12024}{10000} < C.$$

The displayed upper bound is a rational inequality. Finally,

$$p(0) = 0, \qquad p(C) = \frac{18980424935800053 09}{2000000000000000000} < 1 < C.$$

All possible maximizers consequently lie below C. □

Proposition D.2 (Bounded common direction)**.** *For every real M, $\|\mathcal{N}_5(M)\|_2 \le 1$. If $M \neq 0$, the exact map preserves its nonzero singular support, but does not generally set its nonzero singular values to one.*

Proof. For $M \neq 0$, $\|X_0\|_2 \le \|X_0\|_F = 1 < C$. Lemma D.1 and the spectral expression for Ψ imply inductively that every singular value of every X_j lies in $[0, C]$. Division by C proves the norm bound. Strict positivity of p on positive arguments preserves nonzero singular support, and $p(0) = 0$ preserves the zero singular values. The case $M = 0$ follows from the definition. □

More explicitly, if s_i are the singular values of $M \neq 0$, the singular values of $\mathcal{N}_5(M)$, possibly reordered, are

$$f_i = \frac{p^{\circ 5}(s_i/\|M\|_F)}{C}.$$

For isotropic inputs $\Sigma_h = qI_n$,

$$\mathcal{E}(-a\mathcal{N}_5(M)) = \frac{qa^2}{m}\sum_{i=1}^{d} f_i^2.$$

For finite iterations, this sum replaces the rank appearing in Appendix A.2. Thus rank alone does not describe the energy of the implemented direction.

The map $M \mapsto \mathcal{N}_5(M)$ is smooth on $M \neq 0$, including nonzero rank-deficient matrices: normalization is smooth away from zero, and the remaining operations are matrix polynomials. A quantitative, though loose, bound follows by setting

$$L_\Psi = 3.4445 + 3(4.775)C^2 + 5(2.0315)C^4.$$

For $\|X\|_2, \|Y\|_2 \le C$, telescoping the cubic and quintic products gives

$$\|\Psi(X) - \Psi(Y)\|_F \le L_\Psi \|X - Y\|_F.$$

Each cubic product difference has three terms and each quintic difference has five; every unchanged factor is bounded in spectral norm by C. If $\min\{\|M\|_F, \|N\|_F\} \ge \delta > 0$, then

$$\left\| \frac{M}{\|M\|_F} - \frac{N}{\|N\|_F} \right\|_F \le \frac{2}{\delta}\|M - N\|_F,$$

which follows by adding and subtracting $N/\|M\|_F$ and using the reverse triangle inequality. Hence

$$\|\mathcal{N}_5(M) - \mathcal{N}_5(N)\|_F \le \frac{2L_\Psi^5}{C\delta}\|M - N\|_F.$$

This bound does not involve a smallest nonzero singular value, but is not a useful small numerical sensitivity constant. At $M = 0$, a continuous extension is impossible for this normalization: $\mathcal{N}_5(rM) = \mathcal{N}_5(M)$ for $r > 0$ and $M \neq 0$. The exact interval proof excludes FP32 coefficient rounding, matrix products, and underflow; finite-precision constraint checks are separate.

D.2 Spectral Scaling Relations

The momentum recursion is $M_t = \beta M_{t-1} + (1 - \beta)G_t$, with $M_{-1} = 0$, $0 \le \beta < 1$, and $G_t \in \mathbb{R}^{m\times n}$. Let $D_t = \mathcal{N}_5(M_t)$, $\alpha_t = \eta_t S$ with $\eta_t \ge 0$, and $\Delta_t = W_{t+1} - W_t$. The accepted-step statement in Appendix B.3 and Proposition C.2 give

$$\begin{aligned} \text{SBC}: \quad & \|W_{t+1}\|_2 = S, \\ & \frac{\|\Delta_t\|_2}{S} \le \eta_t; \\ \text{RSP}: \quad & \|W_{t+1}\|_2 \le \kappa S, \\ & \frac{\|\Delta_t\|_2}{S} \le 2\eta_t + \frac{\kappa\gamma_t}{1+\gamma_t}. \end{aligned} \tag{D.2}$$

For SBC, a retained feasible point also satisfies these inequalities. They are exact-arithmetic statements, not floating-point equalities.

Define the initialization stable rank

$$\mathrm{sr}(W_0) = \frac{F_0^2}{S^2}.$$

Because $W_0 \neq 0$, $1 \leq \mathrm{sr}(W_0) \leq d$. The calibration satisfies the exact identity

$$\gamma_t = \lambda\eta_t\sqrt{\frac{d}{\mathrm{sr}(W_0)}}. \tag{D.3}$$

To state its width dependence precisely, index models by a width parameter w, with dimensions m_w, n_w and $d_w = \min(m_w, n_w)$. Suppose there is a width-independent constant $\chi_- > 0$ such that

$$\chi_- \leq \frac{\mathrm{sr}(W_{0,w})}{d_w} \leq 1.$$

For common λ and η_t, this gives

$$\lambda\eta_t \leq \gamma_{t,w} \leq \frac{\lambda\eta_t}{\sqrt{\chi_-}}.$$

Consequently, for common κ,

$$\frac{\|\Delta_{t,w}^{\mathrm{RSP}}\|_2}{S_w} \leq \left(2 + \frac{\kappa\lambda}{\sqrt{\chi_-}}\right)\eta_t, \qquad S_w = \|W_{0,w}\|_2.$$

Thus this calibration introduces no additional power of width when the relative initialization stable rank is bounded away from zero. It does not show that every initialization satisfies this condition.

These bounds are conditional on the chosen S_w. For example, if an initialization protocol satisfies $S_w = \Theta(\sqrt{m_w/n_w})$, the displayed bounds imply weights of that order for SBC and weight upper bounds of that order for RSP, together with update upper bounds of order $\eta_t\sqrt{m_w/n_w}$ under the stated stable-rank condition. Here $O(\cdot)$ denotes an upper bound by a width-independent constant, and $\Theta(\cdot)$ additionally requires a positive lower bound. This conditional statement connects the analysis to spectral scaling (Yang et al., 2023); it does not prescribe the initialization or replace the architecture-specific parameterization used in the experiments.

There is also a local scaling identity. If $W_0, W, M, S, F_0, a, \alpha, c$ are multiplied by the same $r > 0$, while η, λ, κ are unchanged, then $D = \mathcal{N}_5(M)$ is unchanged. In SBC, k, H, L, B, K, R and the displacement ratio are unchanged, and each candidate Z is multiplied by r, provided the same admissible eigenbasis selection is used. In RSP, γ is unchanged and $\Pi_{rc}(rX) = r\Pi_c(X)$. This is an algebraic property at a fixed matrix dimension, not an equivalence between the gradient dynamics of networks with different widths.

Neither an upper bound on $\|\Delta_t\|_2$ nor exact preservation of $\|W_t\|_2$ supplies a lower bound on a useful input response. The update can vanish, or the inputs can lie in its null space. Likewise, with zero directions and a fixed positive γ, RSP produces $W_t = (1+\gamma)^{-t}W_0$, so its norm can approach zero while remaining feasible. Cross-width learning-rate transfer is therefore evaluated separately rather than inferred from feasibility.

D.3 Input-Response Diagnostics

Fix a weight and momentum snapshot, and collect $N \geq 1$ input vectors $h_j \in \mathbb{R}^n$ as $\mathsf{H} = [h_1, \ldots, h_N] \in \mathbb{R}^{n\times N}$. These are inputs to the same linear map $h \mapsto Wh$. For a parameter change $A \in \mathbb{R}^{m\times n}$, its response matrix is $A\mathsf{H}$. The empirical second moment is $\widehat{\Sigma}_H = \mathsf{H}\mathsf{H}^\top/N$, giving the exact identity

$$\mathcal{E}_{\mathsf{H}}(A) = \frac{\|A\mathsf{H}\|_F^2}{mN} = \frac{1}{m}\,\mathrm{tr}(A^\top A\widehat{\Sigma}_H) = \frac{1}{mN}\sum_{j=1}^{N}\|Ah_j\|_2^2. \tag{D.4}$$

No independence assumption is required for this identity. If inputs and gradients share a batch, this conditional snapshot need not be an unbiased population estimate.

For a reference change $A_0 \in \mathbb{R}^{m\times n}$, define $r = \|A\mathsf{H}\|_F$ and $r_0 = \|A_0\mathsf{H}\|_F$. The absolute response deviation is $b_H(A; A_0) = \|(A - A_0)\mathsf{H}\|_F$. When their denominators are nonzero, define

$$\begin{aligned} \delta_{\mathsf{H}}(A; A_0) &= \frac{b_H(A; A_0)}{r_0}, \\ \cos_{\mathsf{H}}(A, A_0) &= \frac{\langle A\mathsf{H}, A_0\mathsf{H}\rangle_F}{r\, r_0}, \\ \nu_{\mathsf{H}}(A; A_0) &= \frac{r}{r_0}. \end{aligned} \tag{D.5}$$

They measure relative deviation, directional agreement, and response amplitude ratio, respectively. If $r, r_0 > 0$, expansion of the squared Frobenius norm gives

$$\delta_{\mathsf{H}}(A; A_0)^2 = \nu_{\mathsf{H}}(A; A_0)^2 + 1 - 2\nu_{\mathsf{H}}(A; A_0)\cos_{\mathsf{H}}(A, A_0). \tag{D.6}$$

Thus equal response energy does not imply equal responses, and a larger response can still have poor alignment with the reference.

If $r_0 = 0$, the relative deviation and amplitude ratio are undefined; report the absolute deviation and the two energies instead. If $r = 0 < r_0$, the relative deviation equals one, the amplitude ratio is zero, and the cosine is undefined. If both are zero, the absolute deviation is zero but the cosine and reference-normalized ratios remain undefined. Undefined values are recorded as missing, not assigned a perfect cosine. A numerical denominator threshold, when used, must be reported with the valid-sample count; adding a silent denominator offset changes the metric in (D.5).

For SBC, distinguish the following changes at an accepted trial a:

$$\begin{aligned} A_{\mathrm{req}} &= -\alpha D, \\ A_a &= -aD, \\ A_{\mathrm{block}} &= -a\widehat{D}(a), \\ A_{\mathrm{norm}} &= -aR(a), \\ A_{\mathrm{exec}} &= Z(a) - W. \end{aligned} \tag{D.7}$$

The distinction between A_{req} and A_a separates backtracking from block correction. In particular,

$$\begin{aligned} A_{\mathrm{exec}} - A_{\mathrm{req}} = (\alpha - a)D - a\big(\widehat{D}(a) - D\big) \\ - a\big(R(a) - \widehat{D}(a)\big) + \big(Z(a) - Y_a\big), \end{aligned} \tag{D.8}$$

where $Y_a = W - aR(a)$. The four terms correspond to the changed trial step, block correction, direction normalization, and radial restoration. To isolate block correction, compare A_{block} with A_a, not with a raw proposal at a different step. For a retained update, $A_{\mathrm{exec}} = 0$; a final rejected candidate is not an executed change. Requested, accepted, and rejected-candidate block sizes must be distinguished in logs.

For RSP, the corresponding quantities are

$$\begin{aligned} A_{\mathrm{req}} &= -\alpha D, \qquad Y = W + A_{\mathrm{req}}, \\ X &= \frac{Y}{1+\gamma}, \\ A_{\mathrm{shrink}} &= X - W = \frac{A_{\mathrm{req}} - \gamma W}{1+\gamma}, \\ A_{\mathrm{exec}} &= \Pi_c(X) - W. \end{aligned} \tag{D.9}$$

The decomposition

$$A_{\mathrm{exec}} = A_{\mathrm{req}} + (X - Y) + (\Pi_c(X) - X)$$

separates shrinkage and projection. In particular, A_{shrink} is not merely the raw direction multiplied by $1/(1+\gamma)$; it includes shrinkage of W. The comparison in Appendix C.4 uses the same W, D, α, γ, c and compares $\Pi_c(X) - X$ with $T_c(X) - X$. It does not order the total executed-response energies.

Parameter-space diagnostics use the same changes. For example, $-\langle M, A\rangle_F$ measures alignment of A with negative momentum, and its normalized version is

$$\frac{-\langle M, A\rangle_F}{\|M\|_F\|A\|_F}$$

when both norms are positive. Agreement with momentum is not a certificate of descent for the current true gradient. For $\alpha > 0$, $\|A_{\text{exec}}\|_2/\alpha$ is the executed step ratio. A zero request is reported separately rather than assigned an arbitrary ratio. For SBC, the radius diagnostic is $|\|W\|_2 - S|/S$; for RSP, cap violation is $[\|W\|_2 - c]_+/S$. An interior RSP iterate is not a constraint error. Its projection-active count $\#\{i : \sigma_i(X) > c\}$ is different from SBC's selected block size $k(a)$.

Local comparisons hold weights, momentum, and inputs fixed. Measurements on separate training trajectories have different conditioning states and are reported as trajectory-level evidence. All these diagnostics describe the action and execution of an update; neither increased response energy nor reduced local repair alone establishes improved training loss or generalization.

D.4 Arithmetic Cost and Token-Batch Amortization

The following is an arithmetic-cost analysis of the specified algorithms, not an additional timing experiment. Following the matrix-level comparison used for Muon (Jordan et al., 2024), consider a dense layer $h \mapsto Wh$, with $W \in \mathbb{R}^{m\times n}$ and $d = \min(m, n)$. Let $b \geq 1$ be the number of input vectors processed by this layer between consecutive optimizer updates, including all gradient-accumulation microbatches. For tokenwise Transformer projections, these vectors correspond to token positions, not the number of sequences or the total number of tokens over the full training run.

We count each scalar multiplication or addition as one floating-point operation (FLOP), use conventional dense matrix products, and hold numerical accuracy and decomposition budgets fixed. Thin SVD and spectral-norm computations are assigned their standard dense cost $O(mnd)$, and a symmetric eigendecomposition of size k is assigned $O(k^3)$. These are assumptions of the arithmetic model; solver convergence, precision, communication, and hardware utilization determine actual latency separately. Write F_{fb} for the layer's forward/backward cost and $F_{\text{opt}}^{\text{RSP}}$ and $F_{\text{opt}}^{\text{SBC}}$ for the respective per-update optimizer costs, including their common five-step direction.

Proposition D.3 (Token-batch amortization)**.** *Suppose both weight and input gradients are computed, the common Newton–Schulz direction uses five iterations with the smaller Gram matrix, and SBC evaluates at most $J_{\max}$ candidates. In the dense arithmetic model above,*

$$\begin{aligned} F_{\text{fb}} &= \Theta(mnb), \\ F_{\text{opt}}^{\text{RSP}} &= O(mnd), \\ F_{\text{opt}}^{\text{SBC}} &= O((1 + J_{\max})mnd). \end{aligned} \tag{D.10}$$

Consequently,

$$\begin{aligned} \frac{F_{\text{opt}}^{\text{RSP}}}{F_{\text{fb}}} &= O\left(\frac{d}{b}\right), \\ \frac{F_{\text{opt}}^{\text{SBC}}}{F_{\text{fb}}} &= O\left(\frac{(1 + J_{\max})d}{b}\right). \end{aligned} \tag{D.11}$$

The constants are independent of b and the realized block sizes.

Proof. The forward pass and the two backward products each cost $2mnb$ leading-order FLOPs, giving $6mnb$ in total. Momentum accumulation is $O(mn)$. One Newton–Schulz iteration implemented with the smaller Gram matrix costs $4mnd + 2d^3 \leq 6mnd$ leading-order FLOPs, so five iterations cost $O(mnd)$. RSP adds scalar shrinkage, one thin SVD, singular-value capping, and reconstruction, all within $O(mnd)$.

SBC first computes the SVD of W. At a trial with block size $k \leq d$, extracting the block and forming $U_k K V_k^\top$ cost $O(mnk + (m + n)k^2)$; the local eigendecomposition costs $O(k^3)$. Since $k \leq d \leq m, n$,

$$mnk \leq mnd, \qquad (m + n)k^2 \leq 2mnd, \qquad k^3 \leq mnd.$$

The direction, candidate, and displacement norm checks require at most a fixed number of additional dense spectral computations per candidate, each with cost $O(mnd)$ under the model. Therefore each candidate costs $O(mnd)$, and the full loop costs $O((1 + J_{\max})mnd)$ including the initial decomposition. Dividing by the forward/backward cost proves (D.11). □

For fixed dimensions, numerical budgets, and $J_{\max}$, the worst-case relative arithmetic overhead is thus bounded by a quantity proportional to $1/b$. Increasing gradient accumulation while keeping the microbatch size fixed is one way to increase b without repeating the spectral correction for each microbatch. This does not establish that the realized SBC overhead decreases monotonically, since the selected blocks and trial counts can change with the gradients. Increasing only the total training duration does not yield this amortization, and increasing width at fixed b need not reduce relative overhead.

For a fixed target quality, let r be the ratio of required updates to a baseline and q the ratio of average time per update, measured under the same task and hardware protocol. The ratio of times to target is rq, so fewer updates imply an actual time saving only if $rq < 1$. Changing the token batch can itself change r. The proposition therefore establishes an amortization mechanism, not a wall-clock speedup or a large-batch quality result for the reported experiments.

E Experimental Details

E.1 Models, Data, and Training Protocols

The main language experiment trains BERT-Large (Devlin et al., 2019) from scratch on the English Wikipedia 20231101 snapshot. The model has 24 Transformer layers, hidden width 1,024, 16 attention heads, feedforward width 4,096, and approximately 336M parameters. The vocabulary contains 30,522 tokens. Hidden and attention dropout probabilities are both 0.1. Training combines masked language modeling (MLM), with masking probability 0.15, and next-sentence prediction (NSP). The main validation loss is the sum of the two recorded loss components; the MLM component is also reported separately.

We use sequence length 512, global batch size 192, microbatch size two, FP16 mixed precision, and training seed 210901. The training data and evaluation seeds are 190901 and 190902, respectively. The budget of 2,035 attempted batches corresponds to 200,048,640 input positions, including padding and special tokens, rather than that many non-padding tokens. The learning rate follows cosine decay after 6% warmup, ending at 10% of its nominal value. Gradient clipping uses maximum norm 1.0, and activation recomputation is enabled. Each recorded validation evaluates 4,096 examples.

The vision experiment trains ViT-L/16 (Dosovitskiy et al., 2021) on the 10-class Imagenette subset (Howard, 2019) with 8,000 training and 2,000 validation images. The model has 24 layers, hidden width 1,024, 16 attention heads, feedforward ratio four, and approximately 303M parameters. Images have resolution 224×224 and are divided into 16×16 patches. The training seed is 42. We use BF16 mixed precision, global batch size 64, microbatch size 16, and four accumulation microbatches per update. All runs complete 3,000 successful updates on a single H200 GPU. The learning rate has 100 warmup steps followed by cosine decay to 1% of its nominal value. Gradient clipping uses maximum norm 1.0, and activation recomputation is enabled.

Training augmentation uses Mixup with coefficient 0.8 and CutMix with coefficient 1.0 (Zhang et al., 2018; Yun et al., 2019), label smoothing 0.1, and RandAugment with two operations and magnitude nine. The random-resized-crop scale range is $[0.08, 1]$, color jitter is 0.3, and the evaluation crop ratio is 0.875. Stochastic depth has rate 0.45, and LayerScale is initialized to 10^{-4}. Training uses mixed-label soft-target cross-entropy, whereas validation uses clean hard-label cross-entropy and top-1 accuracy. These two losses are not interchangeable objective values.

Both main tasks are evaluated every 250 steps and at the final step; ViT additionally records its initial validation result. Training curves use a 10-step trailing arithmetic mean. Endpoint comparisons and validation-crossing estimates use unsmoothed recorded evaluations. Each main result is one run, not an average over independently repeated seeds.

Table 6: Nominal matrix learning rates and matrix weight decay in the main experiments.

	BERT-Large		ViT-L/16	
Method	Learning rate	Decay	Learning rate	Decay
AdamW	10^{-4}	0.01	10^{-3}	0.05
Muon	0.02	0	0.02	0.01
SSO	0.02	0	0.01	0
RSP	0.02	0	0.01	0
SBC	0.02	0	0.01	0

Table 7: Successful updates and overflow skips in the main runs.

	BERT: 2,035 attempts		ViT: 3,000 attempts	
Method	Successful	Overflow	Successful	Overflow
AdamW	2,033	2	3,000	0
Muon	1,391	644	3,000	0
SSO	2,035	0	3,000	0
RSP	2,034	1	3,000	0
SBC	2,035	0	3,000	0

E.2 Optimizer Configurations and Run Qualifications

The matrix optimizers act on the attention query, key, value, and output projections and the two feedforward matrices in each Transformer block. Other parameter groups use AdamW; the AdamW baseline also uses AdamW for the matrix group. Table 6 lists the nominal matrix learning rates and matrix weight-decay coefficients recorded in the main runs. Auxiliary learning rates are 10^{-4} for BERT and 10^{-3} for ViT. Auxiliary decay coefficients are 0.01 and 0.05, respectively, with designated bias, normalization, and other no-decay parameters excluded by the implementation. AdamW uses moment coefficients $(0.9, 0.999)$ and $\epsilon = 10^{-8}$. Equal nominal learning rates across optimizers do not imply equal parameter displacements.

SBC, RSP, and the controlled BERT Muon-NS5 baseline use momentum coefficient 0.9 without Nesterov momentum and the five-step direction defined in Section 3.1. RSP uses shrinkage strength $\lambda = 0.05$ and ceiling multiplier $\kappa = 1.1$. The zero matrix-decay entries for RSP do not remove its trial-weight shrinkage. SBC uses the sphere update and acceptance rule in Section 3.3.

ViT's Muon baseline uses the installed native Muon implementation, with momentum 0.95, Nesterov momentum, five Newton–Schulz steps, and numerical epsilon 10^{-7}; it does not use the controlled NS5 direction's additional spectral-radius gain. SSO uses the repository's common-initialization-radius adaptation, with momentum 0.9, Nesterov enabled in its own direction routine, eight Newton–Schulz steps, five power iterations, solver tolerance 10^{-6}, at most 20 solver iterations, and at most ten bracket expansions. This is not the original paper's complete training recipe. Configuration fields for inactive method branches do not define additional operations for the selected optimizer.

Table 7 distinguishes attempted batches from successful optimizer updates. The BERT Muon record contains 644 overflow skips; its separately failed numerical-fix rerun is not substituted for the reported endpoint. SBC's BERT initialization artifact has a different hash from the other runs. A shared seed therefore does not establish that its initial tensors are identical. SBC resumed after step 1,500 and subsequently distributed optimizer matrix work over eight GPUs while the training world size remained one. Its BERT wall-clock total is consequently not used as a matched-hardware comparison with the other main runs.

All five ViT runs share initialization and data manifests. Their AdamW learning rate is fixed at 10^{-3} rather than selected through a comprehensive search. The reported result therefore characterizes this configuration, not the best attainable performance of AdamW. The single-seed comparisons also do not estimate across-seed variability.

Table 8: Qualified 200-update runs in the width-transfer grids.

Method	256	512	1024	2048	Total
AdamW	20	22	22	23	87
Muon-NS5	8	8	8	8	32
SSO	10	8	10	10	38
NR	8	8	8	8	32
RSP	8	8	8	8	32
SBC	10	9	8	7	34
Total	64	63	64	64	255

E.3 Width-Transfer Protocol

The width study uses 24-layer BERT at hidden widths 256, 512, 1,024, and 2,048. Feedforward width is four times the hidden width, and attention-head dimension is 64. Included runs use seed 830101, sequence length 512, global batch size 384, microbatch size 32, FP16, 200 successful updates, and zero overflow. Each run presents 76,800 sequences, or 39,321,600 input positions. Validation uses the same frozen set of 8,192 examples. The auxiliary learning rate is 10^{-4}; training includes NSP, but learning-rate selection and transfer metrics use validation MLM loss only.

Within each width, methods share the recorded `spectral_s1_v1` initialization artifact. For each optimized matrix, $S = \|W_0\|_2$ and $F_0 = \|W_0\|_F$ are obtained from that width's initialized matrix and then kept as reference quantities. They are not copied from the width-256 model. The initialization and evaluation manifests are checked within each width. Different historical runs can nevertheless have different recorded implementation versions; they are not retrospectively described as executions of a later patched version.

Let o index an optimizer, w a hidden width, and $\mathcal{G}_{o,w}$ its qualified learning-rate grid. Write $\mathcal{V}_{o,w}(\eta)$ for the recorded validation MLM loss after 200 successful updates at nominal matrix learning rate η. We select a source minimizer

$$\widehat{\eta}_o \in \arg\min_{\eta \in \mathcal{G}_{o,256}} \mathcal{V}_{o,256}(\eta)$$

and reuse this full-precision value unchanged at widths 512, 1,024, and 2,048. In particular, the procedure does not retune at width 512 before moving to width 1,024. For a target width, the transfer penalty and its mean are

$$\begin{aligned} P_{o,w} &= \mathcal{V}_{o,w}(\widehat{\eta}_o) - \min_{\eta \in \mathcal{G}_{o,w}} \mathcal{V}_{o,w}(\eta), \\ \overline{P}_o &= \frac{1}{3} \sum_{w \in \{512,1024,2048\}} P_{o,w}. \end{aligned} \tag{E.1}$$

All three transferred runs must be available for an optimizer to enter this comparison. Target-width minima serve only as references for scoring; target losses are not used to choose $\widehat{\eta}_o$.

The qualified grids contain 255 runs (Table 8). The numbers of trials and their learning-rate ranges are not equal across methods. A grid-search minimum is the lowest observed loss among the qualified settings, not the optimum over a continuous learning-rate domain. Full-precision rates and losses are retained in the accompanying grid data.

E.4 Validation Crossings and Timing

Let $\mathcal{L}_i$ be a recorded validation loss at step t_i, where $t_0 < \cdots < t_N$. For a target $L_\star$ and a downward-crossing segment satisfying $\mathcal{L}_{i-1} > L_\star \geq \mathcal{L}_i$, define

$$\widehat{t}(L_\star) = t_{i-1} + (t_i - t_{i-1}) \frac{\mathcal{L}_{i-1} - L_\star}{\mathcal{L}_{i-1} - \mathcal{L}_i}. \tag{E.2}$$

We report the first crossing of the piecewise-linear curve, round the resulting step to the nearest integer, and compute relative reductions from the unrounded estimates. For comparison against a baseline with a finite crossing estimate, the reduction is $1 - \widehat{t}_{\text{method}}/\widehat{t}_{\text{baseline}}$.

Table 9: Observed and interpolated ViT crossings of Muon's final validation loss.

Method	Recorded bracket	First observed	Estimated
Muon	$(2750, 3000]$	3000	3000
SSO	$(1750, 2000]$	2000	1906
RSP	$(2000, 2250]$	2250	2027
SBC	$(1750, 2000]$	2000	1768

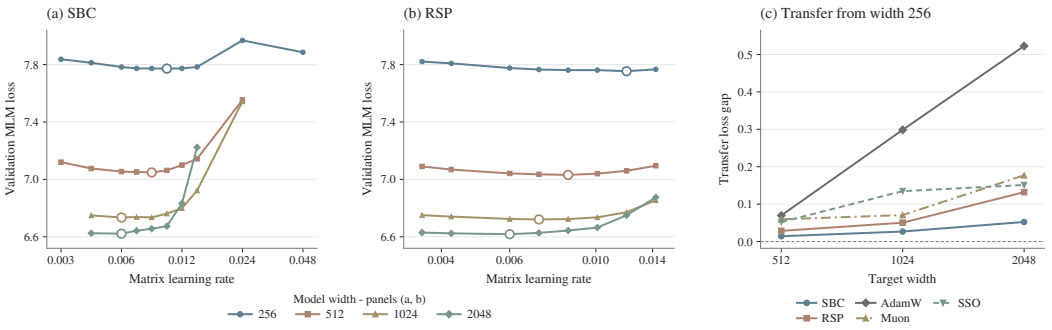

Figure 5: Learning-rate sweeps across widths for SBC (a) and RSP (b), and excess validation MLM loss after learning-rate transfer from width 256 (c).

The targets are post-hoc summaries of the recorded trajectories. Interpolation neither supplies an evaluation at the estimated step nor guarantees that the loss remains below the target between checkpoints. If the first evaluation is already below the target, only attainment by that evaluation is established; we do not extrapolate backward. If no recorded point reaches the target, the crossing is unavailable and is marked "–". This does not exclude an unobserved, temporary crossing.

For Muon's final ViT loss, $L_\star = 1.14657785987854$, Table 9 shows the actual recorded brackets and the interpolated estimates used in the main text. The first-observed crossings for SBC and RSP are 2,000 and 2,250 updates; these differ from the interpolated estimates and are not interchangeable metrics.

Table 3 reports measured 200-step single-V100 acceptance averages for Muon, SSO, and RSP. The SBC entry of 119.60 seconds per step is estimated and is not a directly measured V100 average. These timing references are separate from the 2,035-step BERT quality experiment and the 3,000-step ViT experiment. They are not combined with update reductions from another task or hardware setting to claim an end-to-end speedup. The arithmetic analysis in Appendix D.4 is also separate.

F Additional Experimental Results

F.1 Learning-Rate Sweeps and Transfer Diagnostics

Figure 5 shows the measured learning-rate sweeps for SBC and RSP at all four widths, together with source-256 transfer penalties for the five plotted methods. Table 4 additionally includes NR. The selected learning rates are reused without target-width retuning according to Appendix E.3.

Table 10 reports the absolute MLM losses at the same source-selected rates. It complements penalties measured relative to each optimizer's own grid minimum: a smaller penalty alone does not imply a smaller absolute loss. RSP has the lowest observed absolute loss at widths 256 and 512 in this table, whereas SBC has the lowest at widths 1,024 and 2,048. These are transferred-rate results, not the results of selecting a separate optimum at every width.

Strict source-width minimization can be sensitive to small source-loss differences. For a candidate rate η, define its source excess as

$$\mathcal{V}_{o,256}(\eta) - \min_{\xi \in \mathcal{G}_{o,256}} \mathcal{V}_{o,256}(\xi).$$

Table 10: Absolute validation MLM loss using each optimizer's width-256-selected learning rate at all four widths.

Method	256	512	1024	2048
AdamW	7.8793	7.5597	7.5583	7.5626
Muon-NS5	7.7690	7.1064	6.8033	6.8003
SSO	7.7692	7.1170	6.8780	6.7604
NR	7.7585	7.0679	6.7769	6.7475
RSP	**7.7540**	**7.0599**	6.7713	6.7501
SBC	7.7720	7.0626	**6.7613**	**6.6742**

Table 11: Sensitivity to alternative source-width learning rates.

Method / rate	Learning rate	Source excess	Mean penalty
SSO, selected	0.00848528	0	0.1126823
SSO, alternative	0.00600000	0.0002184	0.0158565
SBC, selected	0.01009076	0	0.0311017
SBC, alternative	0.00848528	0.0008723	0.0117181

Table 11 compares the selected rates with illustrative alternatives already present in the recorded grids. For SSO, increasing source loss by only 0.0002184 changes the mean transfer penalty from 0.1126823 to 0.0158565. SBC likewise has an alternative with source excess 0.0008723 and mean penalty 0.0117181. These post-hoc checks do not replace the source-only selection rule using information from the targets. They show why the main ranking is conditional on the specified grids and selection rule rather than a guarantee for every near-optimal source rate.

F.2 SBC Ablation Protocol and Early-Training Results

The mechanism experiment uses CIFAR-100 (Krizhevsky, 2009), split into 45,000 training and 5,000 validation images with split seed 830000. The ViT uses 32×32 images, patch size four, six layers, hidden width 192, three heads, and feedforward ratio four. We use BF16, global batch size 256, and microbatch size 64. Each epoch contains 175 updates and 44,800 training examples after dropping the incomplete final batch.

The compared methods are SBC, SBC-K1, and NR. SBC-K1 fixes the selected spectral block size to one. NR uses the common Muon proposal without block correction while retaining radial restoration and the final-displacement check. Within each condition, all methods use matrix learning rate 0.02, auxiliary learning rate 0.001, and the same training settings. Original uses seed 831001 and a 50-epoch schedule without Mixup. Mixup uses seed 841001, mixing coefficient 0.2, and a 30-epoch schedule. Both use five warmup epochs and a final learning-rate ratio of 0.01. The conditions have different seeds and schedules, so their cross-condition differences are not attributed solely to Mixup.

The two conditions and the first-25-epoch display were selected exploratorily. Figure 6 shows validation and epoch-mean training losses over this window; Table 12 reports validation loss at epoch 25. The displayed trajectories retain their original learning-rate schedules and are not new runs trained with a 25-epoch schedule.

SBC has the lowest validation loss at epoch 25 in both conditions. Later validation crossings change the ordering, so this early-training result does not establish terminal superiority. These experiments isolate variants of SBC; they do not provide an RSP ablation.

F.3 Input-Response Probe Construction

We use the empirical input-response energy from Section 3.5 and Appendix D.3. For probe p, let $W_p \in \mathbb{R}^{m_p \times n_p}$ be the current weight matrix, and let $\mathsf{H}_p \in \mathbb{R}^{n_p \times N_p}$ collect its sampled input vectors.

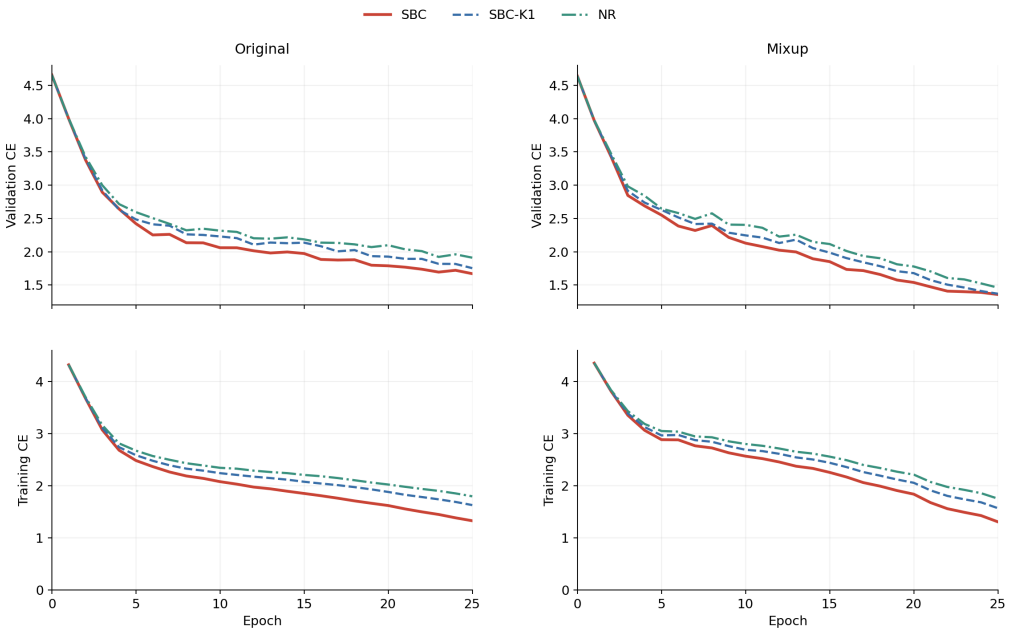

Figure 6: SBC ablations over the first 25 epochs: Original (left) and Mixup (right), with validation cross-entropy above and epoch-mean training cross-entropy below.

Table 12: Validation cross-entropy at epoch 25 for the two ablation conditions.

Method	Original	Mixup
NR	1.9106	1.4607
SBC-K1	1.7531	1.3689
SBC	**1.6700**	**1.3574**

For a parameter change A of the same shape as W_p, define

$$e_p(A) = \frac{\|A\mathsf{H}_p\|_F^2}{m_p N_p}.$$

Let a_p denote the final attempted trial step and R_p its corrected, norm-bounded direction. The compared changes are

$$A_p^{\text{trial}} = -a_p R_p, \qquad A_p^{\text{exec}} = W_p^+ - W_p,$$

where W_p^+ denotes the stored weight after the update. Thus the trial change is measured after direction correction and normalization but before radial restoration. It is not the raw requested change $-\alpha_p D_p$.

For each method and condition, 16 fixed validation images are used to capture layer inputs in evaluation mode with FP32 arithmetic. We probe the query, feedforward-up, and feedforward-down matrices in layers 0, 3, and 5, giving nine matrices per probe step. At most 128 input vectors are retained for each matrix. Diagnostic responses and energies are computed in FP64. The four included probe steps are 876, 1,751, 2,626, and 3,501; these give 36 matrix probes per method and condition. The initialization probe is excluded, and the next probe at step 4,376 falls outside the displayed 25-epoch window of 4,375 updates.

For positive trial energy, the individual ratio and its reported mean over the included probe set $\mathcal{I}$ are

$$r_p = \frac{e_p(A_p^{\text{exec}})}{e_p(A_p^{\text{trial}})}, \qquad \bar{r} = \frac{1}{|\mathcal{I}|}\sum_{p\in\mathcal{I}} r_p. \tag{F.1}$$

Table 13: Mean post-retraction input-response energy as a percentage of corrected-trial energy over the selected probes.

Condition	SBC	SBC-K1	NR
Original	99.9263	97.6529	94.0986
Mixup	99.9508	98.0675	96.1370

Table 14: Recorded constraint-execution statistics for SBC in the two main training runs.

Statistic	BERT	ViT
Matrix updates	293,040	432,000
Rejected matrix updates	0	0
Updates with $k > 1$ (%)	99.2434	96.3470
Mean k	137.3827	22.8945
Mean k/d	0.1342	0.0224
Mean candidate count	1.9447	1.7598
Mean trial/requested step	0.9055	0.9240
Maximum relative radius error	5.95×10^{-8}	9.15×10^{-8}
Maximum displacement/budget	1.00000080	1.00000095

This is an unweighted mean of per-probe ratios, not a ratio of pooled energies. A zero trial-energy denominator does not define a ratio and must not be assigned a perfect-retention value. The reported percentages are $100\bar{r}$. Small points in Figure 4(c) average the nine matrix ratios at each probe step; they are not independent training seeds or confidence intervals.

These measurements quantify the additional response-energy change at radial restoration. They do not measure total retention relative to the raw Muon proposal, identify directional agreement, or establish that greater response energy causes lower validation loss. The image identities are fixed across methods, but the internal activations depend on each trained model. Consequently, the reported values are trajectory-level measurements rather than matched-weight, matched-momentum, matched-input comparisons.

F.4 Constraint Execution in the Main SBC Runs

The main SBC runs additionally record selected block sizes, candidate counts, rejection counts, and numerical constraint checks. Table 14 aggregates these logs over matrix updates. The local dimension is $d = \min(m, n)$ for each matrix, and k is the final selected block size. A candidate count includes the first trial and any subsequent backtracking trials. A rejected matrix update means that no candidate is accepted and the stored weight is retained; this is different from an external mixed-precision overflow skip.

The radius error is the recorded value of $|\|W^+\|_2 - S|/S$, and the displacement ratio is the recorded value of $\|W^+ - W\|_2/\alpha$ for a positive requested budget. The small excesses above one are at the configured floating-point tolerance scale described in Appendix B.4; they are not exact-arithmetic violations of the ideal acceptance rule. Likewise, these numerical measurements are not certified interval bounds on matrix norms.

The high fraction of $k > 1$ updates shows that the multi-direction mechanism is exercised during both main runs. It does not, by itself, establish that selecting more directions improves loss. The counts aggregate operations along two trajectories and are not independent experimental repetitions.