LoRA Initialization Geometry and Learning-Rate Transfer

Anonymous authors
Paper under double-blind review

Abstract

LoRA parameterizes weight updates as a product of low-rank matrices, making the geometric constraints imposed by this structure important for understanding its training dynamics. For nonzero rank-r initial adapter products, initialization determines the initial tangent space and local curvature, constraining early weight changes and their departure from that space. We treat the zero-product initialization of vanilla LoRA separately and analyze how initialization geometry interacts with the learning rate to shape training. For rank-r initializations, we derive bounds on subspace rotation and local departure from the initial tangent space using the initial spectrum and cumulative movement. In our experiments, the loss gap between tangent-constrained and unconstrained LoRA remains small even as tangent-space departure varies substantially. We analyze how initialization scale and the optimizer determine first-order weight update magnitudes, and derive a learning-rate scaling rule that matches these magnitudes across initialization scales. For a fixed optimizer and factor shape, we prove conditions under which the normalized final-loss curves converge over a fixed training horizon, enabling learning-rate transfer with bounded excess loss. Across the tested initialization methods, learning-rate tuning yields similar evaluation performance despite different training geometries. We use these results to guide learning-rate search across initializations and evaluate pilot-based corrections when the optimizer changes.

1 Introduction

LoRA (Hu et al., 2022) has become a standard method for efficiently fine-tuning large language models: it freezes the pretrained weights and trains only the product of two low-rank factors, so the initialization of these factors determines where and in which directions the weight change begins. Vanilla LoRA zeroes one factor and starts from random directions, whereas PiSSA and LoRA-GA initialize the factors from the principal components of the pretrained weights and from the initial gradient, respectively, and report faster convergence or higher performance (Meng et al., 2024; Wang et al., 2024); MiLoRA and LoRA-One likewise use pretrained-weight or gradient information (Wang et al., 2025; Zhang et al., 2025a). Which factor is zeroed changes the training dynamics (Hayou et al., 2024a), and making both factors nonzero changes robustness to the learning rate (Li et al., 2025). Initialization magnitude has been reported to affect update magnitude and performance (Zhang et al., 2025b), and a sufficiently broad learning-rate search substantially narrows the peak-performance gaps among initializations (Lee et al., 2026). Because the learning rate alone does not fix how far LoRA weights move in one step, comparing initializations requires separating the geometry each initialization sets from the update magnitude it induces.

We show that geometric differences among LoRA initializations explain little of the performance they reach, since training restricted to the initial tangent space nearly matches unconstrained LoRA and most initialization–optimizer pairs perform similarly after per-initialization learning-rate tuning, whereas scaling both factors by a scales the first-order update by a^{1+p} but the support rotation rate by a^{p-1}, so an initialization advantage observed at a fixed learning rate is confounded with a difference in effective learning rate. To control for this, we derive the learning-rate rule $\eta \propto a^{-(1+p)}$, which matches the first-order weight change. For magnitude-only changes with the optimizer and initial factor shapes fixed, we prove conditions under which the finite-horizon final-loss curve and the optimal learning rate transfer after this normalization; in AdamW experiments varying only the initial magnitude, the normalized optimal learning rates lie within a factor of about 1.5. When the

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.

initial directions or the optimizer differ, we analyze how far the one-step optimal learning rate can shift and evaluate cross-optimizer transfer using short pilot runs. Our contributions are as follows:

- **Geometry.** We bound support rotation and tangent-space departure by the budget utilization, and observe small loss gaps and similar tuned performance despite large differences in departure.
- **Update magnitude.** We show that initialization magnitude and optimizer response order scale the first-order update and the support rotation with different exponents, which same-learning-rate comparisons confound.
- **Learning-rate transfer.** We derive $\eta \propto a^{-(1+p)}$ with a conditional finite-horizon transfer theorem and correction ranges for direction or optimizer changes, and propose using it to set each initialization's learning-rate search while reporting budget utilization.

2 Preliminaries: LoRA on the Fixed-Rank Manifold

We study LoRA through its adapter product. When that product has rank r, its path lies on the rank-r matrix manifold. The geometric results below require a rank-r initial product; we treat the zero-product vanilla LoRA initialization separately. This section introduces the notation and observables used to describe these dynamics (see Appendix E for the complete notation).

2.1 LoRA Parameterization and Factor Updates

LoRA is attached to several linear layers; we describe one of them and define every quantity per layer. For a layer with pretrained weight $W_{\text{pre}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, factors $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and a fixed scale $s > 0$, let $X_0 := sB_0A_0$. The adapter product and the effective weight are

$$W = W_{\text{frozen}} + X, \qquad X = sBA, \qquad W_{\text{frozen}} := W_{\text{pre}} - X_0, \tag{1}$$

so that training starts from W_{pre} even for nonzero initialization. The subscript 0 denotes initialization, t the step index, and $+$ the next iterate.

Let $G := \nabla_W \mathcal{L}$ be the weight gradient. The factor gradients are $G_B = sGA^\top$ and $G_A = sB^\top G$, and the optimizer maps them, together with its internal state z, to signed update directions $(D_B, D_A) = \Psi(G_B, G_A; B, A, z)$. For SGD, $(D_B, D_A) = (-G_B, -G_A)$. With branch learning rates $\eta_B, \eta_A > 0$, both directions are computed at the same pre-step state and applied simultaneously as $B^+ = B + \eta_B D_B$ and $A^+ = A + \eta_A D_A$. The product change is exactly

$$X^+ - X = s\big(\eta_B D_B A + \eta_A B D_A + \eta_B \eta_A D_B D_A\big). \tag{2}$$

With a common learning rate η, this reads $X^+ - X = \eta F + \eta^2 Q$, where $F := s(D_B A + B D_A)$ is the first-order product response and $Q := sD_B D_A$ is the cross term. The same expansion holds cumulatively: with $\Delta B := B - B_0$ and $\Delta A := A - A_0$, the weight displacement $\Delta W := W - W_{\text{pre}} = X - X_0$ equals $s(\Delta B\, A_0 + B_0\, \Delta A) + s\, \Delta B\, \Delta A$, a term linear in the factor changes plus their cumulative cross term.

2.2 Tangent-Space Decomposition

For a rank-r matrix $X = U\Sigma V^\top$, let $P_U = UU^\top$ and $P_V = VV^\top$ denote the orthogonal projectors onto its *left support* and *right support*, respectively. These support subspaces represent the output and input directions used by X and coincide with $\operatorname{col} B$ and $\operatorname{row} A$ when both factors have rank r. The set of rank-r matrices, $\mathcal{M}_r$, forms a smooth manifold, and its tangent space at X, denoted by T_X, has dimension $r(d_{\text{out}} + d_{\text{in}} - r)$ (Vandereycken, 2013) (see Appendix E for the definition). The tangent space T_X consists of all first-order changes induced by small rotations of the support directions or changes in magnitude. When X_0 has rank r, we write $T_0 := T_{X_0}$ and call it the *initial tangent space*.

Lemma 1 (Initial tangent space)**.** *Let $J_0[Z_B, Z_A] = s(Z_B A_0 + B_0 Z_A)$ be the initial factor Jacobian. If both B_0 and A_0 have rank r, then $\operatorname{im} J_0 = T_0$. Consequently, $\Delta W - s\, \Delta B\, \Delta A \in T_0$.*

Changing B affects only the output directions, whereas changing A affects only the input directions. Consequently, first-order changes in the factors remain within T_0, and changes that simultaneously

introduce new output and input directions arise only from the cross term involving both factors. For vanilla LoRA with $B_0 = 0$, $\operatorname{im} J_0$ reduces to the space in which the input directions are restricted to $\operatorname{row} A_0$. Proofs for both cases are provided in Appendix B.

For a nonzero rank-r initialization X_0, we decompose the displacement into four blocks using the singular vectors of $X_0 = U_0 \Sigma_0 V_0^\top$ and their orthogonal complements $U_{0\perp}$ and $V_{0\perp}$. The *core* $C = U_0^\top \Delta W \, V_0$ represents changes within the initial supports. The two *arms* $L = U_{0\perp}^\top \Delta W \, V_0$ and $R = U_0^\top \Delta W \, V_{0\perp}$ represent changes that rotate the output and input directions, respectively. The block $N = U_{0\perp}^\top \Delta W \, V_{0\perp}$ represents changes that introduce new directions on both sides. The first three blocks constitute T_0 (i.e., $T_0 = \{N = 0\}$), whereas N is the *departure component* lying outside T_0. By Lemma 1, $N = s \, U_{0\perp}^\top \Delta B \, \Delta A \, V_{0\perp}$, so departure from the initial tangent space arises only through the cross term.

We summarize the cumulative change using three observables. We measure support rotation by the sines of the principal angles, $\sin \Theta(U, U_0)$, between $\operatorname{col} U$ and $\operatorname{col} U_0$, and analogously for the right support. The *departure ratio* and the *normalized displacement* are

$$\varepsilon = \frac{\|N\|_F^2}{\|\Delta W\|_F^2}, \qquad x = \frac{\|\Delta W\|_F}{\sigma_r(X_0)}. \tag{3}$$

We set $\varepsilon := 0$ when $\Delta W = 0$. Here, ε is the fraction of the displacement energy lying outside the initial tangent space, while x measures the displacement in units of the smallest singular value of the initial adapter product. These observables are defined per layer; run-level ε and x^2/r aggregate the layerwise values using $\|\Delta W\|_F^2$ weights (Appendix C).

2.3 Initialization Scale and Optimizer Homogeneity

Practical initialization schemes generally differ in both direction and magnitude. For example, PiSSA (Meng et al., 2024) places greater weight on the leading singular directions of the pretrained weights, whereas vanilla LoRA (Hu et al., 2022) uses random directions with $B_0 = 0$. To separate these effects, fix a reference factor pair $(\bar{B}_0, \bar{A}_0)$ with $\|\bar{B}_0 \bar{A}_0\|_F = \sqrt{r}$ and $\bar{X}_0 := s \bar{B}_0 \bar{A}_0$ of rank r, and vary only the *common factor amplitude* $a > 0$ through $(B_0, A_0) = (a\bar{B}_0, a\bar{A}_0)$. Then $X_0 = a^2 \bar{X}_0$: its singular vectors and initial tangent space T_0 stay fixed, while $\sigma_r(X_0) = a^2 \sigma_r(\bar{X}_0)$. Vanilla LoRA is covered by the factor-Jacobian description above.

We measure the initial adapter-product magnitude by

$$\mu_0 := \frac{\|X_0\|_F}{\sqrt{r}} = s \frac{\|B_0 A_0\|_F}{\sqrt{r}}. \tag{4}$$

The magnitude μ_0 scales as a^2 along this family.

The optimizer's response to factor magnitude can be characterized by a single exponent. For a fixed weight gradient G, scaling both factors by a scales the factor gradients to (aG_B, aG_A). Suppose there exists a correspondence between optimizer states, $z \mapsto z_a$, that is preserved after one update step and satisfies

$$\Psi(aG_B, aG_A; aB, aA, z_a) = a^p \, \Psi(G_B, G_A; B, A, z). \tag{5}$$

We call p the optimizer's *response order*.

Lemma 2 (Response Order)**.** *Consider an idealized implementation in which the state correspondence persists after one step, the stabilizing constant, weight decay, and gradient clipping are removed, and all normalization denominators are defined. Then the response order is $p = 1$ for SGD (including momentum), $p = 0$ for Adam, and $p = -1$ for LoRA-Muon and PoLoRA. The Gram inverse square roots and polar inputs in LoRA-Muon must be nondegenerate; PoLoRA's curvature state must be scale covariant, and its direction-norm floor must be inactive.*

For SGD, the update to each factor is proportional to the other factor and therefore scales by a. For Adam, elementwise normalization with positive moment denominators cancels the scaling of the gradient, leaving the update unchanged. Idealized LoRA-Muon and PoLoRA normalize their updates by the factor magnitude, causing the updates to scale by $1/a$. The proof is provided in Appendix B. Implementations with damping, floors, decay, or clipping require checking the corresponding response maps. Here, the response is the direction before the additive factor update; product-preserving rebalancing is a separate map applied afterward (Lemma B.8 in the appendix).

Table 1: First-order scaling under $(B, A) \mapsto (aB, aA)$ (Theorem 1).

Optimizer	p	Fixed η		η_{step}	η_{rot}
		Product change	Support rotation, x		
SGD	1	a^2	a^0	a^{-2}	a^0
Adam	0	a	a^{-1}	a^{-1}	a
LoRA-Muon, PoLoRA	-1	a^0	a^{-2}	a^0	a^2

Table 2: LR-tuned rank-8 PiSSA/LoRA-Muon. ΔCE: T_0-restricted minus LoRA; ε: LoRA's final departure from T_0.

β_{eff}	LoRA CE	T_0 CE	ΔCE	ε
0.099	0.16617	0.16881	+0.00264	14%
0.040	0.16530	0.16862	+0.00332	66%
0.020	0.16554	0.16865	+0.00311	88%

3 How Initialization Constrains the Trajectory

Section 2 showed that a rank-r initialization determines the space T_0 of first-order weight changes. This section derives two constraints on training: the initial spectrum and weight displacement bound departure from T_0 (Section 3.1), and the common factor amplitude scales the first-order product response and instantaneous support rotation with different exponents (Section 3.2). Section 3.3 combines them into bounds expressed through the cumulative product path length and the initial singular values.

3.1 Departure from the Initial Tangent Space

Expressing the current product in block form with respect to the initial coordinates gives

$$X = X_0 + \Delta W = [U_0 \quad U_{0\perp}] \begin{bmatrix} K & R \\ L & N \end{bmatrix} [V_0 \quad V_{0\perp}]^\top, \qquad K := \Sigma_0 + C. \tag{6}$$

Proposition 1 (Schur Representation of Departure). *If $\gamma := \sigma_r(X_0) - \|C\|_2 > 0$, then*

$$N = LK^{-1}R, \qquad \|N\|_F \le \frac{\|\Delta W\|_F^2}{2\gamma}$$

and consequently $\varepsilon \le x^2/(4\bar{\gamma}^2)$. Here $\bar{\gamma} := \gamma/\sigma_r(X_0)$.

Because the rank of X cannot exceed r, the departure component N cannot vary independently and is determined by the two arms and the core. When the displacement is small, so that $K \approx \Sigma_0$, we have $N \approx L\Sigma_0^{-1}R$. Thus, departure is governed by the product of the two arms and is therefore second order in the displacement. Geometrically, $L\Sigma_0^{-1}R$ is the quadratic term of the local Schur graph of $\mathcal{M}_r$ at X_0. When both orthogonal complements exist, $r < \min(d_{\text{out}}, d_{\text{in}})$, the maximum normal curvature is $1/\sigma_r(X_0)$, so $\sigma_r(X_0)$ is the local curvature length scale. The second fundamental form is twice this quadratic term; Lemma B.2 gives its bilinear expression and proves attainment of the curvature bound, connecting the Schur representation to the geometry of fixed-rank matrices (Feppon & Lermusiaux, 2018). The initial spectrum therefore controls the departure bound. When the core change is small relative to $\sigma_r(X_0)$, $\bar{\gamma} \approx 1$ and the upper bound on ε is approximately $x^2/4$. Thus, for $x \ll 1$ with the core gap maintained, training remains close to T_0.

3.2 Two Effective Learning Rates

The common factor amplitude affects the first-order product response and instantaneous support rotation differently. For $X = U\Sigma V^\top$ with learning-rate-free response F, the component determining the derivative of the left support projector is $(I - P_U)FV\Sigma^{-1}$ (Lemma B.3). Scaling both factors by a scales the singular values by a^2. At the same first-order product response, the support derivative therefore scales by a^{-2}.

Theorem 1 (Two Effective Learning Rates). *Suppose $X = sBA$ has rank r, the corresponding states (B, A, z) and (aB, aA, z_a) receive the same weight gradient G, and the response order in Equation 5 is p. The learning-rate-free first-order product response F satisfies $F_a = a^{1+p}F$, while the derivatives at $\eta = 0$ of the left and right support projections are each multiplied by a^{p-1}. Thus the learning rates matching the nonzero first-order response and nonzero instantaneous rotation,*

respectively, are $\eta_{\text{step}} = a^{-(1+p)}\eta$ *and* $\eta_{\text{rot}} = a^{1-p}\eta$*, with ratio* a^2*. The finite-step expansion is* $X_a^+ - X_a = \eta a^{1+p}F + \eta^2 a^{2p}Q$.

The relative factor between the two first-order scales is a^2 for every response order p (Table 1). To first order, the first-step normalized displacement x follows the product-response scale divided by the initial singular value. When $a \neq 1$ and both responses are nonzero, a single learning rate cannot match both the first-order product response and instantaneous support rotation. This is analogous to the angular effective learning rate $\eta/\|w\|^2$ for scale-invariant weights (Hoffer et al., 2018); here it arises from the two LoRA factors and their response order p. With a common exogenous gradient sequence and covariant optimizer states, two runs matched using η_{rot} satisfy $X_k^{(a)} = a^2 X_k$ at every step and undergo identical support rotations (Proposition B.1). In training with live gradients, residualization gives the same initial weight gradient, and subsequent gradients depend on each run's trajectory.

3.3 The Curvature Budget

Sections 3.1 and 3.2 showed that both departure from the initial tangent space and support rotation are governed by the product movement normalized by $\sigma_r(X_0)$. This section combines the two quantities into a single variable. Since the learning rate directly determines the distance traveled at each step, we use the cumulative path length instead of the displacement.

$$\Lambda_t := \sum_{j<t} \|X_{j+1} - X_j\|_F, \qquad \xi_t := \frac{\Lambda_t}{\sigma_r(X_0)} \tag{7}$$

We call $\sigma_r(X_0)$ the initial *budget scale* used to normalize the product path length, and ξ_t the *budget utilization*.

Proposition 2 (Budget Bound)**.** *If* X_0 *and, for every* $k \leq t$*,* X_k *have rank* r*, then, for each* $k \leq t$*,*

$$x_k \leq \xi_t, \qquad \max\{\|\sin\Theta(U_k, U_0)\|_2,\ \|\sin\Theta(V_k, V_0)\|_2\} \leq \frac{\|\Delta W_k\|_2}{\sigma_r(X_0)} \leq \xi_t$$

Moreover, if $\gamma_k := \sigma_r(X_0) - \|C_k\|_2 > 0$*, then* $\varepsilon_k \leq \xi_t^2/(4\bar{\gamma}_k^2)$*. Here* $\bar{\gamma}_k := \gamma_k/\sigma_r(X_0)$*. In particular, if* $\xi_t < 1$*, then for every* $k \leq t$ *we have* $\gamma_k > 0$ *and* $\varepsilon_k \leq \xi_t^2/[4(1-\xi_t)^2]$.

The displacement bound follows from the triangle inequality, and the support bound follows directly from projector identities for rank-r products. Combining these with the Schur bound proves Proposition 2 (Appendix B.2). Small ξ_t bounds support rotation linearly and the departure ratio quadratically.

To bound budget utilization across amplitudes, suppose the actual product steps satisfy $\|X_{k+1}^{(a)} - X_k^{(a)}\|_F \leq M a^{1+p}\eta_k^{(a)}$ for every $k < T$, with M independent of a. This pathwise response bound gives

$$\Lambda_T^{(a)} \leq M a^{1+p} \sum_{k<T} \eta_k^{(a)}, \qquad \xi_T^{(a)} \leq \frac{M a^{p-1}}{\sigma_r(\bar{X}_0)} \sum_{k<T} \eta_k^{(a)}. \tag{8}$$

For step-matched learning rates $\eta_k^{(a)} = a^{-(1+p)}\bar{\eta}_k$, this becomes $\xi_T^{(a)} \leq M \sum_{k<T} \bar{\eta}_k/(a^2\sigma_r(\bar{X}_0))$. The constant M captures the response directions and optimizer states along the path. Under the exogenous-gradient and rotation-matching conditions of Proposition B.1, the exact relation is $\xi_T^{(a)} = \xi_T$ (Corollary 1).

For $0 < \theta < \pi/2$, the condition $\xi_T < \sin\theta$ keeps the largest principal angles from both initial supports below θ for every $k \leq T$. The pathwise bound above gives a sufficient condition for this retention before training. Vanilla LoRA has $X_0 = 0$ and is treated through its factor Jacobian, as in Section 2.2. The ratio $\xi_T = \Lambda_T/\sigma_r(X_0)$ thus connects the realized product path to bounds on geometric change. We next examine how this geometry relates to attained performance.

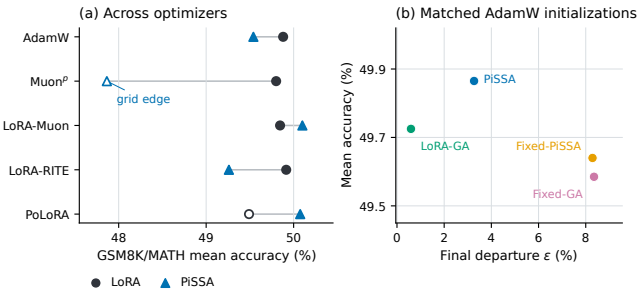

Figure 1: (a) Best GSM8K/MATH mean accuracy over the learning-rate grid for rank-128 LoRA and PiSSA with five optimizers. Open markers indicate LoRA–PoLoRA and PiSSA–Muon^p, whose selected learning rates lie at the boundaries of the grid. (b) Accuracy and final departure ratio from the initial tangent space for the most accurate run among three learning rates for each of four rank-128 AdamW initializations with matched mean initial factor magnitude. The two panels report results from separate learning-rate searches.

4 Tuned Performance and Initial Geometry

4.1 Tuned Performance across Initializations and Optimizers

We fine-tuned Qwen3-0.6B-Base (Yang et al., 2025) on 100K MetaMathQA examples (Yu et al., 2024) and evaluated on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). Across 10 combinations of rank-128 LoRA/PiSSA with AdamW (Loshchilov & Hutter, 2019), Muon^p (Dong & Sawin, 2026), LoRA-Muon (Cesista et al., 2026), LoRA-RITE (Yen et al., 2025), and PoLoRA (Ghosh et al., 2026), the grid-best mean accuracy of 9 ranged from 49.3 to 50.1% (Figure 1(a)); Muon^p–PiSSA reached 47.9% at the lower grid boundary, so its peak remains unresolved.

We also compared four AdamW initializations: PiSSA, LoRA-GA, and their fixed-basis counterparts, which use discrete sine-transform directions (Zhang et al., 2025b) with the layerwise initial product magnitudes of PiSSA or LoRA-GA. Each was rescaled by a common multiplier to match mean factor RMS across 196 adapters, controlling factor magnitude while allowing different product spectra. Their best accuracy over three learning rates ranged only from 49.59 to 49.87%, whereas the final departure ratio ε ranged from 0.59 to 8.35% (Figure 1(b)).

4.2 Training Restricted to the Initial Tangent Space

For rank-8 PiSSA with LoRA-Muon, we compared unconstrained LoRA with a model trained using only $\Delta W_{T_0} = s(\Delta B\, A_0 + B_0\, \Delta A) \in T_0$. This model allows changes in the initial core and the two arms but excludes the cross term between the factor changes. At the original initialization magnitude and under the same learning rate, the cross-entropy (CE) of the restricted model was higher by 0.0014, while that of a model trained only on the core of the initial support was higher by 0.121. Even after searching the learning rate separately for each model at three relative factor amplitudes, $a/a_{\text{orig}} = 1, 0.403, 0.201$, the CE of the restricted model remained higher by 0.00264–0.00332 (Table 2). At all three initialization magnitudes, both models attained their lowest CE at the same learning rate within the interior of the grid ($\eta \simeq 0.0312$). Although ε for unconstrained LoRA ranged from 14 to 88%, it had no monotonic relationship with the CE gap. Because restricting the model also changes its subsequent update trajectory, this gap cannot be attributed solely to the normal component.

5 Learning-Rate Transfer across Initializations

We now consider how a learning rate identified for one initialization can be used to guide the search for another. A reparameterization that preserves the initial product can be absorbed exactly under suitable conditions, whereas changing the magnitude of the product itself also changes the training dynamics. In the latter case, transferring the optimal learning rate requires starting from the scale of

the product step and controlling the discrepancy between the resulting final-loss curves. Additional corrections are required when the initial directions or the optimizer also change.

5.1 Reparameterization Invariance

First, consider the case in which the same initial product is represented using different factor magnitudes and LoRA scales. We assume that the optimizer for each factor does not depend directly on the factor itself, but determines its update from the gradient and internal state. Let q_{g} denote the order such that, when the gradient and its history are scaled by $f > 0$, the update is scaled by $f^{q_{\mathrm{g}}}$ under a corresponding state transformation that is preserved after one step. SGD has $q_{\mathrm{g}} = 1$, while Adam has $q_{\mathrm{g}} = 0$ when its stabilizing constant is removed or transformed accordingly. The order q_{g} describes gradient scaling at fixed factors and is distinct from the common factor-amplitude response order p introduced in Section 2.3. Equation 19 gives the definition, and Lemma B.4 specifies the state transformations for SGD and Adam.

Theorem 2 (Reparameterization Preserving the Initial Product)**.** *Use an optimizer satisfying the gradient–state covariance in equation 19 without weight decay or clipping. If either is used, apply the factorwise transformations in Lemma B.5. For $c_B, c_A > 0$, set*

$$\begin{aligned} B_0' &= c_B B_0, \qquad A_0' = c_A A_0, \qquad s' = \frac{s}{c_B c_A}, \\ \eta_{B,t}' &= c_B^{q_{\mathrm{g}}+1}\eta_{B,t}, \quad \eta_{A,t}' = c_A^{q_{\mathrm{g}}+1}\eta_{A,t} \end{aligned} \tag{9}$$

and transform both factors' optimizer states and any required hyperparameters according to the respective gradient scalings c_B^{-1}, c_A^{-1}. Two runs using the same minibatches and randomness then satisfy $B_t' = c_B B_t$, $A_t' = c_A A_t$, and $X_t' = X_t$ at every step.

This is an exact finite-step equivalence that includes the cross term in equation 2. The proof is provided in Appendix B.3. Fixing the normalized initial factor shapes, write $B = \sqrt{\rho_0} h_0 \bar{B}$ and $A = \sqrt{\rho_0} h_0^{-1} \bar{A}$, where $\rho_0 = \|B_0 A_0\|_F / \sqrt{r}$, $h_0 > 0$ is the initial imbalance, and $\|\bar{B}_0 \bar{A}_0\|_F = \sqrt{r}$. In these coordinates, the magnitude and learning-rate configuration can be expressed as

$$\begin{aligned} \mu_0 &= s\rho_0, \\ \lambda_B &= \eta_B s^{q_{\mathrm{g}}} \rho_0^{(q_{\mathrm{g}}-1)/2} h_0^{-(q_{\mathrm{g}}+1)}, \qquad \lambda_A = \eta_A s^{q_{\mathrm{g}}} \rho_0^{(q_{\mathrm{g}}-1)/2} h_0^{q_{\mathrm{g}}+1}. \end{aligned} \tag{10}$$

For the same normalized shapes, corresponding states, and transformed hyperparameters, matching $(\mu_0, \lambda_B, \lambda_A)$ at every step produces identical weight trajectories (Appendix B.3). These invariants absorb the reparameterization effects of s, ρ_0, h_0. Changing μ_0 instead changes the initial product magnitude and leads to the transfer problem below.

Exact equivalence also requires covariance at the implementation level. For example, when the gradient is scaled by f in Adam, $m, v, \epsilon_{\mathrm{adam}}$ must be scaled by f, f^2, f, respectively. The theorem assumes no weight decay or clipping by default; when these operations are used, the corresponding terms must also be transformed. In particular, two different factor scalings generally cannot be absorbed into a single global clipping threshold. LoRA-Muon and PoLoRA, which depend directly on factor Gram matrices or norms, require separate reparameterization analyses; their $p = -1$ should not be substituted for q_{g} in Equation 10.

5.2 An Update-Scale Transfer Rule

We now fix s and the initial factor shapes and vary only the amplitude. Let η_a be the source learning rate at amplitude a, and let $\eta_{\mathrm{start},b}$ be the search starting point at amplitude b. Since $\mu_{0,a} = a^2 \mu_{0,1}$, Theorem 1 gives

$$\eta_{\mathrm{start},b} = \eta_a \left(\frac{a}{b}\right)^{1+p} = \eta_a \left(\frac{\mu_{0,a}}{\mu_{0,b}}\right)^{(1+p)/2}. \tag{11}$$

This is the size-matched learning rate for a fixed-shape amplitude family. Its amplitude exponent is -2 for SGD, -1 for Adam, and 0 for idealized LoRA-Muon and PoLoRA.

To connect the first-order rule to optimal learning rates, let $\mathcal{L}_{a,\mathcal{O},T}(\eta)$ be the final evaluation loss at learning-rate multiplier η for a fixed training horizon T and schedule. The following theorem controls the difference between the normalized loss curves through regularity of the state dynamics.

Theorem 3 (Finite-Horizon Amplitude Transfer)**.** *Fix the optimizer $\mathcal{O}$, s, T, the normalized initial factor shape, and the minibatch order. Let $\eta = \zeta a^{-(1+p)}$ with $\zeta \in I = [0, Z]$. Assume that the normalized states and responses are bounded in a neighborhood of a common limiting trajectory, and that the one-step state maps are uniformly within $O(a^{-2} + \tau_a)$ of a common Lipschitz limiting map. Here $\tau_a \to 0$. Assume also that the initial-state error is at most τ_a and the evaluation loss is Lipschitz. Suppose that each finite-amplitude and limiting state map is jointly continuous in the state and learning-rate multiplier, and that the initial state is continuous in the learning-rate multiplier. Under the concrete state-map conditions in Appendix B.3, for sufficiently large a,*

$$\sup_{\zeta \in I} \left| \mathcal{L}_{a,\mathcal{O},T}(\zeta a^{-(1+p)}) - \Phi_{\mathcal{O},T}(\zeta) \right| \le \delta_a = O_{T,I}(a^{-2} + \tau_a). \tag{12}$$

For sufficiently large a, b, consider the search interval $I_a := a^{-(1+p)} I$. If the selected learning rate η_a has loss within ϵ_{search} of the optimal loss, then $\widehat{\eta}_b = (a/b)^{1+p} \eta_a$ satisfies

$$\mathcal{L}_{b,\mathcal{O},T}(\widehat{\eta}_b) - \min_{\eta \in I_b} \mathcal{L}_{b,\mathcal{O},T}(\eta) \le 2(\delta_a + \delta_b) + \epsilon_{\text{search}} \tag{13}$$

Moreover, if $\Phi_{\mathcal{O},T}$ has on I a unique positive minimizer $\zeta^*_{\mathcal{O}}$, then any minimizer on I_a, denoted $\eta^*_{a,\mathcal{O}}$, satisfies $a^{1+p} \eta^*_{a,\mathcal{O}} \to \zeta^*_{\mathcal{O}}$.

Lemma B.7 establishes the state-map conditions with $\tau_a = 0$ when the gradients and normalized response and state updates are locally Lipschitz in all inputs, including the factors, with uniformly nondegenerate denominators. For factor rebalancing, Lemma B.8 specifies the initial-gauge fixed-point, $C^{1,1}$ regularity, and state-transport conditions.

The key observation is that both the movement of the normalized factors and the product cross term are $O(a^{-2})$. A finite-step induction controls how this error propagates through the state updates to the final loss. For ideal additive updates without separate rebalancing, $\Phi_{\mathcal{O},T}(\zeta)$ is the loss of LoRA linearized at the normalized initial point, trained for T steps with multiplier ζ and the optimizer's factor-dependent response evaluated at the initial factors. The same interpretation holds under the rebalancing and identity-state-transport conditions in Appendix B.3. In this limit, the weight displacement lies in T_0, connecting the transfer theorem to the restricted model in Section 4.2. The amplitude-independent limiting loss curve yields learning-rate transfer. The argument from uniform convergence to transfer of minimizers is related to width-dependent LR transfer under μP (Hayou, 2026); here the varying quantity is the initial factor amplitude, and the limiting dynamics retain the optimizer state.

Observations across initial magnitudes. We varied a fixed-basis initialization gain $g_{\text{init}} > 0$, defined by $\text{Var}(B_0 A_0) = g_{\text{init}} \text{Var}(W_{\text{pre}})$ in each layer, with entrywise centered variance and $s = 1$. With fixed DST shapes, the factor-amplitude ratio is $(g_{\text{init},2}/g_{\text{init},1})^{1/4}$. Under rank-128 AdamW, the five gains $0.001, 0.01, 0.1, 1, 10$ span relative amplitudes 1–10 with respect to $g_{\text{ref}} = 0.001$ (Figure 2). Log-LR interpolation of the grid minima in mean training loss over the final 100 steps gives a log-LR versus log-amplitude slope of -1.18. The normalized learning rate $\eta^*(g_{\text{init}}/g_{\text{ref}})^{1/4}$ ranges from 2.55 to 3.83×10^{-4}, a factor of approximately 1.5. These measurements include 11 FP32 adapter-branch remeasurements, and clipping histories vary with initialization magnitude.

In a separate 200-step control, scaling both factors by 0.56235 gives same-LR product-step ratios of 0.578–0.586 for AdamW and 0.9999 for LoRA-Muon over steps 10–50. This control supports their distinct first-order response scales (Appendix C.7).

5.3 Directional Effects on the Optimal Learning Rate

Changing the initial direction changes how much a product step of a given size lowers the loss. At a common weight W, let $H = -\nabla_W \mathcal{L}$, and let $F_i \neq 0$ and Q_i be the first-order response and cross term of initialization $i \in \{0, 1\}$. We define the directional efficiency $c_i := \langle H, F_i \rangle_F / (\|H\|_F \|F_i\|_F)$ and the exact one-step path

$$\phi_i(\eta) := \mathcal{L}(W + \eta F_i + \eta^2 Q_i), \qquad k_i(\eta) := \frac{\phi_i''(\eta)}{\|F_i\|_F^2}. \tag{14}$$

The curvature k_i includes both the Hessian along F_i and the bending of the path caused by Q_i.

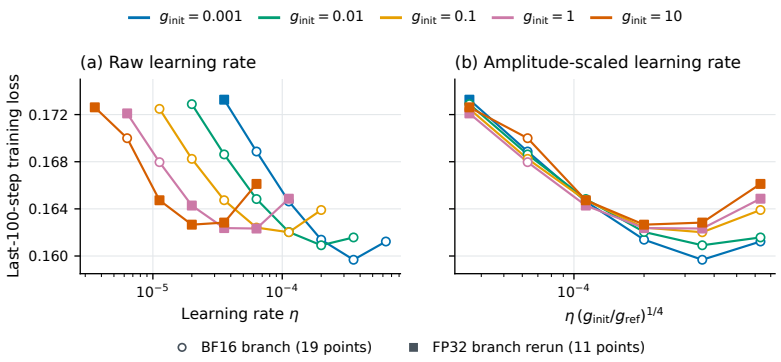

Figure 2: AdamW learning-rate curves with varying initial magnitudes along fixed DST directions. (a) Raw LR and (b) LR multiplied by the relative factor amplitude, $\eta(g_{\text{init}}/g_{\text{ref}})^{1/4}$, for $g_{\text{ref}} = 0.001$. The vertical axis is the mean training loss over the final 100 steps of 1,563-step training. The five gains $g_{\text{init}} = 0.001$–10 span a tenfold factor-amplitude range.

Lemma 3 (Directional Correction for the One-Step Optimal Learning Rate)**.** *Let $H \neq 0$ and $c_i > 0$. Let $\eta_i^\dagger > 0$ be an interior minimizer of ϕ_i, where ϕ_i is C^2 on an open interval containing $[0, \eta_i^\dagger]$ and $0 < \underline{k}_i \leq k_i(\eta) \leq \overline{k}_i$ on $[0, \eta_i^\dagger]$. Then the step-matched learning rate $\eta_{\text{scale}} = \eta_0^\dagger \|F_0\|_F / \|F_1\|_F$, given with $\eta_0 = \eta_0^\dagger$, satisfies*

$$\frac{c_1}{c_0}\frac{\underline{k}_0}{\overline{k}_1} \leq \frac{\eta_1^\dagger}{\eta_{\text{scale}}} \leq \frac{c_1}{c_0}\frac{\overline{k}_0}{\underline{k}_1}. \tag{15}$$

The proof integrates the curvature from $\phi_i'(0) = -\|H\|_F c_i \|F_i\|_F$ to $\phi_i'(\eta_i^\dagger) = 0$ (Appendix B.3). At the same step norm, a more efficient direction allows a longer step, and a more curved path allows only a shorter one. The band applies to a one-step line search from the current state.

We compared LoRA-GA with Fixed-GA, which is matched to it in size. Under the momentum-free LoRA-Muon response at initialization, LoRA-GA had $1.96\times$ the directional efficiency of Fixed-GA, yet its loss-minimizing first-order product step was only $0.483\times$ as long. Under a fresh Adam response, the ratio was $1.31\times$. The directional correction therefore shortens the step under LoRA-Muon and lengthens it under Adam, so it depends on the optimizer. These are FP32 diagnostics that freeze the state and compute fresh responses on 128 training examples. Appendix C.7 lists the tokenizer and numerical differences between the two initial states. We also provide a practical recipe for applying these transfer rules, with full details given in the Appendix.B.4.

6 Conclusion

This paper analyzed how LoRA initialization constrains training geometry and how, together with the optimizer, it sets the update size at a given learning rate. For rank-r initial products, we bounded support rotation and tangent-space departure through the initial spectrum and cumulative movement. In our experiments, training restricted to the initial tangent space reached almost the same loss as unconstrained LoRA even as departure varied substantially, and initializations with tuned learning rates reached similar performance. These results make update size a necessary control in comparisons of initialization. At a common weight gradient and corresponding optimizer states, scaling both factors by a scales the first-order weight response by a^{1+p}, where p is the optimizer's response order. We derived a learning-rate rule that matches this response and proved conditions under which it transfers the final-loss curve and the optimal learning rate over a fixed training horizon when the optimizer and factor shape are fixed. We also derived correction ranges for directional changes and evaluated pilot-based transfer across optimizers. Our comparison procedure uses this rule to guide learning-rate search and reports final performance together with budget utilization and tangent-space departure.

AI USE STATEMENT

We used generative AI tools solely to proofread the manuscript for grammar, wording, and clarity. We did not use generative AI tools to develop research ideas or methods, analyze results, or generate scientific claims. All AI-suggested edits were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. The authors checked the entire manuscript and take full responsibility for its final content.

REPRODUCIBILITY STATEMENT

The setup and theoretical assumptions are specified in Sections 2, 3, and 5. Appendix B contains the proofs and supplementary regularity conditions. Appendix C specifies the data splits, initialization and precision settings, measurement conventions, and evaluation procedures for the reported experiments. The anonymized supplementary material provides the source code, configuration files, and scripts used to reproduce the reported experiments, tables, and figures.

REFERENCES

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. LoRA-Muon: Spectral steepest descent on the low-rank manifold, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Yihe Dong and Will Sawin. Muon^p: Muon with fractional spectral powers, 2026.

Florian Feppon and Pierre F. J. Lermusiaux. A geometric approach to dynamical model order reduction. *SIAM Journal on Matrix Analysis and Applications*, 39(1):510–538, 2018.

Nikhil Ghosh, Tetiana Parshakova, and Robert M. Gower. PoLoRA: A preconditioned orthogonalized LoRA optimizer, 2026.

Soufiane Hayou. A proof of learning rate transfer under μP. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3322–3330. PMLR, 2026. URL https://proceedings.mlr.press/v300/hayou26a.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, 2024a.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning*, 2024b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: Efficient and accurate normalization schemes in deep networks. In *Advances in Neural Information Processing Systems*, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA, 2023.

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

Yu-Ang Lee, Ching-Yun Ko, Pin-Yu Chen, and Mi-Yen Yeh. Learning rate matters: Vanilla lora may suffice for llm fine-tuning, 2026. URL https://arxiv.org/abs/2602.04998.

Shiwei Li, Xiandi Luo, Xing Tang, Haozhao Wang, Hao Chen, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Beyond zero initialization: Investigating the impact of non-zero initialization on lora fine-tuning dynamics. In *The Forty-Second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, 13th-19th July, 2025*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024.

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

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072, 2024.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, 2024.

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

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and René Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, 2025. URL https://api.semanticscholar.org/CorpusID:276903684.

An Yang et al. Qwen3 technical report, 2025.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8df7c2e3c3c3be098ef7b382bd2c37ba-Abstract.html.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA done RITE: Robust invariant transformation equilibration for LoRA optimization. In *International Conference on Learning Representations*, 2025.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, 2024.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *International Conference on Machine Learning*, 2024.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *International Conference on Machine Learning*, 2025a. arXiv:2502.01235.

Zicheng Zhang, Haoran Li, Yifeng Zhang, Guoqiang Gong, Jiaxing Wang, Junxing Hu, Pengzhang Liu, and Qixia Jiang. The primacy of magnitude in low-rank adaptation. *Advances in Neural Information Processing Systems*, 38:39–69, 2025b.

A Related Work

LoRA initialization. LoRA (Hu et al., 2022) starts from $B_0 = 0$ and a random A_0. PiSSA (Meng et al., 2024) and MiLoRA (Wang et al., 2025) initialize the two factors using the leading or trailing singular directions of the pretrained weights, whereas LoRA-GA (Wang et al., 2024) and LoRA-One (Zhang et al., 2025a) use singular directions of the initial gradient. These methods change both the initial directions and the initial magnitude. Hayou et al. (2024a) showed that the training dynamics depend on whether A_0 or B_0 is initialized to zero, and Li et al. (2025) analyzed how initializing both factors to nonzero values changes robustness to the learning rate. Zhang et al. (2025b) explained the effect of initialization through magnitude and proposed an initialization that fixes the directions using a sine-transform basis while controlling the magnitude separately. Our fixed-basis control uses this basis as its initial directions. Lee et al. (2026) reported that sufficiently broad learning-rate searches substantially reduce the performance differences among several LoRA variants and initializations. We explain this observation through the geometry of the initial tangent space and the optimizer response order, and provide conditions for learning-rate transfer across initializations.

LoRA scaling and learning rates. LoRA+ (Hayou et al., 2024b) assigns different learning rates to the two factors, rsLoRA (Kalajdzievski, 2023) adjusts the rank-dependent scale s, and DoRA (Liu et al., 2024) learns the magnitude and direction of a weight separately. The reparameterization theorem in Section 5.1 gives conditions under which s, factor imbalance, and factor-specific learning rates can be exchanged exactly.

LoRA-specific optimizers. Riemannian preconditioned LoRA (Zhang & Pilanci, 2024) derives a preconditioner from a gauge-invariant quotient metric, while LoRA-RITE (Yen et al., 2025) designs updates that are invariant to factor reparameterization. We apply Muon^p (Dong & Sawin, 2026), which extends the matrix spectral transformation of Muon (Jordan et al., 2024) using fractional powers, to the LoRA factors. LoRA-Muon (Cesista et al., 2026) implements spectral steepest descent on the low-rank manifold, and PoLoRA (Ghosh et al., 2026) combines spectral directions, curvature preconditioning, and magnitude-control rules tailored to the low-rank product. The superscript p in Muon^p denotes a spectral exponent and is distinct from the response order p used in this work. Under idealized conditions, we compare the responses to initial amplitude for SGD, Adam (Kingma & Ba, 2015), AdamW without weight decay (Loshchilov & Hutter, 2019), LoRA-Muon, and PoLoRA. Muon^p and LoRA-RITE are included in the empirical comparisons of accuracy and learning-rate sweeps (Appendix).

Theory of LoRA training dynamics. Xu et al. (2025) analyzed the gradient flow of LoRA in matrix factorization problems. Our limiting curve $\Phi_{\mathcal{O},T}$ corresponds to a regime in which large-amplitude LoRA converges to the model linearized at its initialization.

Effective learning rates and learning-rate transfer. For scale-invariant weights, the weight norm determines the effective learning rate (Hoffer et al., 2018). The two effective learning rates in Theorem 1 are the LoRA counterpart of this relationship, expressed in terms of the initial factor amplitude and response order. Width-dependent learning-rate transfer was introduced in μP (Yang et al., 2021), and Hayou (2026) proved that the optimal learning rate converges to a positive constant as width increases in linear multilayer perceptrons parameterized with μP. Theorem 3 instead provides a conditional transfer result in which the initial amplitude, rather than the width, is varied.

Geometry of the fixed-rank manifold. The tangent spaces and projections of the rank-r matrix manifold have long been used in low-rank optimization (Vandereycken, 2013). Feppon & Lermusiaux (2018) analyzed the curvature of this manifold. We distinguish the second fundamental form under the Frobenius metric from the quadratic term of the Schur graph and directly prove that the maximum normal curvature is $1/\sigma_r$ when orthogonal complements exist on both sides (Lemma B.2). The bound on support rotation is related to Wedin's $\sin\Theta$ theorem (Wedin, 1972); Proposition 2 derives the required form using only a projection identity for two products of the same rank. We combine this geometry with the actual product path length and optimizer-specific amplitude responses to derive bounds on support rotation and conditions for learning-rate transfer.

B PROOFS

Table 3 summarizes the proof locations and dependencies for results discussed in the main text and appendix.

Table 3: Proof locations and dependencies for results in the main text and appendix. Rows labeled "derivation" point to the assumptions and calculations behind the corresponding equations.

Result	Main-text use	Proof	Dependencies
Lemma 1 (initial tangent space)	§2.2	B.1	–
Lemma B.1 (projections and distances)	§2.2	B.1	Lemma 1
Lemma 2 (response order)	§2.3	B.1	state covariance
Proposition 1 (Schur representation)	§3.1	B.2	–
Lemma B.2 (normal curvature)	§3.1	B.2	Proposition 1
Lemma B.3 (support-space derivative)	§3.2	B.2	Lemma 1
Theorem 1 (two effective learning rates)	§3.2	B.2	Lemma 2, B.3
Proposition 2 (budget bound)	§3.3	B.2	Proposition 1
Equation 8 (derivation)	§3.3	B.2	uniform bound on actual per-step displacement
Proposition B.1 (amplitude equivariance)	§3.2	B.2	Lemma 2, exogenous gradients
Corollary 1 (ξ invariance)	§3.3	B.2	Proposition B.1
Lemma B.4 (q_g)	§5.1	B.3	Equation 19
Lemma B.5 (decay and clipping)	§5.1	B.3	Equation 19
Theorem 2 (reparameterization)	§5.1	B.3	Lemma B.4, B.5
Equation 10 (derivation)	§5.1	B.3	Theorem 2
Lemma B.6 (displacement coordinates)	§5.2	B.3	state covariance for the response order
Lemma B.7 (regularity)	§5.2	B.3	Lemma B.6
Lemma B.8 (rebalancing)	§5.2	B.3	Lemma B.7
Theorem 3 (amplitude transfer)	§5.2	B.3	state-map conditions (A1)–(A4); the two preceding lemmas give sufficient conditions
Lemma 3 (directional correction)	§5.3	B.3	C^2 regularity and two-sided bounds on path curvature

Unless stated otherwise, $X_0 = U_0\Sigma_0 V_0^\top$ is a rank-r thin singular value decomposition (SVD), and $U_{0\perp}$ and $V_{0\perp}$ are orthonormal bases for the respective orthogonal complements. Since $P_{U_0} = U_0U_0^\top$ and $P_{V_0} = V_0V_0^\top$, we have $I - P_{U_0} = U_{0\perp}U_{0\perp}^\top$ and $I - P_{V_0} = V_{0\perp}V_{0\perp}^\top$. In these bases, we partition a matrix Y into the four blocks $U_0^\top Y V_0$, $U_0^\top Y V_{0\perp}$, $U_{0\perp}^\top Y V_0$, and $U_{0\perp}^\top Y V_{0\perp}$. Because the bases are orthonormal, $\|Y\|_F^2$ is the sum of the squared norms of these four blocks.

B.1 PROOFS FOR SECTION 2

Lemma 1 (Initial tangent space)**.** *Let $J_0[Z_B, Z_A] = s(Z_B A_0 + B_0 Z_A)$ be the initial factor Jacobian. If both B_0 and A_0 have rank r, then* $\operatorname{im} J_0 = T_0$*. Consequently, $\Delta W - s\,\Delta B\,\Delta A \in T_0$.*

Proof. We prove $\operatorname{im} J_0 = T_0$ and also establish the identities $T_0 = \{N = 0\}$ and $N = s\,U_{0\perp}^\top \Delta B\,\Delta A\,V_{0\perp}$ used in the main text. The key point is that B_0 restricts output directions to U_0, while A_0 restricts input directions to V_0.

(1) Form of T_0. By definition, an element of T_0 has the form $P_{U_0}Y + YP_{V_0} - P_{U_0}YP_{V_0}$, which equals $U_0(U_0^\top Y) + \big((I - P_{U_0})YV_0\big)V_0^\top$. Thus an element of T_0 can be represented, for some $M_1 \in$

$\mathbb{R}^{r\times d_{\text{in}}}$ and $M_2 \in \mathbb{R}^{d_{\text{out}}\times r}$, as $U_0M_1 + M_2V_0^\top$, and each such matrix Y satisfies $U_{0\perp}^\top Y V_{0\perp} = 0$. Conversely, for any Y,

$$Y - \big(P_{U_0}Y + YP_{V_0} - P_{U_0}YP_{V_0}\big) = (I - P_{U_0})Y(I - P_{V_0}) = U_{0\perp}\big(U_{0\perp}^\top Y V_{0\perp}\big)V_{0\perp}^\top$$

so $U_{0\perp}^\top Y V_{0\perp} = 0$ implies $Y \in T_0$. Therefore,

$$T_0 = \big\{U_0M_1 + M_2V_0^\top\big\} = \big\{Y : U_{0\perp}^\top Y V_{0\perp} = 0\big\} \tag{16}$$

and $T_0 = \{N = 0\}$ in block coordinates. Counting the dimensions of the remaining three blocks gives $r(d_{\text{out}} + d_{\text{in}} - r)$.

(2) Factors and supports. If $\operatorname{rank} A_0 = r$, then A_0 maps onto $\mathbb{R}^r$, so $\operatorname{col} X_0 = \operatorname{col} B_0$. Likewise, if $\operatorname{rank} B_0 = r$, then $\operatorname{row} X_0 = \operatorname{row} A_0$. Hence $\operatorname{col} U_0 = \operatorname{col} B_0$ and $\operatorname{col} V_0 = \operatorname{row} A_0$, so we may write $B_0 = U_0R_B$ and $A_0 = R_AV_0^\top$. Here $R_B := U_0^\top B_0$ and $R_A := A_0V_0$ are rank-r, $r \times r$ matrices and thus invertible.

(3) Image of J_0. Substituting (2) gives

$$J_0[Z_B, Z_A] = (sZ_BR_A)V_0^\top + U_0(sR_BZ_A)$$

Since R_A and R_B are invertible, the maps $Z_B \mapsto sZ_BR_A$ and $Z_A \mapsto sR_BZ_A$ are bijections on $\mathbb{R}^{d_{\text{out}}\times r}$ and $\mathbb{R}^{r\times d_{\text{in}}}$, respectively. Thus Equation 16 gives $\operatorname{im} J_0 = T_0$.

(4) Cumulative change. Because $\Delta W = s(B_0 + \Delta B)(A_0 + \Delta A) - sB_0A_0 = J_0[\Delta B, \Delta A] + s\,\Delta B\,\Delta A$, we have $\Delta W - s\,\Delta B\,\Delta A \in T_0$. Applying $U_{0\perp}^\top(\cdot)V_{0\perp}$ to both sides, the first term vanishes by Equation 16, leaving $N = s\,U_{0\perp}^\top \Delta B\,\Delta A\,V_{0\perp}$. □

The case $B_0 = 0$ (vanilla LoRA). Since $X_0 = 0$, the space T_0 is undefined, so we consider only $\operatorname{im} J_0$. If $\operatorname{rank} A_0 = r$, choose an orthonormal basis of $\operatorname{row} A_0$, denoted by $V_A \in \mathbb{R}^{d_{\text{in}}\times r}$, and an invertible matrix $R_A := A_0V_A$. Then $A_0 = R_AV_A^\top$ and $J_0[Z_B, Z_A] = (sZ_BR_A)V_A^\top$. Therefore,

$$\operatorname{im} J_0 = \big\{MV_A^\top : M \in \mathbb{R}^{d_{\text{out}}\times r}\big\} = \big\{Y : \operatorname{row} Y \subseteq \operatorname{row} A_0\big\}$$

and its dimension is $d_{\text{out}}r$. First-order changes can alter only the output directions; the input directions remain restricted to $\operatorname{row} A_0$. The identity $\Delta W - s\,\Delta B\,\Delta A = s\,\Delta B\,A_0 \in \operatorname{im} J_0$ also holds.

Lemma B.1 (Tangent-space projections and support-subspace distance)**.** *Let $X = U\Sigma V^\top \in \mathcal{M}_r$ and $P_U = UU^\top$, $P_V = VV^\top$. The orthogonal projections onto T_X and its orthogonal complement under the Frobenius inner product are, respectively,*

$$P_{T_X}Y = P_UY + YP_V - P_UYP_V, \qquad P_{T_X^\perp}Y = (I - P_U)Y(I - P_V)$$

and $\|Y\|_F^2 = \|P_{T_X}Y\|_F^2 + \|P_{T_X^\perp}Y\|_F^2$. Also, if $V, V_0 \in \mathbb{R}^{d_{\text{in}}\times r}$ have orthonormal columns,

$$\sqrt{1 - \frac{\|V^\top V_0\|_F^2}{r}} = \frac{\|\sin\Theta(V, V_0)\|_F}{\sqrt{r}} = \frac{\|P_V - P_{V_0}\|_F}{\sqrt{2r}} \in [0, 1]. \tag{17}$$

Proof. In the $[U, U_\perp]$ and $[V, V_\perp]$ coordinates, T_X consists of all matrices whose lower-right block is 0 (see the proof of Lemma 1). The four blocks are mutually orthogonal under the Frobenius inner product. Thus deleting the lower-right block or retaining only that block gives the two orthogonal projections above. The norm identity follows from the same orthogonal decomposition. If the principal angles are $\theta_1, \ldots, \theta_r$, the singular values of $V^\top V_0$ are $\cos\theta_j$, so $r - \|V^\top V_0\|_F^2 = \sum_j \sin^2\theta_j$. Meanwhile, $\operatorname{tr}(P_V) = \operatorname{tr}(P_{V_0}) = r$ and $\operatorname{tr}(P_VP_{V_0}) = \|V^\top V_0\|_F^2$ give $\|P_V - P_{V_0}\|_F^2 = 2r - 2\|V^\top V_0\|_F^2$. □

The energy captured from each descent field by this projection is no greater than its total energy. Hence the pooled capture ratio C_H in the main text lies in $[0, 1]$ when its denominator is positive. The normalized chordal distance in equation 17 measures the root mean square of the sines of the principal angles, whereas $\|\sin\Theta\|_2$ in Proposition 2 measures the sine of the largest principal angle. The aggregation conventions for these two metrics therefore differ.

Lemma 2 (Response Order). *Consider an idealized implementation in which the state correspondence persists after one step, the stabilizing constant, weight decay, and gradient clipping are removed, and all normalization denominators are defined. Then the response order is $p = 1$ for SGD (including momentum), $p = 0$ for Adam, and $p = -1$ for LoRA-Muon and PoLoRA. The Gram inverse square roots and polar inputs in LoRA-Muon must be nondegenerate; PoLoRA's curvature state must be scale covariant, and its direction-norm floor must be inactive.*

Proof. For each optimizer, define a state correspondence $z \mapsto z_a$ and establish two facts for the scaled input $(aG_B, aG_A; aB, aA, z_a)$. (i) The response is multiplied by a^p, as in Equation 5. (ii) The state remains in correspondence after one step. The latter is the one-step preservation condition in the definition and is used in the proofs of Proposition B.1 and Lemma B.6. We consider only states for which the inverses and normalizations are defined, and, when a denominator is 0, define the resulting zero response to be 0. This convention commutes with positive scalar multiplication, so it does not break algebraic homogeneity, but it does not guarantee continuity near zero or the regularity required by the transfer theorem. Below, $\bullet \in \{B, A\}$, and we repeatedly use two facts. First, momentum and moving-average updates preserve scalar multiplication of their inputs. Second, the matrix sign function defined from a compact SVD retaining only positive singular values, $Y = U_Y \Sigma_Y V_Y^\top$, by $\operatorname{msign}(Y) := U_Y V_Y^\top$ (with $\operatorname{msign}(0) := 0$), is invariant under positive scalar multiplication.

SGD. Let the momentum coefficient be $\beta_1 \in [0, 1)$, and set $z = (m_B, m_A)$ and $z_a = (am_B, am_A)$. The updates are $m_\bullet^+ = \beta_1 m_\bullet + G_\bullet$ and $D_\bullet = -m_\bullet^+$; the Nesterov form is also a fixed linear combination of $G_\bullet$ and $m_\bullet^+$. Since all inputs are multiplied by a, both $m_\bullet^+$ and $D_\bullet$ are multiplied by a, giving $p = 1$.

Adam. Set $z = (m_\bullet, v_\bullet, t)$ and $z_a = (am_\bullet, a^2 v_\bullet, t)$. The update $m_\bullet^+ = \beta_1 m_\bullet + (1 - \beta_1) G_\bullet$ is multiplied by a, while $v_\bullet^+ = \beta_2 v_\bullet + (1 - \beta_2) G_\bullet \odot G_\bullet$ is multiplied by a^2. Thus, for a bias-correction constant independent of a, denoted c_t, the elementwise response $D_\bullet = -c_t\, m_\bullet^+ \oslash \sqrt{v_\bullet^+}$ (with elementwise division and square root) is multiplied by $a/\sqrt{a^2} = 1$, so $p = 0$. Adding ϵ_{adam} to the denominator makes this relation only approximate for entries where $\sqrt{v_\bullet^+} \gg \epsilon_{\text{adam}}$.

LoRA-Muon. When both current factors have rank r, the responses are

$$D_A = -\tfrac{1}{2}(B^\top B)^{-1/2} \operatorname{msign}\big((B^\top B)^{-1/2} m_A\big),$$
$$D_B = -\tfrac{1}{2} \operatorname{msign}\big(m_B (AA^\top)^{-1/2}\big)(AA^\top)^{-1/2}.$$

The correspondence for the momentum buffers $m_\bullet$ is the same as for SGD. Since $\big((aB)^\top (aB)\big)^{-1/2} = a^{-1}(B^\top B)^{-1/2}$,

$$D_A^{(a)} = -\tfrac{1}{2} a^{-1}(B^\top B)^{-1/2} \operatorname{msign}\big(a^{-1}(B^\top B)^{-1/2}\, am_A\big) = a^{-1} D_A$$

and likewise for D_B. Hence $p = -1$. Here $p = -1$ is the order of the additive response before rebalancing. The product-preserving rebalancing $(B, A) \mapsto (B/c, cA)$ uses a coefficient c determined by the ratio of the two factor magnitudes. If the factors correspond by a common scale a after the additive update, the value of c is the same, so rebalancing preserves that correspondence. This condition holds for the rotation-matched learning rates of Proposition B.1. We do not claim that the rebalancing coefficients agree for arbitrary steps in two runs using the same learning rate. Momentum must also be transported according to the rebalancing rule; for example, transport in gradient coordinates, $(m_B, m_A) \mapsto (cm_B, m_A/c)$, commutes with common-amplitude correspondence.

PoLoRA. The state comprises momenta M_B and M_A and positive curvature vectors $u_{\text{out}} \in \mathbb{R}^{d_{\text{out}}}$ and $u_{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$ (Ghosh et al., 2026, Algorithm 1); the correspondence is $z_a = (aM_B, aM_A, u_{\text{out}}, u_{\text{in}})$. The curvature vectors are used in normalized form as $P_{\text{out}} = \operatorname{diag}(u_{\text{out}}/\|u_{\text{out}}\|_\infty)$ and $P_{\text{in}} = \operatorname{diag}(u_{\text{in}}/\|u_{\text{in}}\|_\infty)$. Thus P_{out} and P_{in} remain unchanged, while $C_B := B^\top P_{\text{out}} B$ and $C_A := A P_{\text{in}} A^\top$ are multiplied by a^2. Writing $\hat{M}_\bullet$ for momentum after adding the current gradient, the directions are

$$\check{D}_A = C_B^{-1/2} \operatorname{msign}\big(C_B^{-1/2} \hat{M}_A P_{\text{in}}^{-1/2}\big) P_{\text{in}}^{-1/2},$$
$$\check{D}_B = P_{\text{out}}^{-1/2} \operatorname{msign}\big(P_{\text{out}}^{-1/2} \hat{M}_B C_A^{-1/2}\big) C_A^{-1/2}.$$

Within the matrix sign function, the scales cancel as $a^{-1} \cdot a = 1$, leaving the directions multiplied by a^{-1}. PoLoRA sets their magnitudes using

$$D_\bullet = -\frac{\check{D}_\bullet}{\|\check{D}_\bullet\|_2 \left(\|A\|_2 + \|B\|_2\right)}$$

The normalization removes the scale of $\check{D}_\bullet$, and the denominator $\|aA\|_2 + \|aB\|_2 = a(\|A\|_2 + \|B\|_2)$ leaves a factor of a^{-1}. Hence $p = -1$. The curvature-update terms $\mathrm{diag}(G_A^\top C_B^{-1} G_A)$ and $\mathrm{diag}(G_B C_A^{-1} G_B^\top)$ each scale by $a \cdot a^{-2} \cdot a = 1$, so the curvature vectors are updated identically in both runs and the momenta are multiplied by a. This also establishes (ii).

The states of SGD, Adam, and LoRA-Muon consist of gradient momenta and moving averages, so each component updates with its corresponding scaling order, establishing (ii). Fixed absolute stabilization constants (ϵ_{adam}, absolute damping of the inverse Gram square root, and an absolute floor in a normalization denominator) change their relative size under scaling and generally break this homogeneity. Relative damping that scales by a^2 together with the Gram matrix may be an exception. Fixed decoupled weight decay adds a degree-one term proportional to the factor and is therefore generally incompatible with responses of order $p \neq 1$; fixed-threshold gradient clipping changes the multiplier according to the input norm. Coupled weight decay and the transformations of its hyperparameters must be checked separately with the order of operations, as in Lemma B.5. □

B.2 Proofs for Section 3

Proposition 1 (Schur Representation of Departure). *If $\gamma := \sigma_r(X_0) - \|C\|_2 > 0$, then*

$$N = LK^{-1}R, \qquad \|N\|_F \le \frac{\|\Delta W\|_F^2}{2\gamma}$$

and consequently $\varepsilon \le x^2/(4\bar{\gamma}^2)$. Here $\bar{\gamma} := \gamma/\sigma_r(X_0)$.

Proof. In block coordinates, $X = X_0 + \Delta W$ is the matrix in Equation 6, namely $\begin{bmatrix} K & R \\ L & N \end{bmatrix}$. Weyl's inequality gives $\sigma_{\min}(K) \ge \sigma_r(X_0) - \|C\|_2 = \gamma > 0$, so K is invertible and $\|K^{-1}\|_2 \le 1/\gamma$. Hence

$$\begin{bmatrix} K & R \\ L & N \end{bmatrix} = \begin{bmatrix} I & 0 \\ LK^{-1} & I \end{bmatrix} \begin{bmatrix} K & R \\ 0 & N - LK^{-1}R \end{bmatrix}$$

and, because the first factor is invertible, $\mathrm{rank}\, X = r + \mathrm{rank}(N - LK^{-1}R)$. Since $X = sBA$ has rank at most r, we obtain $N - LK^{-1}R = 0$, or $N = LK^{-1}R$.

The upper bound follows from

$$\|N\|_F \le \|L\|_F \|K^{-1}\|_2 \|R\|_F \le \frac{\|L\|_F \|R\|_F}{\gamma} \le \frac{\|L\|_F^2 + \|R\|_F^2}{2\gamma} \le \frac{\|\Delta W\|_F^2}{2\gamma}$$

where the last inequality uses that L and R are distinct blocks of ΔW. If $\Delta W = 0$, then $\varepsilon = x = 0$ and the conclusion is immediate. Otherwise, squaring both sides and dividing by $\|\Delta W\|_F^2$ gives $\varepsilon \le \|\Delta W\|_F^2/(4\gamma^2) = x^2/(4\bar{\gamma}^2)$. □

Lemma B.2 (Second fundamental form and maximum normal curvature). *Let $1 \le r \le \min(d_{\mathrm{out}}, d_{\mathrm{in}})$ and equip $\mathcal{M}_r$ with the Riemannian metric induced by the Frobenius inner product on the ambient matrix space. In the initial singular-vector coordinates, write tangent vectors as*

$$Z_i = \begin{bmatrix} C_i & R_i \\ L_i & 0 \end{bmatrix} \in T_0 \qquad (i = 1, 2)$$

Then the second fundamental form is

$$\mathrm{II}_{X_0}(Z_1, Z_2) = U_{0\perp}\big(L_1 \Sigma_0^{-1} R_2 + L_2 \Sigma_0^{-1} R_1\big) V_{0\perp}^\top. \tag{18}$$

Thus, for $Z = \left[\begin{smallmatrix} C & R \\ L & 0 \end{smallmatrix}\right]$, the quadratic term of the Schur graph is $\frac{1}{2}\mathrm{II}_{X_0}(Z, Z) = U_{0\perp} L \Sigma_0^{-1} R V_{0\perp}^\top$. Define the maximum normal curvature by

$$\kappa(X_0) := \sup_{\substack{Z \in T_0 \\ \|Z\|_F = 1}} \|\mathrm{II}_{X_0}(Z, Z)\|_F$$

When $r < \min(d_{\mathrm{out}}, d_{\mathrm{in}})$, $\kappa(X_0) = 1/\sigma_r(X_0)$; otherwise, $\kappa(X_0) = 0$.

Proof. By Proposition 1, near X_0 the manifold $\mathcal{M}_r$ is the smooth graph

$$(C, R, L) \longmapsto \begin{bmatrix} \Sigma_0 + C & R \\ L & L(\Sigma_0 + C)^{-1}R \end{bmatrix}$$

At the origin, the first derivative of the normal block is 0, and its second derivative is $L_1\Sigma_0^{-1}R_2 + L_2\Sigma_0^{-1}R_1$. By the standard definition, projecting the ambient derivative onto the normal space gives the second fundamental form; the Taylor expansion carries a factor of $1/2$. In particular, changes in C do not contribute to the second derivative of the normal block.

If $\|Z\|_F = 1$, then

$$\|\mathrm{II}_{X_0}(Z, Z)\|_F \leq \frac{2\|L\|_F\|R\|_F}{\sigma_r(X_0)} \leq \frac{\|L\|_F^2 + \|R\|_F^2}{\sigma_r(X_0)} \leq \frac{1}{\sigma_r(X_0)}.$$

If both orthogonal complements are nonempty, choose unit coordinate vectors u, v in them, respectively. For the smallest singular value of Σ_0, let the corresponding coordinate vector be e_r. Taking $C = 0$, $L = ue_r^\top/\sqrt{2}$, $R = e_r v^\top/\sqrt{2}$ gives $\|Z\|_F = 1$ and $2L\Sigma_0^{-1}R = uv^\top/\sigma_r(X_0)$, attaining the upper bound. If either orthogonal complement is empty, the normal space is $\{0\}$ and the manifold of maximal-rank matrices is open; hence the second fundamental form and normal curvature both equal 0. □

The proof of Theorem 1 requires the first-order rotation of the support over one step.

Lemma B.3 (First-order support variation)**.** *Suppose both factors have rank r. Let $X = U\Sigma V^\top$, and set $B(\eta) = B + \eta D_B$, $A(\eta) = A + \eta D_A$, and $X(\eta) = sB(\eta)A(\eta) = X + \eta F + \eta^2 Q$. For $X(\eta)$, let $P_U(\eta)$ and $P_V(\eta)$ denote the projectors onto its left and right supports, respectively. As $\eta \downarrow 0$,*

$$P_U(\eta) = P_U + \eta\big(L_U U^\top + U L_U^\top\big) + O(\eta^2), \qquad P_V(\eta) = P_V + \eta\big(L_V V^\top + V L_V^\top\big) + O(\eta^2)$$

where $L_U := (I - P_U)FV\Sigma^{-1}$ and $L_V := (I - P_V)F^\top U\Sigma^{-1}$. In particular, $\|\sin\Theta(U(\eta), U)\|_2 = \eta\|L_U\|_2 + O(\eta^2)$, and the same holds for the right support.

Proof. Full rank is an open condition, so for sufficiently small η, $B(\eta)$ and $A(\eta)$ still have rank r. As in part (2) of the proof of Lemma 1, $\operatorname{col} X(\eta) = \operatorname{col} B(\eta)$, and

$$P_U(\eta) = B(\eta)\big(B(\eta)^\top B(\eta)\big)^{-1}B(\eta)^\top$$

is smooth in η. Write $P_U(\eta) = P_U + \eta Z + O(\eta^2)$ and expand $(I - P_U(\eta))X(\eta) = 0$ to first order. Since $(I - P_U)X = 0$, we obtain $(I - P_U)F = ZX$. Multiplying on the right by $V\Sigma^{-1}$ and using $XV\Sigma^{-1} = U$ gives $ZU = L_U$. Because $P_U(\eta)$ is symmetric and idempotent, its first-order term satisfies $Z^\top = Z$ and $P_U Z + Z P_U = Z$. Therefore

$$Z = ZP_U + P_U Z = (ZU)U^\top + U(ZU)^\top = L_U U^\top + U L_U^\top$$

The argument for $P_V(\eta)$ applies the same reasoning to $(I - P_V(\eta))X(\eta)^\top = 0$ and uses $X^\top U\Sigma^{-1} = V$.

For two subspaces of equal dimension, $\|\sin\Theta\|_2$ equals the spectral norm of the difference between their projectors. Since $U^\top L_U = 0$, the matrix $M := L_U U^\top + U L_U^\top$ satisfies $M^2 = L_U L_U^\top + U(L_U^\top L_U)U^\top$, with the two terms supported on orthogonal subspaces. Hence $\|M\|_2 = \|L_U\|_2$, proving the last claim. □

Theorem 1 (Two Effective Learning Rates)**.** *Suppose $X = sBA$ has rank r, the corresponding states (B, A, z) and (aB, aA, z_a) receive the same weight gradient G, and the response order in Equation 5 is p. The learning-rate-free first-order product response F satisfies $F_a = a^{1+p}F$, while the derivatives at $\eta = 0$ of the left and right support projections are each multiplied by a^{p-1}. Thus the learning rates matching the nonzero first-order response and nonzero instantaneous rotation, respectively, are $\eta_{\mathrm{step}} = a^{-(1+p)}\eta$ and $\eta_{\mathrm{rot}} = a^{1-p}\eta$, with ratio a^2. The finite-step expansion is $X_a^+ - X_a = \eta a^{1+p}F + \eta^2 a^{2p}Q$.*

Proof. Denote quantities in the second run by a superscript (a). Because the factor gradients are (aG_B, aG_A) and the states correspond, the definition of the response order gives $D_B^{(a)} = a^p D_B$, $D_A^{(a)} = a^p D_A$. Hence

$$F^{(a)} = s\big(a^p D_B\, aA + aB\, a^p D_A\big) = a^{1+p}F, \qquad Q^{(a)} = s\, a^p D_B\, a^p D_A = a^{2p}Q$$

and, at the same learning rate, the first-order product change $\eta F^{(a)}$ is multiplied by a^{1+p}.

The product in the second run is $X^{(a)} = a^2 X$, so its support is unchanged while its singular values are multiplied by a^2. Thus $U^{(a)} = U$, $V^{(a)} = V$, $\Sigma^{(a)} = a^2\Sigma$. Substitution into L_U in Lemma B.3 gives

$$L_U^{(a)} = (I - P_U)F^{(a)}V\big(\Sigma^{(a)}\big)^{-1} = a^{1+p}a^{-2}L_U = a^{p-1}L_U$$

and the same holds for L_V. Therefore, the first-order changes of P_U and P_V, namely $\eta(L_U U^\top + U L_U^\top)$ and $\eta(L_V V^\top + V L_V^\top)$, are also multiplied by a^{p-1}.

Let η' be the learning rate in the second run. Matching the first-order product change requires $\eta' a^{1+p} = \eta$, that is, $\eta' = a^{-(1+p)}\eta = \eta_{\mathrm{step}}$. Matching the first-order support change requires $\eta' a^{p-1} = \eta$, that is, $\eta' = a^{1-p}\eta = \eta_{\mathrm{rot}}$. Their ratio is $a^{1-p}/a^{-(1+p)} = a^2$. □

At η_{step}, the cross term becomes $\eta_{\mathrm{step}}^2 Q^{(a)} = a^{-2}\eta^2 Q$. Thus the full exact change in one step, $\eta F + \eta^2 Q$, is not matched; as the amplitude increases, the change approaches its first-order term.

Proposition 2 (Budget Bound)**.** *If X_0 and, for every $k \le t$, X_k have rank r, then, for each $k \le t$,*

$$x_k \le \xi_t, \qquad \max\big\{\|\sin\Theta(U_k, U_0)\|_2,\ \|\sin\Theta(V_k, V_0)\|_2\big\} \le \frac{\|\Delta W_k\|_2}{\sigma_r(X_0)} \le \xi_t$$

Moreover, if $\gamma_k := \sigma_r(X_0) - \|C_k\|_2 > 0$, then $\varepsilon_k \le \xi_t^2/(4\bar\gamma_k^2)$. Here $\bar\gamma_k := \gamma_k/\sigma_r(X_0)$. In particular, if $\xi_t < 1$, then for every $k \le t$ we have $\gamma_k > 0$ and $\varepsilon_k \le \xi_t^2/[4(1-\xi_t)^2]$.

Proof. Fix $k \le t$. By the triangle inequality, $\|\Delta W_k\|_F = \big\|\sum_{j<k}(X_{j+1} - X_j)\big\|_F \le \Lambda_k \le \Lambda_t$, so $x_k \le \xi_t$.

To bound $\sin\Theta$, note that P_{U_k} projects onto $\operatorname{col} X_k$. Thus $(I - P_{U_k})X_k = 0$ and, consequently, $(I - P_{U_k})X_0 = -(I - P_{U_k})\Delta W_k$. Since the left-hand side is $(I - P_{U_k})U_0\Sigma_0 V_0^\top$,

$$\sigma_r(X_0)\,\|(I - P_{U_k})U_0\|_2 \le \|(I - P_{U_k})U_0\Sigma_0\|_2 = \|(I - P_{U_k})\Delta W_k\|_2 \le \|\Delta W_k\|_2$$

follows. Both subspaces have dimension r, so $\|(I - P_{U_k})U_0\|_2 = \|\sin\Theta(U_k, U_0)\|_2$, giving the first inequality. The argument for V applies to $X_k^\top$. The second inequality follows from $\|\Delta W_k\|_2 \le \|\Delta W_k\|_F \le \Lambda_t$.

Finally, if $\gamma_k > 0$, Proposition 1 gives $\varepsilon_k \le x_k^2/(4\bar\gamma_k^2) \le \xi_t^2/(4\bar\gamma_k^2)$ (if $\Delta W_k = 0$, then $\varepsilon_k = 0$). Moreover, $\|C_k\|_2 \le \|\Delta W_k\|_2 \le \Lambda_t$, so $\bar\gamma_k \ge 1 - \xi_t$. Therefore, if $\xi_t < 1$, then $\gamma_k > 0$ and $\varepsilon_k \le \xi_t^2/\{4(1-\xi_t)^2\}$. □

Derivation of Equation 8. As in the main text, assume $\eta_k^{(a)} \ge 0$ and that, for some M independent of the amplitude, the actual finite steps satisfy

$$\|X_{k+1}^{(a)} - X_k^{(a)}\|_F \le M a^{1+p}\eta_k^{(a)} \qquad (k < T)$$

Summing these inequalities and dividing by $\sigma_r(X_0^{(a)}) = a^2\sigma_r(\bar X_0) > 0$ gives

$$\Lambda_T^{(a)} \le M a^{1+p}\sum_{k<T}\eta_k^{(a)}, \qquad \xi_T^{(a)} \le \frac{M a^{p-1}}{\sigma_r(\bar X_0)}\sum_{k<T}\eta_k^{(a)}$$

Substituting $\eta_k^{(a)} = a^{-(1+p)}\bar\eta_k$ yields $\xi_T^{(a)} \le M\sum_{k<T}\bar\eta_k/(a^2\sigma_r(\bar X_0))$. This argument assumes neither a positive lower bound on the response norm nor alignment of directions across steps. Thus the displayed relation is an upper bound, not an equality for the realized utilization or a two-sided scaling law.

One sufficient condition for the full-step bound. Set $\bar{F}_k^{(a)} := a^{-(1+p)}F_k^{(a)}$ and $\bar{Q}_k^{(a)} := a^{-2p}Q_k^{(a)}$. If $\|\bar{F}_k^{(a)}\|_F \le M_F$ and $\eta_k^{(a)}a^{p-1}\|\bar{Q}_k^{(a)}\|_F \le M_Q$ hold for all $k < T$ with constants independent of amplitude, the exact product-step identity gives

$$\|\eta_k^{(a)}F_k^{(a)} + (\eta_k^{(a)})^2 Q_k^{(a)}\|_F \le a^{1+p}\eta_k^{(a)}(M_F + M_Q)$$

so we may take $M = M_F + M_Q$. This sufficient condition must also be verified along the full path; homogeneity at the first step alone does not imply uniform constants at later steps.

The rotation-matching result of Theorem 1 extends exactly to entire trajectories under a fixed gradient field.

Proposition B.1 (Amplitude equivariance)**.** *Suppose the weight gradient G_k at each step is prescribed independently of the trajectory (a fixed gradient field). Let an optimizer of response order p train from (B_0, A_0) with learning rates η_k, yielding a trajectory (B_k, A_k). Starting from (aB_0, aA_0) with initial optimizer state $(z_0)_a$ and learning rates $a^{1-p}\eta_k$ gives, for every k, $(B_k^{(a)}, A_k^{(a)}) = (aB_k, aA_k)$ and $X_k^{(a)} = a^2 X_k$.*

Proof. We use induction on k. The case $k = 0$ holds by assumption. At step k, suppose $(B_k^{(a)}, A_k^{(a)}, z_k^{(a)}) = (aB_k, aA_k, (z_k)_a)$. Both runs receive the same G_k, so the factor gradients are $(aG_{B,k}, aG_{A,k})$; by the definition of response order, $D_{\bullet,k}^{(a)} = a^p D_{\bullet,k}$. Therefore

$$B_{k+1}^{(a)} = aB_k + a^{1-p}\eta_k\, a^p D_{B,k} = a\big(B_k + \eta_k D_{B,k}\big) = aB_{k+1}$$

and the same holds for A. Applying a product-preserving rebalancing that is homogeneous of degree one under common amplitude after the additive update also preserves the factor correspondence. We assume that the accompanying state transport commutes with the amplitude correspondence. Since the state correspondence persists after one step by definition (part (ii) of the proof of Lemma 2), $z_{k+1}^{(a)} = (z_{k+1})_a$. The product is bilinear in the two factors, so $X_k^{(a)} = s(aB_k)(aA_k) = a^2 X_k$. □

Corollary 1 (Invariance of movement-budget utilization under rotation matching)**.** *Under the assumptions of Proposition B.1, if $\sigma_r(X_0) > 0$, then for every finite T,*

$$\Lambda_T^{(a)} = a^2\Lambda_T, \qquad \sigma_r(X_0^{(a)}) = a^2\sigma_r(X_0), \qquad \xi_T^{(a)} = \xi_T.$$

Proof. Sum the Frobenius norms of $X_{k+1}^{(a)} - X_k^{(a)} = a^2(X_{k+1} - X_k)$ and divide by the scaling factor of the initial smallest singular value. □

Residualization ensures a shared weight W_{pre} and a shared minibatch gradient at the first step. At later steps, gradients computed during training do not automatically satisfy the exogenous-gradient assumption of Proposition B.1. Under that assumption, the supports of the two rank-preserving trajectories coincide at every step, while the path length needed to reach the same support change is multiplied by a^2.

Distinguishing retention from realignment of initial directions. Proposition 2 gives the sufficient condition that, if $\xi_T < \sin\theta$, then for every $k \le T$ the largest principal angle is less than θ. Conversely, reaching a principal angle of at least θ at any time requires $\xi_T \ge \sin\theta$. However, a path that changes only product magnitude within the same support can have arbitrarily large ξ_T, so a lower bound on utilization cannot guarantee realignment. The invariance in Corollary 1 compares two paths under exogenous gradients and rotation matching; it does not provide a universal realignment threshold for step matching or gradients computed during training. For vanilla LoRA with $B_0 = 0$, utilization normalized by $\sigma_r(X_0)$ is undefined.

B.3 Proofs for Section 5

Definition of gradient homogeneity. For one factor $\theta \in \{B, A\}$, write the response excluding the learning rate and the internal state update as

$$D = \mathcal{D}_\chi(g, z), \qquad z^+ = \mathcal{S}_\chi(g, z), \qquad \theta^+ = \theta + \eta D$$

Here g is the gradient of that factor, χ denotes the optimizer hyperparameters, and the step index may be included in z. Assume that $\mathcal{D}_\chi$ and $\mathcal{S}_\chi$ do not depend directly on θ. If, for every $f > 0$, there are a state correspondence $\mathcal{T}_f$ and a hyperparameter correspondence $\chi \mapsto \chi_f$ such that

$$\begin{aligned} \mathcal{D}_{\chi_f}(fg, \mathcal{T}_f z) &= f^{q_g} \mathcal{D}_\chi(g, z), \\ \mathcal{S}_{\chi_f}(fg, \mathcal{T}_f z) &= \mathcal{T}_f \mathcal{S}_\chi(g, z) \end{aligned} \tag{19}$$

then q_g is the *gradient homogeneity degree*. The correspondences $\mathcal{T}_1 = \mathrm{id}$ and $\mathcal{T}_f \mathcal{T}_h = \mathcal{T}_{fh}$ are taken to be invertible scaling transformations, and the hyperparameter correspondence obeys the same composition law. The second equality is the state-preservation condition of gradient–state covariance. All equalities are required on a common domain where the operations are defined. If $\chi_f = \chi$, this is homogeneity of the optimizer itself with the same hyperparameters; otherwise, it includes the specified hyperparameter transformation. This definition differs from p in equation 5, which also changes the factors.

Lemma B.4 (Gradient homogeneity of SGD and Adam)**.** *Excluding weight decay and clipping, SGD (including momentum and Nesterov momentum) has $q_g = 1$. Adam has $q_g = 0$ under the correspondence*

$$\mathcal{T}_f(m, v, t) = (fm, f^2 v, t), \qquad \epsilon_{\mathrm{adam},f} = f\epsilon_{\mathrm{adam}}$$

If $\epsilon_{\mathrm{adam}} = 0$, the same conclusion holds on the domain where all denominators used are positive. Adam with a fixed positive stabilization constant generally lacks exact gradient homogeneity.

Proof. Write momentum SGD as $m^+ = \beta_1 m + g$, $D = -m^+$. Under $m \mapsto fm$ and $g \mapsto fg$, we simultaneously have $m^+ \mapsto fm^+$ and $D \mapsto fD$. The Nesterov response is likewise a fixed linear combination of g and m^+. SGD without momentum is the special case $D = -g$.

The two Adam moments satisfy

$$m^+ = \beta_1 m + (1 - \beta_1) g, \qquad v^+ = \beta_2 v + (1 - \beta_2) g \odot g$$

and thus scale by f and f^2, respectively. Since the step indices agree across the two runs, the bias-corrected $\widehat{m}^+$ and $\widehat{v}^+$ obey the same scaling. Hence the elementwise response is

$$D_f = -\frac{f\widehat{m}^+}{\sqrt{f^2\widehat{v}^+} + f\epsilon_{\mathrm{adam}}} = -\frac{\widehat{m}^+}{\sqrt{\widehat{v}^+} + \epsilon_{\mathrm{adam}}} = D.$$

The next state of the moments and step index also follows $\mathcal{T}_f$, so both equalities in equation 19 hold. If $\epsilon_{\mathrm{adam}} > 0$ is fixed, the two denominator terms do not scale alike, and the cancellation generally fails. Under the alternative convention that adds the stabilization term *inside* the square root, that term must scale by f^2. □

Lemma B.5 (Transformation of weight decay and gradient clipping under reparameterization)**.** *For one factor in Theorem 2, let $\theta' = c\theta$, $g' = c^{-1}g$, $\eta' = c^{q_g+1}\eta$. If the state and hyperparameters of the base optimizer correspond as in equation 19, the following transformations preserve exact factor correspondence.*

1. *For decoupled weight decay $\theta^+ = (1 - \eta\lambda_{\mathrm{wd}})\theta + \eta D$, use $\lambda'_{\mathrm{wd}} = c^{-(q_g+1)}\lambda_{\mathrm{wd}}$. Equivalently, if the contraction coefficient $d_{\mathrm{wd}} = \eta\lambda_{\mathrm{wd}}$ is specified separately from the learning rate, set $d'_{\mathrm{wd}} = d_{\mathrm{wd}}$.*
2. *For a coupled L^2 term added to the gradient, $g + \lambda_{\mathrm{wd}}\theta$, use $\lambda'_{\mathrm{wd}} = c^{-2}\lambda_{\mathrm{wd}}$.*
3. *For factorwise norm clipping $\mathrm{clip}_\nu(g) := g \min\{1, \nu/\|g\|_F\}$, use $\nu' = c^{-1}\nu$. Here $\mathrm{clip}_\nu(0) = 0$ and $\nu > 0$.*

Apply these rules separately to each factor and each step. If $c_B \neq c_A$, a single finite positive global clipping threshold applied to the joint Euclidean norm of both factor gradients generally cannot be transformed to preserve covariance for every gradient pair.

Proof. For decoupled weight decay, $\eta'\lambda'_{\mathrm{wd}} = \eta\lambda_{\mathrm{wd}}$, so

$$(1 - \eta'\lambda'_{\mathrm{wd}})\theta' + \eta' D' = c\big((1 - \eta\lambda_{\mathrm{wd}})\theta + \eta D\big).$$

For the coupled L^2 term,

$$g' + \lambda'_{\mathrm{wd}}\theta' = c^{-1}g + c^{-2}\lambda_{\mathrm{wd}}\, c\theta = c^{-1}(g + \lambda_{\mathrm{wd}}\theta)$$

so the covariance of the base optimizer applies directly. Factorwise clipping satisfies, for every $f > 0$,

$$\mathrm{clip}_{f\nu}(fg) = fg \min\{1, f\nu/(f\|g\|_F)\} = f\,\mathrm{clip}_{\nu}(g)$$

If the coupled term and clipping are used together, transform the coupled term first as above; the equality then applies under the same operation order.

For the final claim, let $f_B = c_B^{-1}$ and $f_A = c_A^{-1}$. Global clipping multiplies the gradient pair by the common factor

$$\min\left\{1, \frac{\nu}{\sqrt{\|g_B\|_F^2 + \|g_A\|_F^2}}\right\}$$

Suppose one constant ν' satisfies $\mathrm{clip}_{\nu'}(f_B g_B, f_A g_A) = (f_B, f_A)\,\mathrm{clip}_{\nu}(g_B, g_A)$ for every input, where (f_B, f_A) on the right means scalar multiplication of the respective blocks. If $g_A = 0$ and $\|g_B\|_F$ is large enough to activate clipping on both sides, then $\nu' = f_B\nu$ is necessary. Conversely, an input with $g_B = 0$ and sufficiently large $\|g_A\|_F$ requires $\nu' = f_A\nu$. The conditions are incompatible when $f_B \neq f_A$. Equal scaling factors and paths on which clipping is always inactive are outside this counterexample. The claim concerns covariance of the clipping map; it does not rule out coincident final updates for a particular optimizer and input. □

Theorem 2 (Reparameterization Preserving the Initial Product)**.** *Use an optimizer satisfying the gradient–state covariance in equation 19 without weight decay or clipping. If either is used, apply the factorwise transformations in Lemma B.5. For $c_B, c_A > 0$, set*

$$\begin{aligned} B'_0 &= c_B B_0, \qquad A'_0 = c_A A_0, \qquad s' = \frac{s}{c_B c_A}, \\ \eta'_{B,t} &= c_B^{q_{\mathrm{g}}+1}\eta_{B,t}, \quad \eta'_{A,t} = c_A^{q_{\mathrm{g}}+1}\eta_{A,t} \end{aligned} \tag{9}$$

and transform both factors' optimizer states and any required hyperparameters according to the respective gradient scalings c_B^{-1}, c_A^{-1}. Two runs using the same minibatches and randomness then satisfy $B'_t = c_B B_t$, $A'_t = c_A A_t$, and $X'_t = X_t$ at every step.

Proof. Gradient–state covariance refers to the two equalities in Equation 19. Any required hyperparameter transformations are also applied to each factor; the basic proof excludes weight decay and clipping. We proceed by induction on t. At $t = 0$, the definitions give $B'_0 = c_B B_0$ and $A'_0 = c_A A_0$, and the states of both factors correspond. Suppose that, at t, $B'_t = c_B B_t$ and $A'_t = c_A A_t$, and that the states correspond under the respective gradient scalings c_B^{-1} and c_A^{-1}. Then

$$X'_t = s' B'_t A'_t = \frac{s}{c_B c_A}\, c_B c_A\, B_t A_t = X_t$$

and, in particular, $X'_0 = X_0$, so $W'_{\mathrm{frozen}} = W_{\mathrm{frozen}}$ and $W'_t = W_t$. The two runs use the same minibatches and randomness, so their weight gradient G_t is also the same, while their factor gradients satisfy

$$G'_{B,t} = s' G_t (A'_t)^\top = c_B^{-1} G_{B,t}, \qquad G'_{A,t} = s' (B'_t)^\top G_t = c_A^{-1} G_{A,t}$$

Gradient–state covariance gives $D'_{B,t} = c_B^{-q_{\mathrm{g}}} D_{B,t}$ and preserves the correspondence of the B state at the next step. Therefore,

$$B'_{t+1} = c_B B_t + c_B^{q_{\mathrm{g}}+1}\eta_{B,t}\, c_B^{-q_{\mathrm{g}}} D_{B,t} = c_B\big(B_t + \eta_{B,t} D_{B,t}\big) = c_B B_{t+1}$$

and the same holds for A. The second equality in Equation 19 also ensures that the states correspond at the next step. This induction follows the exact update, including the cross term, and involves no approximation. When weight decay or clipping is used, the transformations in Lemma B.5 preserve each induction step. For distinct factor scalings, Adam's stabilizing constant must also be allowed to transform separately for the two factors as $\epsilon'_B = c_B^{-1}\epsilon_B$, $\epsilon'_A = c_A^{-1}\epsilon_A$. The theorem does not apply directly to implementations using a single fixed stabilizing constant or untransformed global clipping. □

Derivation of Equation 10. First compute the update in normalized coordinates directly. Set $b_0 := \sqrt{\rho_0} h_0$ and $a_0 := \sqrt{\rho_0} h_0^{-1}$, and write $B = b_0 \bar{B}$ and $A = a_0 \bar{A}$ at every step. Then $X = \mu_0 \bar{B} \bar{A}$, and the factor gradients are $G_B = s a_0 G \bar{A}^\top$ and $G_A = s b_0 \bar{B}^\top G$. Normalize the states and hyperparameters by the gradient scales $s a_0$ and $s b_0$, respectively. Equation 19 gives

$$\bar{B}^+ = \bar{B} + \frac{\eta_B (s a_0)^{q_g}}{b_0} \bar{D}_B = \bar{B} + \lambda_B \bar{D}_B, \qquad \bar{A}^+ = \bar{A} + \frac{\eta_A (s b_0)^{q_g}}{a_0} \bar{D}_A = \bar{A} + \lambda_A \bar{D}_A,$$

Here $\bar{D}_B$ and $\bar{D}_A$ are the normalized responses to the inputs $G\bar{A}^\top$ and $\bar{B}^\top G$, respectively. Thus μ_0 determines the product representation, while λ_B and λ_A determine the learning rates of the normalized factors. With a learning-rate schedule, these identities and the conditions below must hold at every step. Equal normalized states also require the normalized stabilization constants and other hyperparameters to agree.

Consider two settings with the same normalized shape $(\bar{B}_0, \bar{A}_0)$: $(s, \rho_0, h_0, \eta_B, \eta_A)$ and $(s', \rho_0', h_0', \eta_B', \eta_A')$. Set $c_B := \sqrt{\rho_0'/\rho_0}\, h_0'/h_0$ and $c_A := \sqrt{\rho_0'/\rho_0}\, h_0/h_0'$. Then $B_0' = c_B B_0$, $A_0' = c_A A_0$, and $c_B c_A = \rho_0'/\rho_0$. Hence the condition $s' = s/(c_B c_A)$ of Theorem 2 is equivalent to $s'\rho_0' = s\rho_0$, or $\mu_0' = \mu_0$. The ratios below assume $\eta_B, \eta_A > 0$; when a learning rate is 0, the same conditions follow directly from the normalized update above. If $\mu_0' = \mu_0$, then $s'/s = \rho_0/\rho_0'$, so

$$\frac{\lambda_B'}{\lambda_B} = \frac{\eta_B'}{\eta_B} \Big(\frac{\rho_0}{\rho_0'}\Big)^{q_g} \Big(\frac{\rho_0'}{\rho_0}\Big)^{\frac{q_g - 1}{2}} \Big(\frac{h_0'}{h_0}\Big)^{-(q_g+1)} = \frac{\eta_B'}{\eta_B}\, c_B^{-(q_g+1)}$$

The same calculation gives $\lambda_A'/\lambda_A = (\eta_A'/\eta_A)\, c_A^{-(q_g+1)}$. Therefore, equality of $(\mu_0, \lambda_B, \lambda_A)$ is exactly equivalent to the three conditions of Theorem 2, and the two runs starting from corresponding states have identical weight trajectories.

The proof of Theorem 3 begins by rewriting the training of the amplitude family in coordinates of size $O(1)$.

Lemma B.6 (Displacement coordinates for the amplitude family)**.** *Consider additive updates $B^+ = B + \eta D_B$ and $A^+ = A + \eta D_A$ without separate factor rebalancing, applied by an optimizer of response order p to the amplitude family $(B_0, A_0) = (a\bar{B}_0, a\bar{A}_0)$. For a fixed learning-rate schedule ω_k, set the learning rate at step k to $\eta_k = \zeta \omega_k a^{-(1+p)}$. Let the initial optimizer state be $z_0 = (\bar{z}_0)_a$, and define the displacement coordinates*

$$\beta_k := a(B_k - B_0), \qquad \alpha_k := a(A_k - A_0)$$

and the normalized factors $\tilde{B}_k := \bar{B}_0 + a^{-2}\beta_k$ and $\tilde{A}_k := \bar{A}_0 + a^{-2}\alpha_k$. Then

$$\Delta W_k = s(\beta_k \bar{A}_0 + \bar{B}_0 \alpha_k) + a^{-2} s\, \beta_k \alpha_k$$

holds. Further, let $G_k := \nabla \mathcal{L}_k(W_{\mathrm{pre}} + \Delta W_k)$, $g_{B,k} := s G_k \tilde{A}_k^\top$, and $g_{A,k} := s \tilde{B}_k^\top G_k$, and set

$$(d_{B,k}, d_{A,k}) = \Psi(g_{B,k}, g_{A,k}; \tilde{B}_k, \tilde{A}_k, \bar{z}_k), \qquad \bar{z}_{k+1} = \Psi_z(g_{B,k}, g_{A,k}; \tilde{B}_k, \tilde{A}_k, \bar{z}_k)$$

Then $z_k = (\bar{z}_k)_a$, and

$$\beta_{k+1} = \beta_k + \zeta \omega_k d_{B,k}, \qquad \alpha_{k+1} = \alpha_k + \zeta \omega_k d_{A,k}$$

hold exactly. Here Ψ_z is the optimizer's state update. Thus the amplitude a enters only through terms involving a^{-2}.

Proof. By definition, $B_k = a\tilde{B}_k$ and $A_k = a\tilde{A}_k$, so

$$\Delta W_k = s(B_k A_k - B_0 A_0) = s a^2 \big(\tilde{B}_k \tilde{A}_k - \bar{B}_0 \bar{A}_0\big) = s(\beta_k \bar{A}_0 + \bar{B}_0 \alpha_k) + a^{-2} s\, \beta_k \alpha_k$$

The factor gradients are $G_{B,k} = s G_k A_k^\top = a\, g_{B,k}$ and $G_{A,k} = a\, g_{A,k}$. If $z_k = (\bar{z}_k)_a$, the definition of response order gives $(D_{B,k}, D_{A,k}) = a^p (d_{B,k}, d_{A,k})$, and preservation of the state correspondence after one step gives $z_{k+1} = (\bar{z}_{k+1})_a$. Induction from $k = 0$ shows that, for every k, $z_k = (\bar{z}_k)_a$. Finally,

$$\beta_{k+1} = a\big(B_k + \eta_k D_{B,k} - B_0\big) = \beta_k + a\, \zeta \omega_k a^{-(1+p)}\, a^p d_{B,k} = \beta_k + \zeta \omega_k d_{B,k}$$

and the same argument applies to α. □

This recurrence is an identity for additive factor updates. Composing them with product-preserving rebalancing leaves the product at that step unchanged but can alter the factor displacements and optimizer state. Lemma B.8 treats the displacement map with rebalancing and its limit separately.

Lemma B.7 (Sufficient regularity conditions for finite-amplitude transfer)**.** *Under the assumptions of Lemma B.6, set $h := a^{-2}$ and define the limiting trajectories using the displacement-coordinate map at $h = 0$. Suppose these trajectories have a common bounded neighborhood for $\zeta \in I = [0, Z]$ and $k \le T$. Assume the following conditions hold on that neighborhood for all sufficiently small $0 \le h \le h_0$.*

1. *Each minibatch gradient $\nabla \mathcal{L}_k$ is locally Lipschitz near the corresponding weights.*
2. *The normalized response Ψ and state update Ψ_z are jointly locally Lipschitz in all inputs—the gradient, factors, and state—and those inputs lie in a common compact subset of the interiors of these maps' open domains.*
3. *The normalized initial state is $\bar{z}_0$, independent of amplitude and learning-rate multiplier.*

Then (A1) and (A2) of Theorem 3 hold with $\tau_a = 0$, and both the finite-amplitude and limiting state maps are continuous in (Y, ζ). The same conclusion holds if the initial state and normalized maps admit uniform $O(\tau_a)$ perturbations, with the corresponding τ_a. In that case, continuity is also required separately of the perturbed maps and initial states.

Proof. For $Y = (\beta, \alpha, \bar{z})$, write

$$\tilde{B}_h = \bar{B}_0 + h\beta, \qquad \tilde{A}_h = \bar{A}_0 + h\alpha, \qquad \mathcal{W}_h(Y) = s(\beta \bar{A}_0 + \bar{B}_0 \alpha) + hs\beta\alpha$$

If $\|\beta\|_F, \|\alpha\|_F \le M$ on the common neighborhood, then

$$\|\tilde{B}_h - \bar{B}_0\|_F + \|\tilde{A}_h - \bar{A}_0\|_F \le 2Mh, \qquad \|\mathcal{W}_h(Y) - \mathcal{W}_0(Y)\|_F \le sM^2 h.$$

Let $G_{k,h}(Y) := \nabla \mathcal{L}_k(W_{\mathrm{pre}} + \mathcal{W}_h(Y))$. Condition 1 and the common compact set imply $\|G_{k,h} - G_{k,0}\|_F \le Ch$, and $G_{k,h}$ is uniformly bounded. Consequently,

$$g_{B,k,h} = sG_{k,h}\tilde{A}_h^\top, \qquad g_{A,k,h} = s\tilde{B}_h^\top G_{k,h}$$

also differ from their values at $h = 0$ by a uniform $O(h)$. By condition 2, the response and updated state each differ by $O(h)$. Since $\zeta\omega_k$ is uniformly bounded,

$$\sup_{\zeta,k,Y} \|\mathcal{T}_k^a(Y;\zeta) - \mathcal{T}_k^\infty(Y;\zeta)\| \le C_\chi h.$$

This is (A2), and the initial error is 0. Composing the same polynomial maps with locally Lipschitz maps gives a common Lipschitz constant in Y on the bounded neighborhood. Since $k < T$ is finite, taking the maximum gives (A1). ζ enters through the continuous coefficient $\zeta\omega_k$, so the maps are continuous in (Y, ζ). Uniform $O(\tau_a)$ perturbations add to (A2) by the triangle inequality. □

Optimizer-specific nondegeneracy conditions. Condition 2 of the preceding lemma cannot be replaced by regularity of the gradient with the factors held fixed. In particular, the changes in $\tilde{B}_h = \bar{B}_0 + h\beta$ and $\tilde{A}_h = \bar{A}_0 + h\alpha$ must be included. The following are sufficient conditions for the idealized responses in Lemma 2; every lower bound must be uniformly positive over k, ζ, h.

- The SGD response and momentum update are linear, so no additional denominator condition is needed. For Adam, it suffices that every component of the updated second moment used in the actual response has a positive lower bound. The initial $v_0 = 0$ itself is not the criterion; one must check v_k^+, which enters the denominator.
- LoRA-Muon requires positive lower bounds on the smallest eigenvalues of $\tilde{B}_h^\top \tilde{B}_h$ and $\tilde{A}_h \tilde{A}_h^\top$, and on the smallest singular values of both msign inputs. Here the smallest singular value counts through the smaller matrix dimension, so each polar-decomposition input must have full row or full column rank. A lower bound only on the positive singular values at a fixed rank does not ensure continuity across a whole neighborhood in which the rank may change.

- PoLoRA additionally requires positive lower bounds on $\|u_{\mathrm{out}}\|_\infty$, $\|u_{\mathrm{in}}\|_\infty$, every component of the normalized curvature vector, the smallest eigenvalues of both weighted Gram matrices, each $\|\check{D}_\bullet\|_2$, and $\|\tilde{A}_h\|_2 + \|\tilde{B}_h\|_2$. The curvature-state update and normalization denominators must also be defined throughout the same neighborhood.

Under these conditions, the inverse square root is smooth on the set of positive-definite matrices. For Z of full row rank, the polar factor is $(ZZ^\top)^{-1/2}Z$; for full column rank, it is $Z(Z^\top Z)^{-1/2}$, and is therefore also locally Lipschitz. Matrix norms and elementwise maxima are Lipschitz, and division by denominators with positive lower bounds is also locally Lipschitz. Composing these operations with multiplication and momentum updates gives the sufficient conditions above. The convention of defining a zero response as 0 alone does not make the ideal polar response continuous at zero. A fixed Adam ϵ_{adam} becomes $\epsilon_{\mathrm{adam}}/a$ in the normalized denominator and, under the positive moment lower bound above, gives an example with $\tau_a = O(a^{-1})$. By contrast, a term such as relative decay that survives normalization must be included in the limiting map; we do not assume that an arbitrary stabilization term vanishes.

Lemma B.8 (Displacement coordinates with rebalancing)**.** *Suppose a common product-preserving, degree-one amplitude-homogeneous rebalancing map $\mathcal{R}$ is applied after the additive factor update of the preceding lemma. Assume that the boundedness and regularity conditions of that lemma on the response and state update hold both in a common neighborhood of the* composed *limiting trajectories defined below and at the inputs immediately after the additive update. Write $\mathbf{F}_0 := (\bar{B}_0, \bar{A}_0)$ and assume the following.*

1. *$\mathcal{R}(\mathbf{F}_0) = \mathbf{F}_0$, and $\mathcal{R}$ near $\mathbf{F}_0$ is $C^{1,1}$: it is differentiable with a locally Lipschitz derivative.*
2. *The state is transported to match the rebalancing, and the normalized state transport $\mathcal{C}(\mathbf{F}, \bar{z})$ is jointly locally Lipschitz in a common neighborhood. This transport commutes with the amplitude correspondence of the original state.*

Let the displacement and state immediately after the additive update be $(q_B, q_A, \bar{z}^+)$, with $q := (q_B, q_A)$. Then the exact post-rebalancing displacement map is

$$\mathcal{R}_h(q) := \frac{\mathcal{R}(\mathbf{F}_0 + hq) - \mathbf{F}_0}{h}, \qquad \mathcal{C}(\mathbf{F}_0 + hq, \bar{z}^+), \qquad h = a^{-2} > 0$$

The limiting map uses $D\mathcal{R}(\mathbf{F}_0)q$ and $\mathcal{C}(\mathbf{F}_0, \bar{z}^+)$, respectively. The state map composed with this rebalancing also satisfies (A1) and (A2) with $\tau_a = 0$ and is continuous. The limiting product remains in the initial tangent space. For factor-displacement pairs, write $\mathcal{W}_\infty(q) := s(q_B\bar{A}_0 + \bar{B}_0 q_A)$ as well.

Proof. Immediately before rebalancing, the factors are $a(\mathbf{F}_0 + hq)$. Degree-one homogeneity and $ah = a^{-1}$ give the post-rebalancing displacement $\mathcal{R}_h(q)$ above. By the fundamental theorem of calculus,

$$\mathcal{R}_h(q) = \int_0^1 D\mathcal{R}(\mathbf{F}_0 + thq)q\, dt.$$

If $D\mathcal{R}$ has Lipschitz constant $L_{\mathcal{R}}$, then

$$\|\mathcal{R}_h(q) - D\mathcal{R}(\mathbf{F}_0)q\| \le \tfrac{1}{2}L_{\mathcal{R}}h\|q\|^2.$$

Also, the derivative of $\mathcal{R}_h$ with respect to q is $D\mathcal{R}(\mathbf{F}_0 + hq)$, so its Lipschitz constant is bounded for bounded q, independently of h. Local Lipschitz continuity of the state transport makes its difference from the $h = 0$ value $O(h)$. Composing with the preceding lemma yields both state-map conditions and continuity. Finally, for the product map $\mathcal{P}(B, A) := BA$, differentiate $\mathcal{P} \circ \mathcal{R} = \mathcal{P}$ to obtain

$$D\mathcal{P}(\mathbf{F}_0)D\mathcal{R}(\mathbf{F}_0)q = D\mathcal{P}(\mathbf{F}_0)q, \qquad \mathcal{W}_\infty(D\mathcal{R}(\mathbf{F}_0)q) = \mathcal{W}_\infty(q).$$

Thus the first-order rebalancing term preserves the product's tangent displacement. □

Scope of scalar rebalancing. For the LoRA-Muon map $\mathcal{R}(B, A) = (B/c, cA)$, taking c to be a power of a positive ratio of factor norms makes the map degree-one homogeneous under common amplitude scaling and preserves the product. If both norms have positive lower bounds, then c, $1/c$, and the momentum transport $(m_B, m_A) \mapsto (cm_B, m_A/c)$ are locally Lipschitz. To obtain the

$O(a^{-2})$ conclusion above, the initial gauge must satisfy $c(\mathbf{F}_0) = 1$, and the rebalancing map must satisfy the $C^{1,1}$ condition above. The Frobenius norm has this property near nonzero factors. For the spectral norm, it suffices that the largest singular value of each factor is simple and has a positive gap from the other singular values. Implementations using numerical power iteration, clamping, or threshold branching must verify the same regularity separately in the relevant neighborhood.

Lipschitz continuity of rebalancing alone does not guarantee that $\mathcal{R}_h$ has an $O(h)$ limiting error. Moreover, if $\mathcal{R}(\mathbf{F}_0) \neq \mathbf{F}_0$, then $\mathcal{R}_h(0) = h^{-1}(\mathcal{R}(\mathbf{F}_0) - \mathbf{F}_0)$ diverges, so the theorem above does not apply to displacement coordinates around this fixed initial point. In that case one must align the gauge before training, disable rebalancing, or verify the state-map conditions in different coordinates. For scalar rebalancing with $\mathcal{C}(\mathbf{F}_0, \bar{z}) = \bar{z}$, the product differential identity above makes the limiting product and state dynamics identical to those for the additive update. However, $D\mathcal{R}(\mathbf{F}_0)$ remains in the limit of the factor displacement, so the additive recurrence that omits it must not be applied to post-rebalancing factors.

Theorem 3 (Finite-Horizon Amplitude Transfer)**.** *Fix the optimizer $\mathcal{O}$, s, T, the normalized initial factor shape, and the minibatch order. Let $\eta = \zeta a^{-(1+p)}$ with $\zeta \in I = [0, Z]$. Assume that the normalized states and responses are bounded in a neighborhood of a common limiting trajectory, and that the one-step state maps are uniformly within $O(a^{-2} + \tau_a)$ of a common Lipschitz limiting map. Here $\tau_a \to 0$. Assume also that the initial-state error is at most τ_a and the evaluation loss is Lipschitz. Suppose that each finite-amplitude and limiting state map is jointly continuous in the state and learning-rate multiplier, and that the initial state is continuous in the learning-rate multiplier. Under the concrete state-map conditions in Appendix B.3, for sufficiently large a,*

$$\sup_{\zeta \in I} \left| \mathcal{L}_{a,\mathcal{O},T}(\zeta a^{-(1+p)}) - \Phi_{\mathcal{O},T}(\zeta) \right| \leq \delta_a = O_{T,I}(a^{-2} + \tau_a). \tag{12}$$

For sufficiently large a, b, consider the search interval $I_a := a^{-(1+p)} I$. If the selected learning rate η_a has loss within ϵ_{search} of the optimal loss, then $\widehat{\eta}_b = (a/b)^{1+p} \eta_a$ satisfies

$$\mathcal{L}_{b,\mathcal{O},T}(\widehat{\eta}_b) - \min_{\eta \in I_b} \mathcal{L}_{b,\mathcal{O},T}(\eta) \leq 2(\delta_a + \delta_b) + \epsilon_{\text{search}} \tag{13}$$

Moreover, if $\Phi_{\mathcal{O},T}$ has on I a unique positive minimizer $\zeta^*_{\mathcal{O}}$, then any minimizer on I_a, denoted $\eta^*_{a,\mathcal{O}}$, satisfies $a^{1+p} \eta^*_{a,\mathcal{O}} \to \zeta^*_{\mathcal{O}}$.

To state the state-map conditions in the main text precisely, we use the displacement coordinates of Lemma B.6. When rebalancing is included, the coordinates are defined using the actual factors after rebalancing, and $\mathcal{T}^a_k$ denotes the full step map, including state transport. At learning-rate multiplier ζ, let the state of the amplitude-a run be $Y^a_k := (\beta_k, \alpha_k, \bar{z}_k)$, and write one step as $Y^a_{k+1} = \mathcal{T}^a_k(Y^a_k; \zeta)$. We use the norm $\|Y\| := \|\beta\|_F + \|\alpha\|_F + \|\bar{z}\|$. The weight displacement is $\Delta W_k = \mathcal{W}_a(Y^a_k)$, where

$$\mathcal{W}_a(Y) := s(\beta \bar{A}_0 + \bar{B}_0 \alpha) + a^{-2} s\, \beta\alpha, \qquad \mathcal{W}_\infty(Y) := s(\beta \bar{A}_0 + \bar{B}_0 \alpha)$$

Given a limiting map $\mathcal{T}^\infty_k$, define the limiting trajectory $Y^\infty_k(\zeta)$ from $Y^\infty_0 = (0, 0, \bar{z}_0)$ by iterating $\mathcal{T}^\infty_k$, and, for the evaluation loss $\mathcal{L}_{\text{ev}}$, set

$$\Phi_{\mathcal{O},T}(\zeta) := \mathcal{L}_{\text{ev}}\big(W_{\text{pre}} + \mathcal{W}_\infty(Y^\infty_T(\zeta))\big)$$

The four state-map conditions are as follows. All constants are independent of $\zeta \in I$, $k < T$, and $a \geq a_0$.

- (A1) For every $k \leq T$, throughout the neighborhood of radius $\varrho > 0$ defined by $\mathcal{N}_k(\zeta) := \{Y : \|Y - Y^\infty_k(\zeta)\| \leq \varrho\}$, we have $\|\beta\|_F, \|\alpha\|_F \leq \bar{M}$. For $k < T$, the Lipschitz constant of $\mathcal{T}^\infty_k(\cdot; \zeta)$ is at most $\text{Lip}_{\mathcal{T}} \geq 1$. All state maps are defined on open domains containing these neighborhoods.
- (A2) $\chi_a := \sup_{\zeta \in I,\, k<T,\, Y \in \mathcal{N}_k(\zeta)} \|\mathcal{T}^a_k(Y; \zeta) - \mathcal{T}^\infty_k(Y; \zeta)\| \leq C_\chi(a^{-2} + \tau_a)$, and $\|Y^a_0 - Y^\infty_0\| \leq \tau_a$.
- (A3) $\mathcal{L}_{\text{ev}}$ has Lipschitz constant Lip_{ev} on the weight domain under consideration.
- (A4) $\mathcal{T}^a_k(Y; \zeta)$ and $\mathcal{T}^\infty_k(Y; \zeta)$ are jointly continuous in (Y, ζ), and $Y^a_0(\zeta)$ is continuous in ζ. This condition is used for the existence of finite-amplitude minimizers and convergence of minimizers. Only (A1)–(A3) are needed for the uniform error bound.

For additive factor updates, setting the a^{-2} term in Lemma B.6 to 0 yields $\mathcal{T}_k^\infty$. Lemma B.7 proves (A1) and (A2) with $\tau_a = 0$ under Lipschitz conditions that include factor dependence. This does not mean that the two maps are identical; rather, there is no additional error beyond the $O(a^{-2})$ error specified in (A2). Sufficient conditions that include rebalancing, together with the modified limiting map, are given in Lemma B.8. If stabilizing constants, damping, clipping, or weight decay are present, the convergence error and continuity must be checked for the actual normalized maps.

Proof. **Uniform error.** Fix $\zeta \in I$ and let $E_k := \|Y_k^a - Y_k^\infty\|$. Take a sufficiently large that $K_a := \mathrm{Lip}_{\mathcal{T}}^T(\tau_a + T\chi_a) \le \varrho$. If $E_k \le \varrho$, then $Y_k^a \in \mathcal{N}_k(\zeta)$, and hence

$$E_{k+1} \le \big\|\mathcal{T}_k^a(Y_k^a) - \mathcal{T}_k^\infty(Y_k^a)\big\| + \big\|\mathcal{T}_k^\infty(Y_k^a) - \mathcal{T}_k^\infty(Y_k^\infty)\big\| \le \chi_a + \mathrm{Lip}_{\mathcal{T}} E_k$$

Starting from $E_0 \le \tau_a$, induction gives, for every $k \le T$, $E_k \le \mathrm{Lip}_{\mathcal{T}}^k \tau_a + \sum_{j<k} \mathrm{Lip}_{\mathcal{T}}^j \chi_a \le K_a \le \varrho$, so the trajectory remains within the neighborhoods. The final loss satisfies

$$\begin{aligned}
&\big|\mathcal{L}_{a,\mathcal{O},T}(\zeta a^{-(1+p)}) - \Phi_{\mathcal{O},T}(\zeta)\big| \\
&\quad\le \mathrm{Lip}_{\mathrm{ev}}\Big(\big\|(\mathcal{W}_a - \mathcal{W}_\infty)(Y_T^a)\big\|_F + \big\|\mathcal{W}_\infty(Y_T^a) - \mathcal{W}_\infty(Y_T^\infty)\big\|_F\Big) \\
&\quad\le \mathrm{Lip}_{\mathrm{ev}}\, s\Big(a^{-2}\bar{M}^2 + \big(\|\bar{A}_0\|_2 + \|\bar{B}_0\|_2\big)K_a\Big).
\end{aligned}$$

Denote the right-hand side by δ_a. It is independent of ζ and is $O_{T,I}(a^{-2} + \tau_a)$, which proves Equation 12.

Existence of minimizers and transfer error. Let $f_a(\zeta) := \mathcal{L}_{a,\mathcal{O},T}(\zeta a^{-(1+p)})$. Then $\sup_I |f_a - \Phi_{\mathcal{O},T}| \le \delta_a$. By (A4), the state recursion, and (A3), both f_a and $\Phi_{\mathcal{O},T}$ are continuous in ζ. They therefore attain their minima on the compact interval I, and the continuous one-to-one change of variables $\eta = a^{-(1+p)}\zeta$ gives the existence on I_a of $\eta_{a,\mathcal{O}}^*$. Let $\zeta_a := a^{1+p}\eta_a \in I$. Then $\widehat{\eta}_b = b^{-(1+p)}\zeta_a \in I_b$, and $f_a(\zeta_a) \le \inf_I f_a + \epsilon_{\mathrm{search}}$ implies

$$\begin{aligned}
f_b(\zeta_a) &\le \Phi_{\mathcal{O},T}(\zeta_a) + \delta_b \le f_a(\zeta_a) + \delta_a + \delta_b \\
&\le \inf_I \Phi_{\mathcal{O},T} + \epsilon_{\mathrm{search}} + 2\delta_a + \delta_b \\
&\le \inf_I f_b + \epsilon_{\mathrm{search}} + 2\delta_a + 2\delta_b.
\end{aligned}$$

Since $\inf_I f_b = \min_{\eta \in I_b} \mathcal{L}_{b,\mathcal{O},T}(\eta)$, this proves Equation 13.

Convergence of minimizers. We use the continuity and existence of minimizers established above. Let $\zeta_a^* := a^{1+p}\eta_{a,\mathcal{O}}^*$. Then

$$\Phi_{\mathcal{O},T}(\zeta_a^*) \le f_a(\zeta_a^*) + \delta_a \le f_a(\zeta_{\mathcal{O}}^*) + \delta_a \le \Phi_{\mathcal{O},T}(\zeta_{\mathcal{O}}^*) + 2\delta_a$$

If ζ_a^* does not converge to $\zeta_{\mathcal{O}}^*$, some subsequence converges to $\zeta' \ne \zeta_{\mathcal{O}}^*$. Continuity then gives $\Phi_{\mathcal{O},T}(\zeta') \le \Phi_{\mathcal{O},T}(\zeta_{\mathcal{O}}^*)$, contradicting uniqueness of the minimizer. □

Limiting curve. For ideal additive updates without separate factor rebalancing, $\mathcal{T}_k^\infty$ is the map in Lemma B.6 with a^{-2} set to 0. Here $\Delta W = s(\beta\bar{A}_0 + \bar{B}_0\alpha)$, and the response is computed with the factors fixed at $(\bar{B}_0, \bar{A}_0)$. Thus $\Phi_{\mathcal{O},T}(\zeta)$ is the evaluation loss obtained by training LoRA linearized at the normalized initialization with $\mathcal{O}$ and learning-rate multiplier ζ for T steps. When the response is factor-independent, as for SGD and Adam, the linearized model follows the same trajectory in displacement coordinates regardless of a. That is, when trained at any amplitude with step-matched learning rates, the linearized model has exactly the loss curve $\Phi_{\mathcal{O},T}$.

When the rebalancing of Lemma B.8 is used, its derivative is included in the factor coordinates. If state transport is the identity at the initial gauge, the limiting product and state coincide with those of the linearized model above. For general rebalancing, however, the theorem's state-map conditions alone do not establish identity with the linearized optimizer.

Lemma 3 (Directional Correction for the One-Step Optimal Learning Rate)**.** *Let $H \ne 0$ and $c_i > 0$. Let $\eta_i^\dagger > 0$ be an interior minimizer of ϕ_i, where ϕ_i is C^2 on an open interval containing $[0, \eta_i^\dagger]$ and $0 < \underline{k}_i \le k_i(\eta) \le \overline{k}_i$ on $[0, \eta_i^\dagger]$. Then the step-matched learning rate $\eta_{\mathrm{scale}} = \eta_0^\dagger \|F_0\|_F / \|F_1\|_F$, given with $\eta_0 = \eta_0^\dagger$, satisfies*

$$\frac{c_1}{c_0}\frac{\underline{k}_0}{\overline{k}_1} \le \frac{\eta_1^\dagger}{\eta_{\mathrm{scale}}} \le \frac{c_1}{c_0}\frac{\overline{k}_0}{\underline{k}_1}. \tag{15}$$

Proof. The assumed $\phi_i \in C^2$ permits the integration of derivatives below. In particular, if $\mathcal{L} \in C^2$ near the path, the chain rule for $W_i(\eta) := W + \eta F_i + \eta^2 Q_i$ gives

$$\phi_i''(\eta) = \big\langle F_i + 2\eta Q_i, \nabla^2 \mathcal{L}(W_i(\eta))[F_i + 2\eta Q_i] \big\rangle_F + 2\langle \nabla \mathcal{L}(W_i(\eta)), Q_i \rangle_F$$

Thus k_i includes both the directional Hessian contribution and the path's own quadratic term; a positive semidefinite Hessian alone does not imply $\underline{k}_i > 0$.

The two paths start at the same W, so they share H. We have $\phi_i'(0) = \langle \nabla \mathcal{L}(W), F_i \rangle_F = -\langle H, F_i \rangle_F = -\|H\|_F c_i \|F_i\|_F$, and at an interior minimizer $\phi_i'(\eta_i^\dagger) = 0$. Hence

$$\|H\|_F c_i \|F_i\|_F = \phi_i'(\eta_i^\dagger) - \phi_i'(0) = \int_0^{\eta_i^\dagger} \phi_i''(u)\, du = \|F_i\|_F^2 \int_0^{\eta_i^\dagger} k_i(u)\, du$$

On the integration interval, $\underline{k}_i \le k_i \le \overline{k}_i$, so

$$\frac{\|H\|_F c_i}{\overline{k}_i \|F_i\|_F} \le \eta_i^\dagger \le \frac{\|H\|_F c_i}{\underline{k}_i \|F_i\|_F}$$

Since $\eta_1^\dagger / \eta_{\text{scale}} = (\eta_1^\dagger \|F_1\|_F)/(\eta_0^\dagger \|F_0\|_F)$, applying the bound for $i = 1$ to the numerator and that for $i = 0$ to the denominator gives Equation 15. □

B.4 Practical Recipe

We apply cross-initialization transfer in five steps. (1) Residualize so that the initial function is unchanged, and check the layerwise s and μ_0, the factor shapes, and the common amplitude ratio. (2) If the change preserves the initial product, transfer the factorwise LRs and optimizer states exactly through the optimizer's covariance (Theorem 2). (3) If only the amplitude changes, start the search from Equation 11. (4) If the direction or optimizer changes, recalibrate the LR range with a short pilot that measures response size and loss. (5) If the search minimum lies on the grid boundary, extend the range. Report held-out performance together with the budget utilization ξ, which records the change in geometry after transfer.

Changing the optimizer. Theorem 3 gives each optimizer its own limiting curve $\Phi_{\mathcal{O},T}$, so we calibrate the gap with a short pilot. We fix a rank-8 PiSSA initialization with both factors scaled by about 0.403, and transfer the LR from LoRA-Muon or AdamW to PoLoRA with the pilot rule

$$\widehat{\eta}_{\text{tgt},300} = \eta^{\mathcal{V}}_{\text{src},300} \, \frac{\widehat{\eta}^{\mathcal{C}}_{\text{tgt},100}}{\widehat{\eta}^{\mathcal{C}}_{\text{src},100}}, \tag{20}$$

where $\mathcal{C}$ and $\mathcal{V}$ are separate sets of 256 calibration and 256 LR-selection examples, and subscripts give the number of training steps. We also test two other rules. The norm rule rescales the LR by the median ratio of the single-step movement $\|X_{t+1} - X_t\|_F / \eta_t$ at steps 50, 75, and 100 of the central-LR runs. The same-LR rule reuses the source LR. With all predictions frozen, we train from scratch at each exact predicted LR. We then compare each model against the PoLoRA reference selected by a $\mathcal{V}$ sweep, using 512 held-out examples $\mathcal{H}$.

The prespecified primary rule, the LoRA-Muon pilot ratio, predicts LR 0.06614 (reference 0.09155) and raises CE by 0.000338 (Table 4). This is 0.164% of the reference model's improvement over initialization, and the upper end of the 95% paired-example interval is also within the prespecified 2% tolerance. The norm rule passes when transferring from AdamW but fails the 2% tolerance when transferring from LoRA-Muon, so its validity depends on the optimizer pair.

As an auxiliary check, using the target pilot optimum 0.11144 directly also passes. Its CE minus that of the LoRA-Muon pilot ratio is -0.000177 (95% CI $[-0.000633, 0.000270]$), so the horizon correction borrowed from the source adds no detectable gain. On the target, the pilot ratio costs 1,200 updates (nine 100-step pilots plus one 300-step run), compared with 2,700 for a full nine-point sweep; source tuning is not counted. The norm rule needs one 100-step run at the central target LR.

Table 4: Cross-optimizer LR transfer for rank-8 PiSSA on the last four blocks (single seed, 300 steps). ΔCE is the CE increase on $\mathcal{H}$ over the PoLoRA reference selected on $\mathcal{V}$ (LR 0.09155, CE 0.196827). Brackets show 95% paired-example bootstrap intervals from 5,000 resamples. Passing requires both the point estimate and CI upper bound to meet the prespecified tolerance: 0.004128 for LoRA-Muon and 0.010320 for AdamW.

Source	Rule	Predicted LR	ΔCE [95% CI]	Decision
LoRA-Muon	pilot ratio	0.06614	0.000338 $[0.000015, 0.000684]$	Pass
	norm ratio	0.01195	0.007495 $[0.006391, 0.008641]$	Fail
	same LR	0.00779	0.010839 $[0.009515, 0.012232]$	Fail
AdamW	pilot ratio	0.05572	0.000696 $[0.000305, 0.001103]$	Inconclusive[†]
	norm ratio	0.07043	0.000213 $[-0.000056, 0.000499]$	Pass
	same LR	0.00114	0.131065 $[0.124477, 0.137941]$	Fail

[†]The AdamW pilot minimum remained at the search boundary, so its ratio was inconclusive. The norm rule is independent of this minimum and passes the stricter 2% tolerance. The prespecified primary comparison was the LoRA-Muon pilot ratio.

C Experimental Details

C.1 Common Setup and Aggregation Conventions

The experiments use Qwen3-0.6B-Base and the first 100,000 training examples from MetaMathQA. Full training uses a global batch size of 64, a maximum sequence length of 512, 1,563 optimizer steps, and cosine decay after a 3% warmup; weight decay and LoRA dropout are both set to 0. Unless otherwise specified, the gradient-clipping threshold is 1, and we use native BF16 forward computation with FP32 master factors. The shorter 100-, 200-, and 300-step experiments run the corresponding initial segments of this 1,563-step schedule. The rank-128 experiments use 196 adapters corresponding to `q/k/v/o_proj` and `gate/up/down_proj` in all 28 Transformer blocks, whereas the rank-8 experiments use 28 adapters corresponding to the same seven projections in the last four blocks. Both settings use $s = \alpha/r = 1$. Thus, both the rank and the adapter placement change between the two settings. For learning-rate grid comparisons, we select the highest accuracy or lowest CE among the completed runs for each condition. Unless stated otherwise in the main text, all reported results are observations from a single training seed. The training and data-shuffling seed is 42, and "trial 1" in run names is a repetition identifier. The GSM8K/MATH mean accuracy is the arithmetic mean of the accuracies on the two tasks. CE is token-weighted cross-entropy: we sum the negative log-likelihoods of all supervised tokens and divide by the number of such tokens, rather than taking the arithmetic mean of per-example averages.

Index the adapters by $\ell = 1, \ldots, m$, and let $w_\ell(t) := \|\Delta W_\ell(t)\|_F^2$ and $\sigma_{\ell,0} := \sigma_{r_\ell}(X_{\ell,0}) > 0$. The run-level aggregations of the observables in Equation 3 are

$$\varepsilon_{\mathrm{run}}(t) = \frac{\sum_\ell w_\ell(t)\varepsilon_\ell(t)}{\sum_\ell w_\ell(t)} = \frac{\sum_\ell \|N_\ell(t)\|_F^2}{\sum_\ell \|\Delta W_\ell(t)\|_F^2}, \tag{21}$$

$$\left(\frac{x^2}{r}\right)_{\mathrm{run}}(t) = \frac{\sum_\ell w_\ell(t) x_\ell(t)^2/r_\ell}{\sum_\ell w_\ell(t)} = \frac{\sum_\ell w_\ell(t)^2/(r_\ell \sigma_{\ell,0}^2)}{\sum_\ell w_\ell(t)}. \tag{22}$$

If all $w_\ell(t)$ are zero, both aggregate values are defined to be zero. The second expression is a weighted average of quantities that are first normalized by the smallest singular value of each layer. It therefore generally differs from dividing the global displacement norm by a single average singular value. We do not define this x_ℓ for vanilla LoRA, whose initial product does not have rank r_ℓ, nor do we add an arbitrary small constant to its zero singular values and aggregate it together with full-rank initializations.

C.2 Fixed-Basis Initialization and Numerical Precision

The fixed-basis initialization in the main text uses the DST-I basis of Zhang et al. (2025b). For $k \geq r$, define $\Phi_k \in \mathbb{R}^{k \times r}$ by

$$(\Phi_k)_{ij} = \sqrt{\frac{2}{k+1}} \sin\left(\frac{ij\pi}{k+1}\right), \qquad 1 \leq i \leq k, \quad 1 \leq j \leq r,$$

so that $\Phi_k^\top \Phi_k = I_r$. For a positive layerwise coefficient b_ℓ, we set $B_{\ell,0} = b_\ell \Phi_{d_{\mathrm{out},\ell}}$ and $A_{\ell,0} = b_\ell \Phi_{d_{\mathrm{in},\ell}}^\top$. Here, b_ℓ is the positive coefficient of the layerwise DST basis. When varying the common gain, we apply the same relative amplitude multiplier to both factors in every layer.

As in the main text, the gain used in the magnitude sweep is denoted by g_{init}. Define $\mathrm{Var}(M)$ as the centered sample variance of the matrix entries, namely $\sum_{ij}(M_{ij} - \overline{M})^2/(n_M - 1)$ for a matrix with n_M entries. Then

$$b_\ell^4 = g_{\mathrm{init}} \frac{\mathrm{Var}(W_{\mathrm{pre},\ell})}{\mathrm{Var}(\Phi_{d_{\mathrm{out},\ell}} \Phi_{d_{\mathrm{in},\ell}}^\top)},$$

and therefore $\mathrm{Var}(B_{\ell,0}A_{\ell,0}) = g_{\mathrm{init}} \mathrm{Var}(W_{\mathrm{pre},\ell})$. The sweep in the main text uses the common gains $0.001, 0.01, 0.1, 1, 10$ for all layers. The symbols `Q` and `Q_init` in earlier run records are implementation names for this gain. They should not be confused with the cross term $Q = sD_BD_A$ in Equation 2.

Fixed-PiSSA and Fixed-GA are constructed to match the layerwise product magnitudes of their reference initializations. For a reference initialization $(s_{\mathrm{ref}}, B_{\mathrm{ref}}, A_{\mathrm{ref}})$, we set

$$b_\ell^4 = \frac{\mathrm{Var}(s_{\mathrm{ref}} B_{\mathrm{ref},\ell} A_{\mathrm{ref},\ell})}{s^2 \mathrm{Var}(\Phi_{d_{\mathrm{out},\ell}} \Phi_{d_{\mathrm{in},\ell}}^\top)}$$

to match the centered variance of the layerwise products. This matching does not require the full singular-value distribution, Frobenius norm, or factor RMS in every layer to exactly equal the corresponding reference value. In the comparison of four initializations, we additionally apply a common multiplier to both factors of each method so that the mean, over 196 adapters, of the sum of the entrywise RMS values of the two stored initial factors is approximately 0.0181. For every nonzero initialization, we reconstruct the frozen weight from the original pretrained weight as $W_{\mathrm{frozen}} = W_{\mathrm{pre}} - sB_0A_0$. Because BF16 rounding occurs during storage, we distinguish exact equality of the mathematical initial functions from numerical equality.

Initial factor magnitude metric. The effective initialization scale used in the run records is

$$\beta_{\mathrm{eff},\ell} := \mathrm{rms}(A_{\ell,0}) + \mathrm{rms}(B_{\ell,0}), \qquad \mathrm{rms}(M) := \frac{\|M\|_F}{\sqrt{n_M}}.$$

When reporting a single value for a run, we use the arithmetic mean of this layerwise quantity over adapters. For the DST initialization above, exact arithmetic gives $\beta_{\mathrm{eff},\ell} = b_\ell(d_{\mathrm{in},\ell}^{-1/2} + d_{\mathrm{out},\ell}^{-1/2})$. Thus, scaling the common amplitude of fixed factor shapes by a also scales this value by a. However, equal means across different factor shapes do not imply equal product spectra.

FP32 adapter branch. For remeasurements with `LORA_FP32_BRANCH=1`, the base layer is evaluated using the existing BF16 path, while autocast is disabled for the two adapter matrix multiplications and the input, A, and B are computed in FP32. The FP32 adapter output is converted to the dtype of the base-layer output before addition. Thus, the base model and stored residual remain in BF16; this is not a full-network FP32 experiment. Among the 30 markers in Figure 2, 11 prespecified FP32-branch remeasurements take precedence, and the remaining 19 use BF16 measurements. This precision change reduces the loss of small factor changes relative to large initial entries before BF16 multiplication, but does not remove residual rounding or the output dtype conversion.

C.3 Support Distance and Capture Ratio of the Same Descent Field

Let $V_{\ell,t}, V_{\ell,u}$ be orthonormal bases of the current and reference right supports. The normalized chordal distance is

$$d_{V,\ell}(t,u) := \frac{\|P_{V_{\ell,t}} - P_{V_{\ell,u}}\|_F}{\sqrt{2r_\ell}} = \sqrt{1 - \frac{\|V_{\ell,t}^\top V_{\ell,u}\|_F^2}{r_\ell}} = \frac{\|\sin\Theta(V_{\ell,t}, V_{\ell,u})\|_F}{\sqrt{r_\ell}}.$$

The quantity $d_V(t,u) = m^{-1}\sum_\ell d_{V,\ell}(t,u)$ is an arithmetic mean over adapters and differs from the energy-weighted aggregation in Equation 21. Each layerwise distance lies in $[0,1]$, and its square is the mean squared sine of the principal angles. This identity and the orthogonal tangent-space projection are proved in Lemma B.1.

The descent field $H = -\nabla_W \mathcal{L}$ is the negative full-matrix gradient with respect to the adapted weight, before clipping, momentum, or adaptive preconditioning. Using $P_{T(U,V)}H = P_U H + HP_V - P_U H P_V$, the quantity C_H is computed immediately after the intervention described in the next section. The actual branched model is held fixed, and the same $H_{\ell,b}$ is evaluated over the next 10 training batches:

$$C_H(U,V) = \frac{\sum_{b=1}^{10}\sum_{\ell=1}^{28}\|P_{T(U_\ell,V_\ell)}H_{\ell,b}\|_F^2}{\sum_{b=1}^{10}\sum_{\ell=1}^{28}\|H_{\ell,b}\|_F^2}.$$

We compare the actual post-step support with eight randomized post-step supports, while keeping U common. Because energies are summed across layers and batches before taking the ratio, this quantity is not a simple average of layerwise capture ratios. The model is not updated on the subsequent batches during this measurement. This should be distinguished from a comparison in which the actual and randomized branches are trained separately and evaluated using their different current gradients.

C.4 Work-Matched Right-Rotation Intervention

The intervention in the main text is performed at step 100 of a 200-step trajectory using a rank-8 initialization obtained by jointly scaling the two original LoRA-GA factors by approximately 0.118 and AdamW with peak LR 2.84528×10^{-4}. Clipping was inactive throughout all 200 steps of this trajectory. Let (B^-, A^-) denote the pre-step factors in each layer and (B^+, A^+) the actual post-step factors, and let $P_- = P_{\mathrm{row}\,A^-}$. In the decomposition $A^+ = A^+P_- + E$, where $E := A^+(I - P_-)$, we hold B^+ and A^+P_- fixed and change only the direction of E. The eight controls are generated using fixed random seeds selected without inspecting the loss.

Specifically, for the current pre-step descent field H, define $\mathcal{S} = \mathrm{row}\,A^- + \mathrm{row}(s(B^+)^\top H)$ and hold the $\mathcal{S}$ component of E fixed. We construct E' by randomizing the right orthonormal basis of the $\mathcal{S}^\perp$ component while preserving its row Gram matrix. The resulting $A' = A^+P_- + E'$ preserves the following quantities layerwise in exact arithmetic:

$$A'A'^\top = A^+(A^+)^\top, \qquad A'P_- = A^+P_-,$$
$$\langle H, sB^+A' - sB^-A^-\rangle_F = \langle H, sB^+A^+ - sB^-A^-\rangle_F.$$

The second condition, together with $A^-P_- = A^-$, also preserves $A'(A^-)^\top = A^+(A^-)^\top$. Consequently, the first two conditions preserve $X^+(X^+)^\top$ and $(X^+ - X^-)(X^+ - X^-)^\top$, and therefore preserve the singular values of the product and product step, the Frobenius norm of the step, the distribution of pre/post support principal angles, and $d_{V,\ell}$. Preservation of the principal angles follows because the singular values of the cross Gram matrix between the orthonormal row bases of the two full-row-rank factors,

$$(A^+(A^+)^\top)^{-1/2}A^+(A^-)^\top(A^-(A^-)^\top)^{-1/2},$$

remain unchanged. The final condition matches the current raw-gradient work. The implementation checks these invariants and the current work error using FP32 factors. In the step-100 comparison reported in the main text, the maximum absolute matching error in $d_{V,\ell}$ is 1.50×10^{-9}.

After the intervention, each branch starts from the same Adam moments, scheduler state, and RNG state as immediately after the actual step and uses the same future minibatch order. We evaluate CE on 128 examples selected in advance from outside the training split after excluding duplicates

based on normalized question strings. The horizon $h = 1$ is immediately after the intervention, while $h = 10$ is immediately after the tenth step including the intervention, corresponding to nine additional updates. A loss difference $\mathrm{CE}_{\mathrm{random}} - \mathrm{CE}_{\mathrm{actual}} > 0$ indicates an advantage for the actual direction. This matching does not simultaneously match the cumulative departure ratio ε relative to the initial tangent space or the native-BF16 output-change norm. The eight directions are conditional controls constructed from a single trajectory, rather than eight independent training seeds. The paired bootstrap over evaluation examples likewise does not represent variation across training seeds.

C.5 Initial-Tangent-Space-Restricted and Core-Only Models

Starting from the same $B_0, A_0, W_{\mathrm{frozen}}$ as unconstrained LoRA, we implement the forward pass of the initial-tangent-space-restricted model as

$$W_{T_0} = W_{\mathrm{frozen}} + s\{B_0 A + (B - B_0)A_0\} = W_{\mathrm{pre}} + s\{B_0 \Delta A + \Delta B A_0\}.$$

The factors A, B remain trainable, while A_0, B_0 are stored as fixed buffers. Thus, $\Delta W_{T_0} \in T_0$, and the cumulative cross term is excluded. This differs from an algorithm that projects the result of each unconstrained step after the update, and $X_0 + \Delta W_{T_0}$ need not retain rank r throughout training. The initial output and first factor gradients are identical to those of unconstrained LoRA. For the LoRA-Muon restricted model, the preconditioner Gram matrices are also computed from the fixed $B_0^\top B_0, A_0 A_0^\top$. Because the linearized forward pass is not invariant to factor gauge transformations, we disable scalar rebalancing and also disable it in the corresponding unconstrained primary control. Under this condition, we verified that the first nonzero factor update is bitwise identical between the two models. Because the unconstrained product step still contains $\eta^2 Q$, this does not imply that the product steps are identical.

The core-only model fixes both A_0, B_0 and trains only $M \in \mathbb{R}^{r \times r}$ from $M_0 = 0$:

$$W_{\mathrm{core}} = W_{\mathrm{frozen}} + sB_0(I_r + M)A_0 = W_{\mathrm{pre}} + sB_0 M A_0.$$

If B_0, A_0 have full rank, this correction represents every core component within the initial supports. For the LoRA-Muon comparison, the core optimizer uses momentum $Z^+ = 0.95Z + 0.05G_M$ for $G_M = \nabla_M \mathcal{L}$ and computes

$$M^+ = M - \frac{\eta}{2} S_B \,\mathrm{msign}(S_B Z^+ S_A) S_A, \qquad S_B = (B_0^\top B_0)^{-1/2}, \quad S_A = (A_0 A_0^\top)^{-1/2}.$$

The implementation uses the same Gram damping and polar computation as the compared LoRA-Muon model. The core-only comparison uses this optimizer with the same LR specified for the unconstrained model.

The PiSSA comparison in the main text uses rank 8, 28 adapters, native BF16, and 1,563 steps. We searched the learning rate separately for the unconstrained and restricted models using the original PiSSA factors and initializations obtained by jointly scaling them by 0.4026 and 0.2013. BF16 rounding was applied during storage. There are grid points on both sides of the selected LR of approximately 0.031178. For each model, LR selection and reported CE use the same 128 examples from outside the training split, containing 26,656 supervised tokens. Thus, this CE is not an independent final evaluation separate from LR selection. Because the Gram responses and training trajectories of the two models may diverge from the second step onward, we do not interpret their final CE difference as the causal effect of the normal component alone.

C.6 Cross-Optimizer LR Transfer Using Short Pilots

We fix the 28 adapters of a rank-8 initialization obtained by jointly scaling the two original PiSSA factors by 0.4026 and transfer from LoRA-Muon or AdamW to PoLoRA. Each run starts with fresh optimizer states using the same initialization and data order. We select 1,024 new examples outside the 100,000 training examples and the existing 128-example probe, and divide them into 256 calibration examples $\mathcal{C}$, 256 LR-selection examples $\mathcal{V}$, and 512 final-evaluation examples $\mathcal{H}$. Before selection, we remove duplicate question strings after normalizing whitespace and case across the training set, the existing probe, and the three splits. We retain only examples with supervised tokens remaining after the 512-token limit. This procedure does not guarantee the removal of paraphrases or questions derived from a common source.

We begin with seven points ranging from $1/8$ to 8 times the central LR of each optimizer. For the internal grid minimum of the 100-step $\mathcal{C}$ loss, we perform a quadratic interpolation in log-LR using the minimum and its two neighboring points. If the minimum is at the boundary, we inspect only $\mathcal{C}$ and expand the range by factors of two at most twice. If the minimum remains at the boundary, the success decision for the corresponding pilot ratio is left inconclusive. Equation 20 uses only this calibration result and the source model's 300-step grid minimum on $\mathcal{V}$. The norm rule measures

$$n_{\mathcal{O},t} := \frac{\left(\sum_\ell \|X_{\ell,t+1} - X_{\ell,t}\|_F^2\right)^{1/2}}{\eta_t}$$

at $t = 50, 75, 100$ in the central-LR run and multiplies the source LR by the median of $n_{\mathrm{src},t}/n_{\mathrm{tgt},t}$. This is an empirical correction based on the actual finite step, including the cross term. After freezing the predicted LRs and input hashes, we train the target from scratch for 300 steps using each exact predicted LR. The reference is the model with the lowest $\mathcal{V}$ loss in a separate 300-step target sweep. The final target losses on $\mathcal{V}$ and $\mathcal{H}$ are not used to compute the predicted LRs.

Let $L_{\mathcal{H},0}$ and $L_{\mathcal{H},\mathrm{ref}}$ denote the CE values of the initial and reference models, and define $\mathcal{I}_\mathcal{H} := L_{\mathcal{H},0} - L_{\mathcal{H},\mathrm{ref}}$. For the primary LoRA-Muon→PoLoRA comparison, the prespecified loss tolerance is

$$\delta_{\mathrm{main}} = \max\{0.002,\ 0.02\mathcal{I}_\mathcal{H}\} \quad \text{nats/token},$$

while the auxiliary AdamW comparison uses 0.05 instead of 0.02. We do not declare success if the reference improvement satisfies $\mathcal{I}_\mathcal{H} \le 0.02$ or if the target reference LR lies at a grid boundary. Otherwise, we compare both the point estimate of $L_{\mathcal{H},\mathrm{transfer}} - L_{\mathcal{H},\mathrm{ref}}$ and the upper endpoint of its 95% interval from 5,000 paired-example bootstrap resamples against the tolerance. The bootstrap resamples examples in pairs and recomputes the token-weighted CE. The realized tolerance for the primary comparison is 0.004128, and the interval upper bound reported in the main text, 0.000684, is below this value.

The pilot-only control that directly uses the target pilot LR was registered separately before analyzing the final target results. This control also passed the criterion, so no additional benefit from correcting for the source training horizon was established. After range expansion, the target cost of the pilot-ratio procedure is $9 \times 100 + 300 = 1,200$ updates, whereas the complete target sweep costs $9 \times 300 = 2,700$ updates. Source-tuning cost is separate, and the target cost of the norm rule is $100 + 300 = 400$ updates. For validation, we also performed the full target sweep used to obtain the reference. All decisions are conditional results from the first 300 steps of the original schedule and a single training seed.

C.7 Fixed-State GA Directional Diagnostic and 200-Step Amplitude Comparison

The GA-versus-Fixed-GA directional diagnostic in the main text fixes the rank-128 initial checkpoint and uses the same 128 examples sampled without replacement from the 100,000 training examples with seed 1234. The full model is loaded in FP32, and forward and gradient computations are performed in evaluation mode with TF32 disabled. Factor directions and matrix diagnostics are computed in FP64, while CE is evaluated by inserting candidate factors into the FP32 model. The fresh Adam response is $D = -G/(|G| + 10^{-8})$, and the LoRA-Muon response uses Gram inverse square roots and polar directions without momentum. Neither case reuses accumulated training states or retrains the optimizer for each candidate.

We fix the once-computed (D_B, D_A) and evaluate the exact candidate $(B_0 + \eta D_B, A_0 + \eta D_A)$, so the cross term $\eta^2 Q$ is included in the forward pass. The network-level quantities $\|H\|_F$, $\|F\|_F$, and $\langle H, F\rangle_F$ are computed by first summing squared norms or inner products over adapters. For the first-order product-step length $u := \eta\|F\|_F$, we evaluate 34 logarithmically spaced candidates for which $u/\{\|H\|_F \max(c, 10^{-6})\}$ ranges from 10^{-4} to $10^{1.5}$, and stop the search if the loss exceeds the initial loss by more than 1 nat/token. An interior grid minimum is refined by quadratic interpolation in log-u using the minimum and its two neighbors. If the interpolated minimum lies outside the neighboring interval or the quadratic coefficient is not positive, we use the grid minimum.

In this record, the LoRA-Muon ratio of c between LoRA-GA and Fixed-GA is approximately 1.96, the ratio of u at the minimum loss is approximately 0.483, and the corresponding fresh-Adam ratio of u is approximately 1.31. Using the tokenizer stored with each initialization, the numbers of

supervised tokens differ: 22,100 for GA and 22,251 for Fixed-GA. Small differences also remain in the numerical initial weights and gradient norms. These values, including the effective curvature inferred from the curves, are diagnostics comparing the optimizer-specific local responses of the stored initializations. They do not verify the common-W and common-loss conditions or the derivative bounds over the full interval in Lemma 3.

200-step amplitude comparison. A separate experiment compares eight conditions formed by crossing AdamW and LoRA-Muon, peak LRs $1.1246970065 \times 10^{-4}$ and $1.9999946768 \times 10^{-4}$, and relative factor amplitudes 1 and 0.56235 along the same DST directions of rank-128 Fixed-PiSSA. Each condition uses the first 200 steps of the same 1,563-step schedule. The two AdamW conditions reuse prefixes of completed runs with identical settings. For the smaller initialization, we apply the multiplier 0.56235 to both factors and reconstruct the residual from the original pretrained weight. BF16 rounding during storage introduces a small error into the realized factor-magnitude ratio. We use native BF16 forward computation and do not enable the FP32 adapter branch. Both optimizers use a clipping threshold of 1 and weight decay 0, while LoRA-Muon uses momentum 0.95 and rebalancing at every step.

The reported product-step norm is $\{\sum_\ell \|X_{\ell,t+1} - X_{\ell,t}\|_F^2\}^{1/2}$. When measuring the first-order factor response of LoRA-Muon, we first undo scalar rebalancing to remove pure gauge changes. The values 0.578–0.586 and 0.9999 reported in the main text are the ratios of the sampled-step norms for the smaller and larger initializations under the same LR, averaged over steps 10–50. In increasing-LR order, the clipping counts were $7/5$ for the larger AdamW initialization, $0/0$ for the smaller AdamW initialization, $29/22$ for the larger LoRA-Muon initialization, and $2/1$ for the smaller LoRA-Muon initialization. These values therefore include differences in the gradient and clipping histories of the respective trajectories. This experiment compares amplitude-dependent product-step magnitudes over the initial 200-step interval.

D Limitations

Our experiments use a single model, Qwen3-0.6B-Base, and a single fine-tuning task, MetaMathQA with evaluation on GSM8K and MATH. Each configuration was trained once with a single seed. Extending the experiments to larger models, other tasks, and multiple seeds would strengthen our conclusions. In addition, our learning-rate transfer analysis focuses on changes in initialization scale, and a fuller treatment of changes in initialization direction and optimizer is left for future work.

E Notation

Tables 5–9 collect the notation used in the paper, grouped by topic. The subscript 0 denotes initialization, the superscript $+$ the next iterate, and the subscript t the step index; the current index is omitted when no confusion arises. Throughout a run, s and W_{frozen} are fixed. We write Δ for cumulative changes from initialization and δ for one-step changes. Quantities defined through singular vectors, such as T_X and Θ, are used on the rank-r manifold $\mathcal{M}_r$.

Table 5: Weights, factors, and initialization.

Symbol	Meaning	Definition or domain
$d_{\text{out}}, d_{\text{in}}$	Output/input dimensions	$W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$.
r	LoRA rank	Internal factor dimension; $r \leq \min(d_{\text{out}}, d_{\text{in}})$.
B, A	Output-/input-side factors	$B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$.
s	LoRA scale	Positive multiplier in $X = sBA$, fixed during a run. Standard LoRA uses $s = \alpha/r$.
α	Scaling hyperparameter	Sets s at a given rank; distinct from the learning rate.
W_{pre}	Pretrained weight	Original weight of the layer.
X	Adapter product	$sBA \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$.
X_0	Initial adapter product	sB_0A_0.
W_{frozen}	Frozen base weight	$W_{\text{pre}} - X_0$; fixed during training.
W	Effective weight	$W_{\text{frozen}} + X$; equals W_{pre} at initialization.
ΔW	Weight displacement	$W - W_{\text{pre}} = X - X_0$.
$\Delta B, \Delta A$	Cumulative factor changes	$B - B_0$ and $A - A_0$. Then $\Delta W = s(\Delta B\, A_0 + B_0\, \Delta A) + s\, \Delta B\, \Delta A$, and only the last term can leave T_0.
$\delta B, \delta A, \delta W$	One-step changes	$\delta B = B^+ - B$, $\delta A = A^+ - A$, and $\delta W = W^+ - W = X^+ - X$.

Table 6: Tangent-space geometry and the curvature budget.

Symbol	Meaning	Definition or domain
Rank-r manifold and supports		
$\mathcal{M}_r$	Rank-r manifold	$\{Y \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}} : \operatorname{rank} Y = r\}$.
U, Σ, V	Thin SVD of X	$X = U\Sigma V^\top$ with $U^\top U = V^\top V = I_r$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$.
$\sigma_i(\cdot)$	Singular values	$\sigma_1 \geq \cdots \geq \sigma_r$; $\sigma_r(X) > 0$ on $\mathcal{M}_r$.
U_0, Σ_0, V_0	Initial SVD	$X_0 = U_0 \Sigma_0 V_0^\top$.
$U_{0\perp}, V_{0\perp}$	Orthogonal complements	Orthonormal bases of the complements of $\operatorname{col} U_0$ and $\operatorname{col} V_0$.
P_U, P_V	Left/right support projections	$UU^\top$ and $VV^\top$; projections onto $\operatorname{col} X$ and $\operatorname{row} X$.
Tangent space and block decomposition		
T_X	Tangent space at X	$\{P_U Y + Y P_V - P_U Y P_V\}$; dimension $r(d_{\text{out}} + d_{\text{in}} - r)$.
P_{T_X}	Tangent projection	$P_{T_X} Y = P_U Y + (I - P_U) Y P_V$.
T_0	Initial tangent space	$T_{X_0} = \{Y : U_{0\perp}^\top Y V_{0\perp} = 0\}$; the image of J at (B_0, A_0) when both factors have full rank.
J	Factor-to-product Jacobian	$J[Z_B, Z_A] = s(Z_B A + B Z_A)$ at the current (B, A).
C	Core	$U_0^\top \Delta W \, V_0 \in \mathbb{R}^{r \times r}$.
L, R	Arms	$L = U_{0\perp}^\top \Delta W \, V_0$ and $R = U_0^\top \Delta W \, V_{0\perp}$; each rotates one support.
N	Departure component	$U_{0\perp}^\top \Delta W \, V_{0\perp}$; the normal-space component, zero on T_0.
K	Core matrix	$U_0^\top X V_0 = \Sigma_0 + C$. If $\operatorname{rank} X = r$ and K is invertible, $N = LK^{-1}R$.
Observables and budget		
$\Theta(U, U_0)$	Principal angles	Between $\operatorname{col} U$ and $\operatorname{col} U_0$; $\Theta(V, V_0)$ likewise.
$d_{V,\ell}(t, 0)$, $d_V(t, 0)$	Right-support chordal distance	$d_{V,\ell}(t, 0) = \sqrt{1 - \lVert V_{\ell,t}^\top V_{\ell,0} \rVert_F^2 / r}$ for adapter ℓ; for multi-adapter plots, d_V is the simple adapter mean.
ε	Departure ratio	$\lVert N \rVert_F^2 / \lVert \Delta W \rVert_F^2 \in [0, 1]$; 0 when $\Delta W = 0$. Write ε_t at step t and aggregate layers by displacement-energy weights.
x	Normalized displacement	$\lVert \Delta W \rVert_F / \sigma_r(X_0)$.
$c_\perp$	Departure coefficient	$\varepsilon r / x^2$ for $x > 0$.
γ	Schur gap	$\sigma_r(X_0) - \lVert C \rVert_2$.
II_X	Second fundamental form	Normal projection of the ambient derivative of tangent vector fields; Equation 18.
κ	Maximum normal curvature	$\sup_{Z \in T_X, \, \lVert Z \rVert_F = 1} \lVert \mathrm{II}_X(Z, Z) \rVert_F = 1/\sigma_r(X)$ when $r < \min(d_{\text{out}}, d_{\text{in}})$; otherwise 0.
$\sigma_r(X_0)$	Budget scale	Initial spectral scale used to normalize the product path length for rank-r X_0.
Λ_t	Product path length	$\sum_{j<t} \lVert X_{j+1} - X_j \rVert_F$.
ξ_t	Budget usage	$\Lambda_t / \sigma_r(X_0)$ for $\sigma_r(X_0) > 0$, computed from the actual product path.

Table 7: Gradients and optimizer response.

Symbol	Meaning	Definition or domain
Gradients and optimizer response		
$\mathcal{L}$	Loss	A scalar objective; dependence on data and other layers is stated when needed.
G	Weight gradient	$\nabla_W \mathcal{L} = \nabla_X \mathcal{L}$ for fixed W_{frozen}.
H	Descent field	$-G$; not a Hessian.
G_B, G_A	Factor gradients	$sGA^\top$ and $sB^\top G$.
z	Optimizer state	Moments, momentum, step count, and other state required by the update rule.
Ψ	Optimizer response map	$(D_B, D_A) = \Psi(G_B, G_A; B, A, z)$. For SGD, $(D_B, D_A) = (-G_B, -G_A)$.
D_B, D_A	Update directions	Optimizer outputs with the learning rates excluded; not necessarily unit-norm.
η_B, η_A	Factor learning rates	$B^+ = B + \eta_B D_B$ and $A^+ = A + \eta_A D_A$.
η	Common learning rate	Used when $\eta_B = \eta_A = \eta$.
F	Product response	$s(D_B A + B D_A)$, the first-order change of X per unit learning rate.
Q	Cross term	$sD_B D_A$. With a common η, $X^+ - X = \eta F + \eta^2 Q$ exactly.
m, v	Adam moments	Factor-shaped first and second raw moments.
β_1, β_2	Moment decay rates	Scalars in $[0, 1)$.
ϵ_{adam}	Adam stabilizer	Positive constant in the denominator; unrelated to the departure ratio ε.
λ_{wd}	Weight-decay coefficient	Coupled and decoupled weight decay are distinguished.

Table 8: Initialization scales and normalized learning rates.

Symbol	Meaning	Definition or domain
Scale variables and response order		
a	Common amplitude	For a fixed reference pair with $\|\bar{B}_0\bar{A}_0\|_F = \sqrt{r}$, $(B_0, A_0) = (a\bar{B}_0, a\bar{A}_0)$, so $X_0 = a^2\bar{X}_0$.
p	Response order	Degree of positive homogeneity under corresponding optimizer states: $\Psi(aG_B, aG_A; aB, aA, z_a) = a^p\Psi(G_B, G_A; B, A, z)$ for a state map preserved by one step. Ideal covariant SGD 1, Adam 0, LoRA-Muon and PoLoRA -1.
q_{g}	Gradient-scaling order	Degree under gradient/state scaling at fixed factors for factor-independent optimizers, with a state map preserved by one step: SGD 1, matched Adam 0 (Section 5.1). Distinct from the amplitude response order p; see Lemma B.4.
ρ	Raw product magnitude	$\|BA\|_F/\sqrt{r}$.
μ	Adapter magnitude	$s\rho = \|X\|_F/\sqrt{r}$; μ_0 at initialization.
h	Factor imbalance	Positive scalar that multiplies B by h and A by h^{-1} relative to a chosen normalized factor shape, at fixed X.
g_{init}	Fixed-basis size gain	Positive experiment-defined gain targeting $\mathrm{Var}(B_0A_0) = g_{\mathrm{init}}\,\mathrm{Var}(W_{\mathrm{pre}})$ per layer before dtype rounding; distinct from the cross term Q.
g_{ref}	Reference gain	10^{-3} for the relative-amplitude axis of Figure 2; the ratio $(g_{\mathrm{init}}/g_{\mathrm{ref}})^{1/4}$ is invariant to a common rescaling of both gains.
λ_B, λ_A	Normalized learning rates	$\lambda_B = \eta_B s^{q_{\mathrm{g}}} \rho_0^{(q_{\mathrm{g}}-1)/2} h_0^{-(q_{\mathrm{g}}+1)}$ and $\lambda_A = \eta_A s^{q_{\mathrm{g}}} \rho_0^{(q_{\mathrm{g}}-1)/2} h_0^{q_{\mathrm{g}}+1}$, under Theorem 2.
η_{step}	Step-matched learning rate	Matches $\eta\|F\|_F$ across amplitudes at fixed s and gradient field; $\propto a^{-(1+p)}$.
$\eta_{\mathrm{start},b}$	Transfer search start	$\eta_a(a/b)^{1+p}$ in a common-amplitude family; a candidate matching the first-order product response.
η_{rot}	Rotation-matched learning rate	Matches the first-order change of P_U, P_V at fixed s and gradient field; $\propto a^{1-p}$.
η^*	Tuned learning rate	A loss-minimizing learning rate in the specified search interval and at the stated horizon.
ζ	Normalized learning rate	$\zeta = a^{1+p}\eta \in I = [0, Z]$ for a fixed optimizer, s, and horizon; the limit minimizer is optimizer-dependent.

Table 9: Directional quantities and conventions.

Symbol	Meaning	Definition or domain
Direction and curvature of one step		
C_H	Pooled capture ratio	$\sum_{b,\ell}\|P_{T_\ell}H_{\ell b}\|_F^2 / \sum_{b,\ell}\|H_{\ell b}\|_F^2$ over stated batches and adapters.
c_i	Directional efficiency	$\cos(H, F_i) = \langle H, F_i\rangle_F/(\|H\|_F\|F_i\|_F)$ for nonzero H, F_i.
$k_i(\eta)$	Path effective curvature	$\phi_i''(\eta)/\|F_i\|_F^2$ along the exact one-step path $\phi_i(\eta) = \mathcal{L}(W + \eta F_i + \eta^2 Q_i)$; includes the factor cross term.
$\underline{k}_i, \overline{k}_i$	Curvature bounds	$0 < \underline{k}_i \le k_i(\eta) \le \overline{k}_i$ for $\eta \in [0, \eta_i^\dagger]$.
$\eta_i^\dagger$	One-step optimal learning rate	Interior minimizer of ϕ_i; satisfies Lemma 3 and is distinct from the full-training optimum η^*.
η_{scale}	Size-matched learning rate	$\eta_0\|F_0\|_F/\|F_1\|_F$ for a source learning rate η_0; using $\eta_0^\dagger$ gives the comparator in Lemma 3.
Operations and conventions		
$\langle Y_1, Y_2\rangle_F$	Frobenius inner product	$\operatorname{tr}(Y_1^\top Y_2)$ for equal-sized matrices.
$\cos(Y_1, Y_2)$	Matrix cosine	$\langle Y_1, Y_2\rangle_F/(\|Y_1\|_F\|Y_2\|_F)$ for nonzero arguments.
$\|\cdot\|_F, \|\cdot\|_2$	Frobenius/spectral norms	The spectral norm is the largest singular value.
$\operatorname{rms}(Y)$	Entry root-mean-square	$\|Y\|_F$ divided by the square root of the number of entries.
Γ	Gauge transformation	Invertible $r \times r$ matrix acting as $(B, A) \mapsto (B\Gamma, \Gamma^{-1}A)$; leaves X unchanged.
$\odot, \oslash$	Elementwise operations	Multiplication and division, respectively.