SPECTRAL OPTIMIZATION IN RESIDUAL NETWORKS: LEARNING ANISOTROPIC CORRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral optimizers such as Muon are often studied through deep linear models, where repeated matrix products create mode-dependent learning rates that become more uneven with depth. Modern pre-norm residual networks operate in a different regime. Their blocks can remain close to the identity map, which makes the classical product-induced factor nearly constant across modes. Meanwhile, the identity path can hide substantial anisotropy in the trainable residual correction. We argue that residual connections therefore shift, rather than remove, the relevant conditioning problem: the identity path stabilizes the full map, while the optimizer must learn a residual correction that has anisotropic structure. Under anisotropic residual-branch geometry in deep linear networks, gradient descent learns correction modes unevenly, while spectral updates balance progress across unlearned directions. Our results suggest that Muon's benefit in deep residual architectures arises because residual parameterization exposes a local spectral correction problem well matched to spectral optimization.

1 INTRODUCTION

Residual connections are a central component of modern deep architectures. They stabilize signal propagation and make very deep networks trainable. In modern large Transformer-based models, residual parameterization and pre-layer normalization have become standard design choices (Xiong et al., 2020; Bachlechner et al., 2021; Touvron et al., 2021; Wang et al., 2024). Recent research also finds that matrix-valued spectral optimizers such as Muon (Jordan et al., 2024) can outperform vector-valued optimization methods when singular directions learn at uneven rates (Vasudeva et al., 2026; Wang et al., 2026a).

A natural explanation for the benefits of spectral optimizers comes from plain deep linear networks. In the standard aligned setting with balanced initialization, an end-to-end target mode $t_i^\star$ is represented multiplicatively across L layers, so that the corresponding layerwise target is $(t_i^\star)^{1/L}$ (Vasudeva et al., 2026). As depth increases, these layerwise target modes become more similar. SPECGD is well matched to this structure because it updates active modes at the same absolute rate, while GD retains the products of the remaining layers in its gradients and therefore gives different modes different learning rates (Saxe et al., 2013). Residual parameterization changes this depth effect. For a near-identity target $t_i^\star = 1 + \delta_i$, the target of each layer's trainable residual mode becomes δ_i/L approximately. Depth therefore mainly divides the correction across layers rather than compressing its relative mode imbalance through an L-th root. This weakens the classical depth-based explanation for the advantage of spectral optimization in residual architectures.

This leads to our primary motivation: we argue that the relevant conditioning problem is not removed by residual connections, but rather shifted to and hidden within the trainable residual branch. Deep pre-norm blocks can have Jacobians close to the identity, especially in deeper layers (Sun et al., 2025). A small residual Jacobian guarantees that the full block is well-conditioned (i.e, isometry), but the optimizer updates parameters in the residual branch which can be highly anisotropic. We observe that the presence of the skip connection naturally promotes a highly anisotropic spectrum in this trainable branch. The identity path absorbs and transmits isotropic components of the representation, the residual branch is left to learn the anisotropic corrections.

Empirical results in LLMs confirm that this phenomenon occurs in practice. Even when the full block remains close to an isometry, the raw gradients in the residual branch remain highly spectrally

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.

anisotropic (Huang et al., 2026). Under this anisotropic residual-branch geometry, standard gradient descent learns correction modes unevenly, heavily favoring directions with larger signal magnitudes. Spectral optimizers, on the other hand, perfectly align with this ill-conditioned optimization problem by balancing progress across unlearned directions.

Our analysis utilizes gradient descent and an idealized spectral update (i.e., SPECGD) to isolate the geometric and spectral allocation mechanisms at play. We formalize the perspective that Muon's benefit in deep residual architectures arises because residual parameterization exposes a highly anisotropic spectral correction problem that is fundamentally well matched to spectral optimization.

In summary, our main **contributions** are as follows:

- Through an exact analysis of an aligned deep residual linear network under isotropic inputs, we show that residual parameterization changes the depth allocation of near-identity modes from multiplicative factorization to approximately additive factorization, which weakens the classical depth-based advantage of spectral updates.
- We derive an exact conditional block decomposition that identifies the local geometry of the residual branch. We provide a matched velocity condition proving exactly when spectral equalization (i.e., SPECGD) reduces correction risk faster than gradient descent.
- Through controlled skip-strength ablations and LLM experiments, we provide evidence that skip connections promote an anisotropic residual-branch Jacobian while the full-block Jacobian remains near isotropic.

2 RELATED WORK

Muon and Spectral Optimization. Muon applies approximate orthogonalization to matrix-valued momentum updates and has demonstrated strong empirical performance (Jordan et al., 2024; Liu et al., 2025; Shah et al., 2025). One line of research explains such orthogonal updates through non-Euclidean optimization geometry. Bernstein & Newhouse (2024) develop a norm-dependent steepest-descent view, under which spectral normalization arises naturally for matrix parameters; while Chen et al. (2026) further connect Muon to implicit spectral-norm constraints. From a complementary perspective, Pethick et al. (2025) formulate optimization through norm-constrained linear minimization oracles (LMOs) and introduce Scion. Muon has also been interpreted as structured matrix preconditioning. Lau et al. (2025) develop a unified preconditioning perspective and derive PolarGrad, while Ma et al. (2026) characterize how spectral orthogonalization mitigates ill-conditioning in structured problems. Another closely related line studies how spectral updates redistribute learning across modes. Vasudeva et al. (2026) show that SpecGD equalizes learning rates across principal components on imbalanced data, with depth amplifying this effect in standard deep linear networks. Wang et al. (2026a) obtain analogous balancing across frequency components in associative-memory learning and derive corresponding scaling laws.

Learning Dynamics in Deep (Linear) Networks. Saxe et al. (2013) derive exact mode-wise dynamics for aligned deep linear networks. Arora et al. (2019a) establish convergence results, while Du et al. (2018) show that gradient flow preserves layer balance. This balance is not only critical for trainability (Bartlett et al., 2018) but also implicit regularization and its effect on model selection (Marcotte et al., 2024; Wang & Klabjan, 2025). Arora et al. (2019b) and Jacobs et al. (2025) study such implicit regularization in deep matrix factorization and Gidel et al. (2019) analyze discrete gradient dynamics in linear networks. Shamir (2019) and Jacobs et al. (2026) show that deep linear parameterization can also produce very slow optimization. We use the same aligned-mode methodology, but center the analysis at identity factors. This changes the size of the product-induced prefactor without removing the exact deep coupling.

Residual Networks and Near-identity Blocks. Residual connections (He et al., 2016) change optimization and signal propagation in deep networks. Identity initialization can simplify optimization (Hardt & Ma, 2017; Bartlett et al., 2018), while residual structure can mitigate shattered gradients and loss-surface singularities (Balduzzi et al., 2017; Orhan & Pitkow, 2018). Dynamical isometry can arise under suitable residual scaling or initialization (Tarnowski et al., 2019; Gadhikar & Burkholz, 2022). Continuous-depth limits provide another view of deep residual models (Chen et al., 2018;

Marion et al., 2024). Pre-LN Transformers, ReZero, LayerScale, and DeepNet explicitly control the size of residual updates (Xiong et al., 2020; Bachlechner et al., 2021; Touvron et al., 2021; Wang et al., 2024). Most directly related to our motivation, Sun et al. (2025) report that the derivatives of later Pre-LN Transformer blocks can approach the identity. This observation motivates separating the full block Jacobian from the residual-branch Jacobian.

3 PRELIMINARIES

We consider teacher-student regression with input $x \in \mathbb{R}^d$, covariance $\Sigma_x := \mathbb{E}[xx^\top]$, and target

$$y = T^\star x + \xi, \qquad T^\star \in \mathbb{R}^{d\times d}, \tag{1}$$

where ξ denotes optional zero-mean noise. We assume that the target transformation $T^\star$ is close to the identity matrix, which suggests that $\|T^\star - I\|_2$ is small enough. This assumption is consistent with the empirical phenomena that the Jacobian matrices of deep layers in LLM are close to identity matrix (Sun et al., 2025). For the theoretical statements below, the noise can be omitted without changing the deterministic mode dynamics.

The linear residual block of ℓ-th layer acts as

$$h_{\ell+1} = (I + W_\ell)h_\ell, \tag{2}$$

where h_ℓ is the identity path or skip path, and $W_\ell h_\ell$ represents the ℓ-th layer's residual branch. The global student across all the layers is

$$f_\Theta(x) = F_\Theta x, \qquad F_\Theta := \prod_{\ell=1}^{L}(I + W_\ell), \tag{3}$$

where the product is ordered from the input side to the output side. For a matrix A, $\|A\|_2$ denotes the spectral norm and $\|A\|_F$ denotes the Frobenius norm. If $G = U\Sigma V^\top$ is a compact singular value decomposition, we define the matrix sign as

$$\operatorname{msgn}(G) := UV^\top, \tag{4}$$

with zero singular directions omitted. Standard gradient descent uses

$$W_{t+1} = W_t - \eta G_t, \qquad G_t = \nabla_W \mathcal{L}(W_t). \tag{5}$$

It is the steepest step associated with a quadratic Frobenius penalty (Bernstein & Newhouse, 2024)

$$-\eta G_t = \arg\min_{\Delta} \left\{ \langle G_t, \Delta\rangle + \frac{1}{2\eta}\|\Delta\|_F^2 \right\}. \tag{6}$$

The steepest descent on a unit ball of Frobenius norm would instead be the normalized gradient $-G_t/\|G_t\|_F$ (Boyd & Vandenberghe, 2004). This distinction matters when comparing normalized steepest directions.

The idealized spectral update used in this paper, which we call SPECGD (Carlson et al., 2015), is defined as

$$W_{t+1} = W_t - \eta \operatorname{msgn}(G_t). \tag{7}$$

It solves the steepest descent problem on a unit ball of spectral norm (Bernstein & Newhouse, 2024)

$$-\operatorname{msgn}(G_t) \in \arg\min_{\|\Delta\|_2\le 1} \langle G_t, \Delta\rangle. \tag{8}$$

The above discussion is based on a first-order expansion of the loss function, under which Muon can be interpreted as taking a steepest-descent direction on a unit-norm ball. When second-order Hessian and higher order terms are taken into account, however, theoretical analysis suggests that the optimal update need not correspond to exact orthogonalization; instead, it favors a more homogeneous singular value spectrum (Su, 2025). Consistent with this perspective, several recent works have shown that explicitly reshaping the singular values of the momentum or weight matrices, rather than merely whitening them, can further improve the performance of Muon-style optimizers (Pang et al., 2026; Wang et al., 2026b; Shumaylov et al., 2026; Dong & Sawin, 2026).

The exact Frobenius-normalized analogue of SPECGD is normalized GD instead of standard GD. We retain standard GD as the main Euclidean baseline because its mode dynamics are the conventional deep linear reference and Frobenius normalization multiplies all active modes by the same scalar, leaving their relative allocation unchanged. Section 5 compares GD and SPECGD under matched initial Frobenius velocity when an absolute performance comparison is needed.

Practical Muon orthogonals the momentum matrix. Ignoring details of the Nesterov variant, a simplified update has the form

$$M_t = \mu M_{t-1} + (1-\mu)G_t, \qquad W_{t+1} = W_t - \eta\,\widetilde{\mathrm{msgn}}(M_t),$$

where $\widetilde{\mathrm{msgn}}$ is the approximation of exact matrix sign operation by Newton–Schulz iterations. Our SPECGD analysis sets momentum to zero and uses the exact matrix sign operation. In a fixed aligned basis, momentum remains diagonal mode by mode. If active gradient signs remain unchanged over the averaging window, the matrix sign of the momentum has the same active signs as the current gradient. Momentum changes temporal smoothing in this regime without changing the spectral allocation mechanism studied below. For reference, Li et al. (2026) discusses the role of momentum in Muon through a perspective of spectral filtering in details.

4 FROM MULTIPLICATIVE TO ADDITIVE MODE ALLOCATION

We study the depth effect in an aligned model, where the target and all layer matrices share a fixed orthogonal basis. This reduces the matrix dynamics to scalar mode dynamics and allows us to compare the plain and residual parameterizations in the same coordinates.

Assumption 1 (Aligned isotropic model)**.** *The inputs are isotropic, $\Sigma_x = I$, and there is an orthogonal matrix Q such that*

$$T^\star = QD^\star Q^\top, \qquad D^\star = \mathrm{diag}(t_1^\star, \dots, t_d^\star). \tag{9}$$

For the plain network, the layer matrices are initialized as

$$V_\ell(0) = QB_\ell(0)Q^\top, \qquad B_\ell(0) = \mathrm{diag}(q_{\ell 1}(0), \dots, q_{\ell d}(0)). \tag{10}$$

For the residual network,

$$W_\ell(0) = QA_\ell(0)Q^\top, \qquad A_\ell(0) = \mathrm{diag}(a_{\ell 1}(0), \dots, a_{\ell d}(0)). \tag{11}$$

Under this alignment, the gradient remains diagonal in the same basis Q for both GD and SPECGD. The basis therefore does not rotate during training, and the evolution is entirely described by the diagonal mode amplitudes. We use balanced initialization in the standard layerwise sense: for each mode i, its scalar factor is initially equal across all layers. Thus $q_{\ell i}(0) = q_i(0)$ for the plain network and $a_{\ell i}(0) = a_i(0)$ for the residual network. Both optimization flows preserve this structure in the aligned model.

4.1 MULTIPLICATIVE MODE ALLOCATION IN A PLAIN DEEP LINEAR NETWORK

Consider the plain deep linear networks $F_{\mathrm{plain}} = \prod_{\ell=1}^{L} V_\ell$. On the aligned balanced structure, the i-th mode has the form $s_i = q_i^L$. To represent a target mode $t_i^\star > 0$, each layer therefore approaches $q_i^\star = (t_i^\star)^{1/L}$. For two target modes,

$$q_i^\star / q_j^\star = \left(t_i^\star / t_j^\star\right)^{1/L}. \tag{12}$$

Thus depth compresses the relative difference between the layerwise targets even though the end-to-end target spectrum is fixed. The two optimizers respond differently to this factorization. For one layerwise mode q_i,

$$\dot{q}_i^{\mathrm{GD}} = \eta_G (t_i^\star - s_i) q_i^{L-1}, \tag{13}$$

$$\dot{q}_i^{\mathrm{Spec}} = \eta_S\, \mathrm{sign}(t_i^\star - s_i). \tag{14}$$

The factor q_i^{L-1} in GD is the product of the corresponding modes in the other $L-1$ layers. SPECGD removes this magnitude dependence and gives every active mode the same absolute layerwise velocity.

If all modes start from a common value α below their targets, SPECGD reaches the i-th balanced factor after

$$T_i^{\mathrm{Spec}} = \frac{(t_i^\star)^{1/L} - \alpha}{\eta_S}. \tag{15}$$

which can be easily derived from Eq. (14) since the velocity is constant. The difference between these completion times therefore decreases with depth because the layerwise targets are compressed by the L-th root. In other words, if the original target $t_i^\star > 1$, the new learning target $(t_i^\star)^{1/L}$ will be reduced; conversely, if the target $t_i^\star < 1$, it will be amplified. GD does not have the same mode equalization because its velocity still contains q_i^{L-1}. Such kind of equalization of the mode-learning speed has been shown to accelerate the learning the weak components and decrease the balanced risk in the early training (Vasudeva et al., 2026).

4.2 ADDITIVE ALLOCATION OF A NEAR-IDENTITY CORRECTION

Residual parameterization changes what each trainable layer must represent. The i-th end-to-end mode is now

$$s_i = (1 + a_i)^L. \tag{16}$$

Consider a target close to identity,

$$t_i^\star = 1 + \delta_i, \qquad |\delta_i| \ll 1. \tag{17}$$

The corresponding balanced residual mode is

$$a_i^\star = (1 + \delta_i)^{1/L} - 1 = \frac{\delta_i}{L} - \frac{L-1}{2L^2}\delta_i^2 + O(\delta_i^3). \tag{18}$$

To first order,

$$\delta_i \simeq L a_i^\star. \tag{19}$$

Thus the target correction δ_i is allocated approximately additively across the L layers.

This allocation differs from the plain product model. For two nonzero correction modes,

$$\frac{a_i^\star}{a_j^\star} = \frac{\delta_i}{\delta_j}\left(1 + O(\rho)\right), \qquad |\delta_i|, |\delta_j| \le \rho. \tag{20}$$

Depth reduces the absolute scale of both residual modes by approximately $1/L$, but their relative difference is unchanged at leading order. The L-th-root compression that benefits SPECGD in the plain DLN is therefore absent from the trainable residual correction.

4.3 EXACT RESIDUAL MODE DYNAMICS NEAR IDENTITY

The additive allocation above provides an intuition on the benefits attenuation of the depth near-identity regime. The exact optimization dynamics of residual network is provided in the following.

Proposition 2 (Exact aligned residual mode dynamics)**.** *Under Assumption 1, suppose the residual initialization is balanced across layers,*

$$a_{\ell i}(0) = a_i(0).$$

Continuous-time GD and SPECGD *preserve the aligned balanced structure. The modes satisfy*

$$\dot{r}_i^{\mathrm{GD}} = \eta_G L (1 + r_i)^{2-2/L}(\delta_i - r_i), \tag{21}$$

$$\dot{r}_i^{\mathrm{Spec}} = \eta_S L (1 + r_i)^{1-1/L}\,\mathrm{sign}(\delta_i - r_i). \tag{22}$$

until the corresponding gradient mode reaches zero.

In the near-identity regime, these dynamics lead to a different mode-learning picture from the plain deep linear model. For GD, let $T_{i,\epsilon}^{\mathrm{GD}}$ denote the time required for the i-th mode to be ϵ close to the target. Then

$$T_{i,\epsilon}^{\mathrm{GD}} = \frac{\log(1/\epsilon)}{\eta_G L}\left(1 + O(\rho)\right). \tag{23}$$

In particular, for two active modes i and j, $T^{\mathrm{GD}}_{i,\epsilon}/T^{\mathrm{GD}}_{j,\epsilon} = 1 + O(\rho)$. Thus different residual modes make approximately the same fractional progress under GD simultaneously, even when their target corrections have very different magnitudes.

For SPECGD, starting from zero residual initialization, the exact saturation time of the i-th mode is

$$T^{\mathrm{Spec}}_i = \frac{\left|(1+\delta_i)^{1/L} - 1\right|}{\eta_S} = \frac{|\delta_i|}{\eta_S L}\left(1 + O(\rho)\right), \tag{24}$$

under the condition that $|\delta_i| \le \rho \ll 1$. Therefore, for two nonzero correction modes i and j, $T^{\mathrm{Spec}}_i/T^{\mathrm{Spec}}_j = |\delta_i|/|\delta_j|\,(1 + O(\rho))$.

These time scales contrast directly with the plain deep linear model. There, SPECGD moves all layerwise modes from a common initialization toward $(t^\star_i)^{1/L}$ with the same velocity, so increasing depth compresses the differences between their saturation times. In the residual model, the identity path already supplies the common unit component. The trainable part only needs to learn the residual correction $(1+\delta_i)^{1/L} - 1$, whose relative magnitude differences are preserved to first order. Equal absolute velocity under SPECGD therefore no longer produces balanced completion times across modes, while the behavior of GD changes in the opposite direction.

This reverses the mode-balance picture used in the plain deep linear explanation. The balanced-learning mechanism studied by Vasudeva et al. (2026), where reducing the learning disparity between strong and weak spectral components benefits weak-mode learning and balanced risks, no longer favors SPECGD in the near-identity residual regime. The classical depth-based explanation is therefore insufficient to account for the benefit of spectral optimization in residual networks. Detailed derivations of (23) and (24) are given in Appendix A.3.

5 RESIDUAL ANISOTROPY AND LOCAL CORRECTION GEOMETRY

Section 4 isolates the effect of residual parameterization under aligned modes and isotropic inputs. In that setting, the residual correction can be highly anisotropic in magnitude, while the geometry multiplying the GD error remains independent of the mode magnitude. GD therefore makes approximately equal progress across correction modes.

A general residual block contains an additional source of mode dependence. The gradient of its trainable branch depends not only on the correction that remains to be learned, but also on the geometry of the signals in and out the branch. This section separates these two effects. We first describe how an anisotropic residual transformation can be hidden by the identity path, and then derive the conditional optimization geometry that determines the mode-wise learning rates.

5.1 RESIDUAL-BRANCH ANISOTROPY AND IDENTITY MASKING

Consider the Jacobian of a residual block

$$J_{\mathrm{block},\ell} = I + J_{\mathrm{res},\ell}, \tag{25}$$

where $J_{\mathrm{res},\ell}$ is the Jacobian of the trainable residual branch. In the near-identity regime

$$\|J_{\mathrm{res},\ell}\|_2 \le \rho < 1, \tag{26}$$

the singular values of the full block satisfy

$$1 - \rho \le \sigma_i(J_{\mathrm{block},\ell}) \le 1 + \rho. \tag{27}$$

Hence

$$\kappa(J_{\mathrm{block},\ell}) \le \frac{1+\rho}{1-\rho}. \tag{28}$$

These bounds constrain the absolute fluctuation of the full block, but not the relative variation among the singular values of $J_{\mathrm{res},\ell}$. For example, residual singular values of sizes ρ and $\varepsilon\rho$ can differ by a factor of $1/\varepsilon$, while the corresponding full-block singular values remain within an $O(\rho)$ neighborhood of one. The identity path can therefore mask strong relative anisotropy in the residual transformation. We refer to this separation between the residual and full-block spectra as *identity masking*.

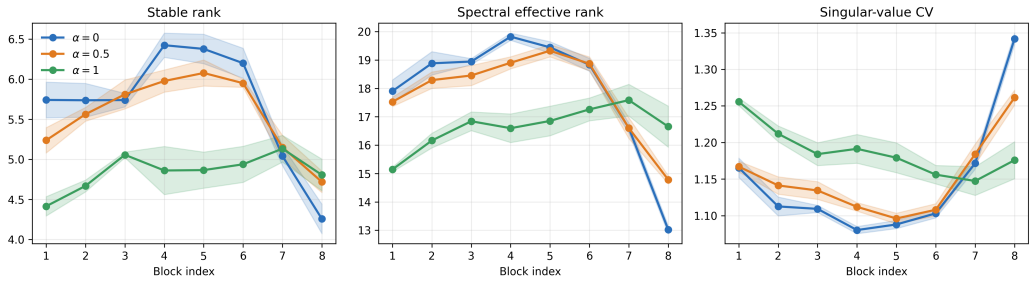


Figure 1: Residual-branch Jacobian anisotropy in the eight-layer pre-norm MLP. A unit skip produces lower stable rank, lower spectral effective rank, and larger singular-value variation across most layers.

The same separation appears in the aligned model of Section 4. A target mode $1 + \delta_i$ can remain close to one for every i even when the residual amplitudes $|\delta_i|$ differ greatly across modes. Thus anisotropy of the residual correction does not by itself imply anisotropic learning under GD.

Residual parameterization also changes what the trainable branch needs to represent. The identity path already supplies the common isotropic part of the block transformation, while the residual branch represents deviations from that map. These deviations can vary strongly across singular modes even when their absolute magnitudes remain small. A residual branch can therefore become substantially more anisotropic without making the complete block poorly conditioned.

We test this effect by varying the skip strength while keeping the branch architecture and training procedure fixed. Figure 1 reports three complementary measures of the residual-Jacobian spectrum. With a unit skip, the residual branch has lower stable rank, lower spectral effective rank, and larger singular value variation across most layers than the zero-skip and half-skip models. At the same time, the corresponding full-block Jacobians become more isotropic, as shown in Appendix C.4. The experiment therefore shows the two sides of identity masking in the same trained network: the residual branch becomes more anisotropic while the complete block becomes better conditioned.

5.2 Exact conditional geometry

Residual-branch anisotropy describes the target learned by the branch, but it does not determine the optimization dynamics on it own. The latter also depends on signals entering and leaving the branch. In other words, the residual correction determines how much must be learned in each mode, while the surrounding signal geometry determines how strongly optimizers weights the error between the target and the learned mode.

This distinction connects the present analysis to Section 4. In the aligned isotropic model, the input and output geometry is generated only by the remaining near-identity factors. For the i-th mode of a balanced residual network, the resulting geometry coefficient is $c_i = (1 + a_i)^{2(L-1)} = s_i^{2-2/L}$. Near identity, $a_i = O(\delta_i/L)$, so $c_i \approx 1$ even when the residual amplitudes δ_i have large relative differences.

Consider $h_{\ell+1} = h_\ell + W_\ell h_\ell$. Let

$$A_{<\ell} := \prod_{j<\ell}(I + W_j), \qquad A_{>\ell} := \prod_{j>\ell}(I + W_j),$$

so $h_\ell = A_{<\ell}x$. Condition on all blocks except W_ℓ and define the output residual with the ℓ-th branch removed as $r_\ell := y - A_{>\ell}h_\ell$. Conditioned on the current values of all other blocks, the loss as a function of W_ℓ is exactly

$$\mathcal{L}_\ell(W) = \frac{1}{2}\mathbb{E}\,\|A_{>\ell}Wh_\ell - r_\ell\|_2^2\,. \tag{29}$$

Define

$$H_{\text{out},\ell} := A_{>\ell}^\top A_{>\ell}, \qquad H_{\text{in},\ell} := \mathbb{E}[h_\ell h_\ell^\top], \qquad C_{rh,\ell} := \mathbb{E}[r_\ell h_\ell^\top],$$

direct differentiation gives

$$\nabla_W \mathcal{L}_\ell = H_{\text{out},\ell}WH_{\text{in},\ell} - A_{>\ell}^\top C_{rh,\ell}. \tag{30}$$

If $H_{\mathrm{out},\ell}$ and $H_{\mathrm{in},\ell}$ are invertible, define

$$\Delta_\ell^\star := H_{\mathrm{out},\ell}^{-1} A_{>\ell}^\top C_{rh,\ell} H_{\mathrm{in},\ell}^{-1}. \tag{31}$$

The gradient becomes

$$\nabla_W \mathcal{L}_\ell = H_{\mathrm{out},\ell}(W - \Delta_\ell^\star) H_{\mathrm{in},\ell}. \tag{32}$$

Up to a term independent of W, the same conditional loss is

$$\mathcal{L}_\ell(W) = \frac{1}{2} \operatorname{tr}\left[H_{\mathrm{out},\ell}(W - \Delta_\ell^\star) H_{\mathrm{in},\ell}(W - \Delta_\ell^\star)^\top\right] + \text{const}. \tag{33}$$

For the mode analysis below, we omit the layer index.

Assumption 3 (Aligned conditional geometry)**.** *There exist orthogonal matrices U, V such that*

$$H_{\mathrm{out}} = U\Gamma U^\top, \qquad \Delta^\star = UDV^\top, \qquad H_{\mathrm{in}} = V\Lambda V^\top, \tag{34}$$

where

$$\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_r), \qquad D = \operatorname{diag}(d_1, \ldots, d_r), \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r),$$

with $\gamma_i > 0$, $\lambda_i > 0$, and $d_i > 0$ on the active modes.

Set $c_i := \gamma_i \lambda_i$ on the active modes. Here d_i measures the correction amplitude of mode i, while c_i measures its local optimization geometry. These two sources of anisotropy play different roles in the learning dynamics.

Theorem 4 (Mode dynamics in an aligned residual branch)**.** *Under Assumption 3 with $W(0) = 0$, continuous-time GD and* SPECGD *preserve the aligned diagonal manifold. The i-th mode satisfies*

$$w_i^{\mathrm{GD}}(t) = d_i\left(1 - e^{-\eta_G c_i t}\right), \tag{35}$$

$$w_i^{\mathrm{Spec}}(t) = \min(\eta_S t, d_i). \tag{36}$$

For a target fraction $q \in (0, 1)$,

$$t_i^{\mathrm{GD}}(q) = \frac{-\log(1-q)}{\eta_G c_i}, \qquad t_i^{\mathrm{Spec}}(q) = \frac{q\, d_i}{\eta_S}. \tag{37}$$

Thus GD has mode-dependent fractional learning times when the local geometry c_i is anisotropic, whereas SPECGD removes this geometry factor and gives every unfinished mode the same absolute velocity. Its fractional completion times can still differ because the target amplitudes d_i can differ.

This also explains the relation to Section 4. In the aligned isotropic residual model, $c_i \approx 1$ near identity, so the geometry is nearly mode independent and GD has approximately common fractional convergence times. The present conditional model allows H_{in} and H_{out} to introduce additional anisotropy, making the c_i substantially different across modes. The resulting GD imbalance therefore comes from the surrounding optimization geometry rather than from the residual correction amplitudes alone.

Equal absolute mode speed does not by itself imply better performance, such as a lower risk. We therefore compare the unweighted risk

$$\mathcal{R}(W) := \frac{1}{2}\|W - \Delta^\star\|_F^2 = \frac{1}{2}\sum_{i=1}^{r}(d_i - w_i)^2 \tag{38}$$

after matching the initial Frobenius speed of the two flows.

Theorem 5 (Initial risk comparison at matched Frobenius speed)**.** *Under Theorem 4, set*

$$\eta_G = \frac{1}{\|c \odot d\|_2}, \qquad \eta_S = \frac{1}{\sqrt{r}}. \tag{39}$$

Then

$$\left\|\dot{W}_{\mathrm{GD}}(0)\right\|_F = \left\|\dot{W}_{\mathrm{Spec}}(0)\right\|_F = 1. \tag{40}$$

At $t = 0$, SPECGD *decreases $\mathcal{R}$ faster than GD if and only if*

$$\frac{1}{\sqrt{r}}\sum_i d_i > \frac{\sum_i c_i d_i^2}{\|c \odot d\|_2}. \tag{41}$$

If the inequality is strict, then $\mathcal{R}_{\mathrm{Spec}}(t) < \mathcal{R}_{\mathrm{GD}}(t)$ on a nonzero early-time interval.

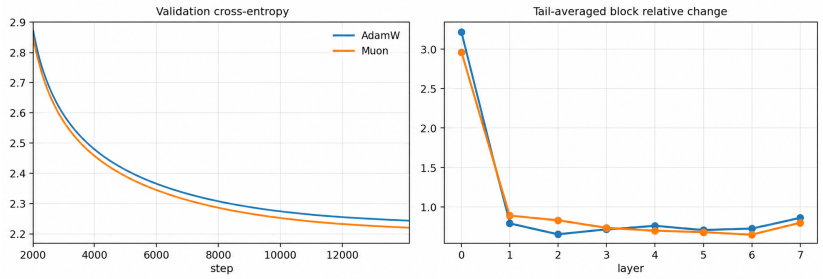


Figure 2: Decoder pretraining diagnostics. Left: validation cross-entropy under AdamW and Muon. Right: tail-averaged relative block change, which is small in the deeper blocks for both optimizers.

Under the matched learning rates, the initial mode velocities are

$$\dot{w}^{\mathrm{GD}}(0) = \frac{c \odot d}{\|c \odot d\|_2}, \qquad \dot{w}^{\mathrm{Spec}}(0) = \frac{1}{\sqrt{r}}\mathbf{1}. \tag{42}$$

Since $-\nabla_w \mathcal{R}(0) = d$, the initial decrease of $\mathcal{R}$ is the projection of each unit-speed velocity onto the correction vector d. GD allocates more of its update speed to modes with larger $c_i d_i$, whereas SpecGD allocates the same absolute speed to every active mode. Condition (41) states exactly when the latter allocation makes faster initial progress toward the residual correction.

The condition can hold even when the correction itself is strongly anisotropic. For example, $d = (1, \varepsilon)$ and $c = (\varepsilon^2, 1)$ place the larger correction in the slower GD geometry, and SpecGD has a strict initial advantage for $0 < \varepsilon < 1$; Appendix A.6 gives the calculation. Appendix B gives an approximate-alignment bound and a factorized-branch extension.

6 Experiments

The skip-strength experiment in Figure 1 tests how the identity path changes residual-branch and full-block Jacobians. Its experimental details and additional block-level spectra are given in Appendix C.4. Here we examine whether the same near-identity and anisotropic regimes appear during decoder pretraining.

We train an eight-layer pre-norm decoder for one epoch with AdamW or Muon. Figure 2 shows that Muon reaches lower validation cross entropy, while the relative block change $\|h_{\mathrm{out}} - h_{\mathrm{in}}\| / \|h_{\mathrm{in}}\|$ is small in the deeper blocks for both runs, consistent with a near-identity forward regime. In an additional experiment with a 130M parameter model on the C4 dataset (Raffel et al., 2020), the combined effective rank averaged over representative layers is 2.48% under AdamW and 4.00% under Muon. The low effective-rank fractions show that strongly anisotropic raw-gradient geometry persists under both optimizers. Experimental settings and metric definitions are given in Appendix C.

7 Conclusion and Limitations

Residual parameterization changes the mode-learning picture of spectral optimization. Near identity suppresses the product-induced rate differences of plain deep linear networks, so anisotropic correction amplitudes alone need not produce uneven fractional learning speed under GD. In a general residual block, anisotropic input and output geometry introduces additional mode-dependent factors; GD inherits these factors, whereas SpecGD removes their magnitudes from the active update and can reduce unweighted correction risk faster at matched initial speed. Our Jacobian and decoder diagnostics support the coexistence of near-isometric full blocks, anisotropic residual transformations, and anisotropic optimizer-facing gradients. The theoretical results use aligned linear models and SpecGD, while practical Muon additionally includes momentum, approximate orthogonalization, and stochastic gradients.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the manuscript structuring and editing, experimental planning and code generation, and LaTeX formatting. The authors reviewed the AI-assisted material and take responsibility for the final content of this work, including its text, claims, proofs, and artifacts.

REPRODUCIBILITY STATEMENT

The main assumptions and closed-form dynamics are stated in Sections 4 and 5, with complete proofs and extensions in Appendices A and B. Appendix C reports architectures, data splits, hyperparameters, seed aggregation, Jacobian validation, and compute. The supplementary material includes the experiment drivers, environment requirements, fixed sample indices, per-seed summaries, checkpoints, and plotting utilities used to produce the reported figures.

REFERENCES

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019b.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International conference on machine learning*, pp. 342–350. PMLR, 2017.

Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pp. 521–530. PMLR, 2018.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

Stephen Boyd and Lieven Vandenberghe. Convex optimization, 2004.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Artificial intelligence and statistics*, pp. 111–119. PMLR, 2015.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yihe Dong and Will Sawin. Muon^p: Muon with fractional spectral powers. In *Third Conference on Language Modeling*, 2026.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

Advait Gadhikar and Rebekka Burkholz. Dynamical isometry for residual networks. *arXiv preprint arXiv:2210.02411*, 2022.

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

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhendong Huang, Hengjie Cao, Fang Dong, Ruijun Huang, Mengyi Chen, Yifeng Yang, Xin Zhang, Anrui Chen, Mingzhi Dong, Yujiang Wang, et al. Spectra: Rethinking optimizers for llms under spectral anisotropy. *arXiv preprint arXiv:2602.11185*, 2026.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? In *Forty-second International Conference on Machine Learning*, 2025.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Never saddle for reparameterized steepest descent as mirror flow. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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

Ren-Cang Li. New perturbation bounds for the unitary polar factor. *SIAM Journal on Matrix Analysis and Applications*, 16(1):327–332, 1995. doi: 10.1137/S0895479893256359.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in muon via spectral filtering. *arXiv preprint arXiv:2606.03899*, 2026.

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

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Keep the momentum: Conservation laws beyond euclidean gradient flows, 2024.

Pierre Marion, Yu-Han Wu, Michael Eli Sander, and Gérard Biau. Implicit regularization of deep residual networks towards neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024.

Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. In *International Conference on Learning Representations*, 2018.

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu, and Yaoqing Yang. HTMuon: Improving muon via heavy-tailed spectral correction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36504–36535. Association for Computational Linguistics, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Forty-second International Conference on Machine Learning*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pp. 2691–2713. PMLR, 2019.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

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

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrzębski, Jacek Tabor, and Maciej Nowak. Dynamical isometry is achieved in residual networks in a universal way for any activation function. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2221–2230. PMLR, 2019.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 32–42, 2021.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6761–6774, 2024.

Kaifei Wang, Binghui Li, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *Forty-third International Conference on Machine Learning*, 2026a.

Senmiao Wang, Tiantian Fang, Haoran Zhang, Yushun Zhang, Kunxiang Zhao, Alex Schwing, and Ruoyu Sun. Pc layer: Polynomial weight preconditioning for improving llm pre-training. *arXiv preprint arXiv:2606.06470*, 2026b.

Shuyang Wang and Diego Klabjan. A mirror descent perspective of smoothed sign descent. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

A Deferred proofs

A.1 Proof of Proposition 2

Under Assumption 1, all target and layer matrices are diagonal in the basis Q. For the plain network, write

$$V_\ell = QB_\ell Q^\top, \qquad B_\ell = \mathrm{diag}(q_{\ell 1}, \dots, q_{\ell d}).$$

The end-to-end map is

$$F_{\mathrm{plain}} = Q \left[\prod_{\ell=1}^{L} B_\ell \right] Q^\top,$$

and its i-th mode is

$$s_i = \prod_{\ell=1}^{L} q_{\ell i}.$$

Using $\Sigma_x = I$,

$$\mathcal{L} = \frac{1}{2} \sum_i (s_i - t_i^\star)^2.$$

Therefore

$$\frac{\partial \mathcal{L}}{\partial q_{\ell i}} = (s_i - t_i^\star) \prod_{j \neq \ell} q_{ji}. \tag{43}$$

The matrix gradient is diagonal in the same basis Q, so both GD and SpecGD preserve the aligned structure.

If $q_{\ell i}(0) = q_i(0)$ for all ℓ, the right-hand sides are equal across layers and the balanced structure is preserved. Hence

$$s_i = q_i^L,$$

and

$$\dot{q}_i^{\mathrm{GD}} = \eta_G (t_i^\star - s_i) q_i^{L-1}, \tag{44}$$

$$\dot{q}_i^{\mathrm{Spec}} = \eta_S \,\mathrm{sign}(t_i^\star - s_i). \tag{45}$$

For the residual network,

$$W_\ell = QA_\ell Q^\top, \qquad A_\ell = \mathrm{diag}(a_{\ell 1}, \dots, a_{\ell d}),$$

so

$$F_{\mathrm{res}} = Q \left[\prod_{\ell=1}^{L} (I + A_\ell) \right] Q^\top.$$

Its i-th mode is

$$s_i = \prod_{\ell=1}^{L} (1 + a_{\ell i}).$$

Differentiation gives

$$\frac{\partial \mathcal{L}}{\partial a_{\ell i}} = (s_i - t_i^\star) \prod_{j \neq \ell} (1 + a_{ji}). \tag{46}$$

The gradient again remains diagonal in the same basis Q. If $a_{\ell i}(0) = a_i(0)$ for all ℓ, this equality is preserved and

$$s_i = (1 + a_i)^L.$$

The layerwise dynamics are

$$\dot{a}_i^{\mathrm{GD}} = \eta_G (t_i^\star - s_i)(1 + a_i)^{L-1}, \tag{47}$$

$$\dot{a}_i^{\mathrm{Spec}} = \eta_S \,\mathrm{sign}(t_i^\star - s_i). \tag{48}$$

For either parameterization, differentiating the end-to-end mode gives

$$\dot{s}_i^{\mathrm{GD}} = \eta_G L s_i^{2-2/L} (t_i^\star - s_i),$$

$$\dot{s}_i^{\mathrm{Spec}} = \eta_S L s_i^{1-1/L} \,\mathrm{sign}(t_i^\star - s_i).$$

This proves Proposition 2 and the layerwise equations used in Section 4.

A.2 A COMMUTING NON-ISOTROPIC EXTENSION

The isotropic-input assumption removes any preferred input basis from the main result. If instead

$$\Sigma_x = Q\Lambda_x Q^\top, \qquad \Lambda_x = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \qquad \lambda_i > 0,$$

in the same basis as the target and initialization, the aligned manifold remains invariant. The GD layer dynamics and balanced end-to-end dynamics become

$$\dot{a}_{\ell i} = \eta_G \lambda_i (t_i^\star - s_i) \prod_{j \neq \ell} (1 + a_{ji}), \qquad \dot{s}_i = \eta_G L \lambda_i s_i^{2-2/L} (t_i^\star - s_i).$$

The positive factor λ_i does not change the active SPECGD signs. This extension recovers the earlier covariance-weighted equations, but it requires the commuting structure that the main theorem avoids.

A.3 MODE-LEARNING TIMES IN THE NEAR-IDENTITY RESIDUAL MODEL

We derive the mode-learning times used in Section 4. Recall the exact residual-mode dynamics

$$\dot{r}_i^{\mathrm{GD}} = \eta_G L (1 + r_i)^{2-2/L} (\delta_i - r_i), \tag{49}$$

$$\dot{r}_i^{\mathrm{Spec}} = \eta_S L (1 + r_i)^{1-1/L} \operatorname{sign}(\delta_i - r_i). \tag{50}$$

We assume zero residual initialization, so $r_i(0) = 0$, and consider trajectories satisfying $|r_i(t)| \leq \rho < 1$.

For GD, define

$$e_i(t) := \delta_i - r_i(t).$$

Equation (49) gives

$$\dot{e}_i(t) = -\eta_G L (1 + r_i(t))^{2-2/L} e_i(t). \tag{51}$$

Therefore

$$\frac{d}{dt} \log |e_i(t)| = -\eta_G L (1 + r_i(t))^{2-2/L}. \tag{52}$$

Let

$$T_{i,\epsilon}^{\mathrm{GD}} := \inf \left\{ t \geq 0 : |e_i(t)| \leq \epsilon |e_i(0)| \right\}, \qquad 0 < \epsilon < 1. \tag{53}$$

Integrating (52) from 0 to $T_{i,\epsilon}^{\mathrm{GD}}$ gives

$$\log(1/\epsilon) = \eta_G L \int_0^{T_{i,\epsilon}^{\mathrm{GD}}} (1 + r_i(t))^{2-2/L} \, dt. \tag{54}$$

Since $|r_i(t)| \leq \rho$,

$$(1 - \rho)^{2-2/L} \leq (1 + r_i(t))^{2-2/L} \leq (1 + \rho)^{2-2/L}.$$

Substituting these bounds into (54) yields

$$\frac{\log(1/\epsilon)}{\eta_G L (1 + \rho)^{2-2/L}} \leq T_{i,\epsilon}^{\mathrm{GD}} \leq \frac{\log(1/\epsilon)}{\eta_G L (1 - \rho)^{2-2/L}}. \tag{55}$$

For $\rho \ll 1$,

$$(1 \pm \rho)^{-(2-2/L)} = 1 + O(\rho),$$

so

$$T_{i,\epsilon}^{\mathrm{GD}} = \frac{\log(1/\epsilon)}{\eta_G L} \left(1 + O(\rho)\right). \tag{56}$$

The bound is uniform over modes. Therefore, for any two active modes i and j,

$$\frac{T_{i,\epsilon}^{\mathrm{GD}}}{T_{j,\epsilon}^{\mathrm{GD}}} = 1 + O(\rho). \tag{57}$$

We next consider SPECGD. Before saturation, the sign in (50) is constant. Suppose first that $\delta_i > 0$. Separating variables gives

$$dt = \frac{(1 + r_i)^{-1+1/L}}{\eta_S L} \, dr_i.$$

Integrating from $r_i(0) = 0$ to $r_i(T_i^{\mathrm{Spec}}) = \delta_i$ gives

$$T_i^{\mathrm{Spec}} = \frac{1}{\eta_S L} \int_0^{\delta_i} (1+r)^{-1+1/L}\, dr = \frac{(1+\delta_i)^{1/L} - 1}{\eta_S}. \tag{58}$$

The same calculation for $\delta_i < 0$ gives

$$T_i^{\mathrm{Spec}} = \frac{\left|(1+\delta_i)^{1/L} - 1\right|}{\eta_S}. \tag{59}$$

Using the expansion

$$(1+\delta_i)^{1/L} - 1 = \frac{\delta_i}{L} - \frac{L-1}{2L^2}\delta_i^2 + O(\delta_i^3),$$

we obtain

$$T_i^{\mathrm{Spec}} = \frac{|\delta_i|}{\eta_S L}\left(1 + O(\rho)\right), \tag{60}$$

for $|\delta_i| \le \rho \ll 1$. Consequently, for two nonzero correction modes,

$$\frac{T_i^{\mathrm{Spec}}}{T_j^{\mathrm{Spec}}} = \frac{|\delta_i|}{|\delta_j|}\left(1 + O(\rho)\right). \tag{61}$$

Equations (57) and (61) give the mode-balance comparison used in the main text. Near identity, GD has approximately common fractional convergence times across modes, whereas the SPECGD saturation times retain the relative magnitude differences of the residual corrections.

A.4 Proof of Theorem 4

Write

$$W = U\bar{W}V^\top.$$

Using Assumption 3,

$$\mathcal{L}_\ell = \frac{1}{2}\operatorname{tr}\left[\Gamma(\bar{W} - D)\Lambda(\bar{W} - D)^\top\right].$$

The gradient in the aligned basis is

$$\nabla_{\bar{W}}\mathcal{L}_\ell = \Gamma(\bar{W} - D)\Lambda.$$

Diagonal initialization is therefore preserved by both update rules. On the i-th mode,

$$g_i = \gamma_i\lambda_i(w_i - d_i) = c_i(w_i - d_i).$$

GD gives

$$\dot{w}_i = -\eta_G g_i = \eta_G c_i(d_i - w_i),$$

whose solution at $w_i(0) = 0$ is (35). Since $c_i > 0$, the sign of the active gradient is the sign of $w_i - d_i$. The partial matrix sign maps this active scalar to its sign and maps a zero gradient to zero. Hence

$$\dot{w}_i^{\mathrm{Spec}} = \eta_S \mathbf{1}\{w_i < d_i\},$$

which gives (36).

A.5 Proof of Theorem 5

At initialization,

$$\dot{w}_i^{\mathrm{GD}}(0) = \eta_G c_i d_i, \qquad \dot{w}_i^{\mathrm{Spec}}(0) = \eta_S.$$

Therefore

$$\left\|\dot{w}^{\mathrm{GD}}(0)\right\|_2 = \eta_G \left\|c \odot d\right\|_2, \qquad \left\|\dot{w}^{\mathrm{Spec}}(0)\right\|_2 = \eta_S\sqrt{r}.$$

The choices in (39) make both norms equal to one. Since

$$\dot{\mathcal{R}} = -\sum_i (d_i - w_i)\dot{w}_i,$$

evaluation at zero gives

$$-\dot{\mathcal{R}}_{\mathrm{GD}}(0) = \frac{\sum_i c_i d_i^2}{\|c \odot d\|_2}, \qquad -\dot{\mathcal{R}}_{\mathrm{Spec}}(0) = \frac{1}{\sqrt{r}} \sum_i d_i.$$

This proves (41). A strict initial gap persists on a sufficiently small time interval by continuity.

A.6 An imbalanced correction family

The condition in Theorem 5 includes imbalanced targets. Let $d_i > 0$ and set $c_i = \kappa/d_i^2$ for a constant $\kappa > 0$. If d^{-1} denotes the elementwise reciprocal of d, the initial decreases under SpecGD and GD are

$$-\dot{\mathcal{R}}_{\mathrm{Spec}}(0) = \frac{\|d\|_1}{\sqrt{r}}, \qquad -\dot{\mathcal{R}}_{\mathrm{GD}}(0) = \frac{r}{\|d^{-1}\|_2}.$$

The inequality $\|d\|_1 \left\|d^{-1}\right\|_2 \geq r^{3/2}$ shows that the spectral rate is at least as large, with equality only when all d_i are equal. For $d = (1, \varepsilon)$ and $c = (\varepsilon^2, 1)$ with $0 < \varepsilon < 1$, the two rates are $(1+\varepsilon)/\sqrt{2}$ and $2\varepsilon/\sqrt{1+\varepsilon^2}$. Their strict ordering follows from

$$(1+\varepsilon)^2(1+\varepsilon^2) - 8\varepsilon^2 = (1-\varepsilon)^2(\varepsilon^2 + 4\varepsilon + 1) > 0.$$

B Extensions of the local model

B.1 Finite-time stability under approximate local alignment

Proposition 6 (Finite-time stability under approximate local alignment)**.** *Work in the reference bases (U, V) from Assumption 3 and write*

$$U^\top H_{\mathrm{out}} U = \Gamma + B_L, \qquad V^\top H_{\mathrm{in}} V = \Lambda + B_R, \qquad U^\top \Delta^\star V = D + C,$$

with $\|B_L\|_2 \leq \delta_L$, $\|B_R\|_2 \leq \delta_R$, and $\|C\|_F \leq \delta_\Delta$. Assume the geometry matrices and trajectories remain bounded on $[0, T]$. Let $W_{\mathrm{GD},0}$ and $W_{\mathrm{Spec},0}$ denote the aligned reference trajectories from the same initialization. Then

$$\sup_{t\in[0,T]} \|W_{\mathrm{GD}}(t) - W_{\mathrm{GD},0}(t)\|_F \leq C_T(\delta_L + \delta_R + \delta_\Delta). \tag{62}$$

If the active reference gradient satisfies

$$\inf_{t\in[0,T]} \sigma^+_{\min}(\nabla\mathcal{L}_\ell, 0(W_{\mathrm{Spec},0}(t))) \geq m > 0, \tag{63}$$

then, before the first saturation event,

$$\sup_{t\in[0,T]} \|W_{\mathrm{Spec}}(t) - W_{\mathrm{Spec},0}(t)\|_F \leq C_{T,m}(\delta_L + \delta_R + \delta_\Delta). \tag{64}$$

The result is local in time. The lower singular-value bound is needed because the polar map ceases to be smooth when an active gradient singular value approaches zero.

Proof of Proposition 6

In the reference basis, define

$$E := \bar{W} - D.$$

The perturbed gradient is

$$\bar{G} = (\Gamma + B_L)(E - C)(\Lambda + B_R),$$

while the aligned reference gradient is

$$G_0 = \Gamma E \Lambda.$$

Expanding the difference and using submultiplicativity gives

$$\begin{aligned}\left\|\bar{G} - G_0\right\|_F \leq\ & \left(\delta_L \left\|\Lambda\right\|_2 + \left\|\Gamma\right\|_2 \delta_R + \delta_L \delta_R\right) \left\|E\right\|_F \\ & + \left(\left\|\Gamma\right\|_2 + \delta_L\right) \left(\left\|\Lambda\right\|_2 + \delta_R\right) \delta_\Delta. \end{aligned} \tag{65}$$

On a bounded finite-time trajectory, the right-hand side is bounded by a constant times

$$\delta_L + \delta_R + \delta_\Delta.$$

For GD, vectorizing the homogeneous part gives the reference generator

$$K_0 = \Lambda \otimes \Gamma$$

and perturbed generator

$$K = (\Lambda + B_R) \otimes (\Gamma + B_L).$$

Their difference satisfies

$$\left\|K - K_0\right\|_2 \leq \delta_R \left\|\Gamma\right\|_2 + \left\|\Lambda\right\|_2 \delta_L + \delta_L \delta_R.$$

Variation of constants, together with bounded semigroups on $[0, T]$ and the forcing induced by C, gives (62).

For SPECGD, perturbation bounds for the unitary polar factor imply local Lipschitz continuity when the active singular values are bounded below by $m > 0$ (Li, 1995). Condition (63) therefore gives

$$\left\|\operatorname{msgn}(\bar{G}) - \operatorname{msgn}(G_0)\right\|_F \leq C_m \left\|\bar{G} - G_0\right\|_F$$

on the interval before rank loss. Combining this bound with (65) and applying Grönwall's inequality proves (64).

B.2 FACTORIZED RESIDUAL BRANCHES

Theorem 7 (Mode dynamics for a balanced factorized correction)**.** *Consider an aligned local mode with $c_i := \gamma_i \lambda_i > 0$ and a K-factor residual correction $u_i = \prod_{k=1}^{K} q_{ki}$. Suppose $q_{1i}(0) = \cdots = q_{Ki}(0) > 0$. While the factors remain positive, both GD and factorwise* SPECGD *preserve this balance, and*

$$\dot{u}_i^{\mathrm{GD}} = \eta_G K c_i u_i^{2-2/K} (d_i - u_i), \tag{66}$$

$$\dot{u}_i^{\mathrm{Spec}} = \eta_S K u_i^{1-1/K} \operatorname{sign}(d_i - u_i) \tag{67}$$

until the active mode saturates.

PROOF OF THEOREM 7

On one aligned mode, let

$$u_i = \prod_{k=1}^{K} q_{ki}$$

and

$$\ell_i = \frac{1}{2} c_i (u_i - d_i)^2.$$

The gradient of factor q_{ki} is

$$\frac{\partial \ell_i}{\partial q_{ki}} = c_i (u_i - d_i) \prod_{m \neq k} q_{mi}.$$

If all factors are equal and positive, every factor has the same gradient. Both GD and factorwise SPECGD therefore preserve the balanced manifold.

Write $q_{ki} = q_i$ on this manifold, so $u_i = q_i^K$. Under GD,

$$\dot{q}_i = \eta_G c_i (d_i - u_i) q_i^{K-1}.$$

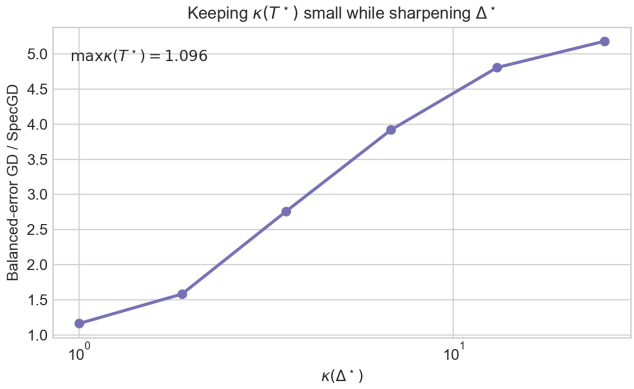


Figure 3: A controlled correction-spectrum sweep. The correction becomes more concentrated while the full target remains close to identity.

Hence

$$\begin{aligned}\dot{u}_i &= K q_i^{K-1} \dot{q}_i \\ &= \eta_G K c_i q_i^{2K-2} (d_i - u_i) \\ &= \eta_G K c_i u_i^{2-2/K} (d_i - u_i),\end{aligned}$$

which proves (66). Under factorwise SPECGD, positivity of the factors gives

$$\dot{q}_i = \eta_S \operatorname{sign}(d_i - u_i).$$

Therefore

$$\dot{u}_i = \eta_S K u_i^{1-1/K} \operatorname{sign}(d_i - u_i),$$

which proves (67).

C EXPERIMENTAL DETAILS AND SYNTHETIC CHECKS

C.1 CODE ORGANIZATION

The supplementary material contains the synthetic experiment driver, decoder pretraining launcher, Muon implementation, skip-strength sweep, plotting utilities, environment records, data indices, per-seed summaries, and trained MLP checkpoints.

C.2 SYNTHETIC LINEAR CHECKS

The synthetic studies are retained as controlled checks rather than the main empirical evidence. The depth comparison appears with the near-identity analysis in Section 4. The correction-spectrum sweep checks the separation between a near-isometric full map and an anisotropic correction, while the geometry experiment checks the mode times in Theorem 4.

C.3 SYNTHETIC SETTINGS

Depth comparison. Dimension $d = 32$. Depths $L \in \{2, 4, 8, 16, 32\}$. The plain model uses a diagonalized product parameterization, the residual model uses a diagonalized near-identity residual parameterization, and both are evaluated in an early-time regime.

Sharp-correction sweep. Dimension $d = 64$. The correction spectrum is $\delta_i = 0.1\rho^{i-1}$ with $\rho \in \{1.0, 0.99, 0.98, 0.97, 0.96, 0.95\}$. The local geometry is fixed to a geometric spectrum $\lambda_i = 0.9^{i-1}$. We report early unweighted correction error after 40 steps.

Isotropic versus anisotropic geometry. Dimension $d = 64$. The target correction is constant across modes, $d_i = 0.05$. We compare $\Sigma_z = I$ against a geometric spectrum with ratio 0.95.

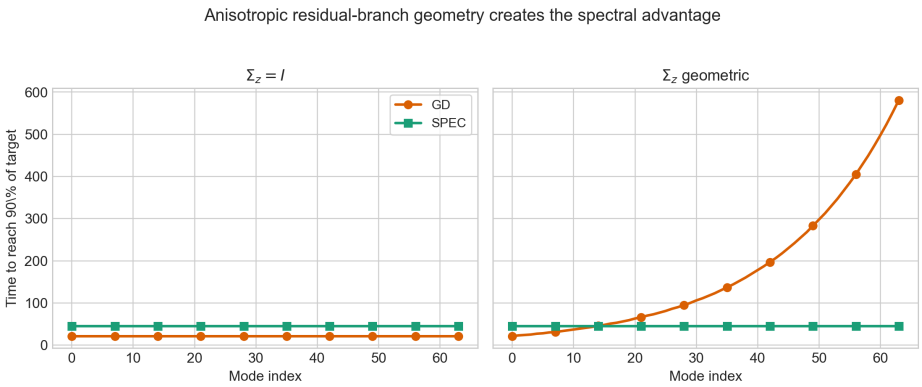


Figure 4: Mode-learning times under isotropic and anisotropic branch geometry with equal target amplitudes. GD time constants differ by $27.7\times$ under geometric covariance, while SPECGD keeps a common absolute velocity. Equal amplitudes are used here only to make absolute and fractional SPECGD times coincide.

Decoder pretraining diagnostic. The experiment uses an 8-layer pre-norm decoder with hidden size 384, 6 attention heads, MLP width 1536, context length 512, and vocabulary size 6400, for a total of 19.8M trainable parameters. The pretraining corpus is tokenized into 854,253 training chunks and 8,628 validation chunks of length 512, which corresponds to a single epoch of 13,347 optimizer steps at batch size 16 with gradient accumulation factor 4. Both optimizers use learning rate 3×10^{-4}, weight decay 0.1, and Adam-style β values $(0.9, 0.95)$. The Muon run applies Muon to all two-dimensional matrices except embeddings, layer norms, biases, and the tied LM head. These excluded parameters remain on AdamW. The Muon branch uses momentum 0.95, Nesterov acceleration, and 5 Newton–Schulz iterations per update.

The logger records training and validation losses every 100 optimizer steps, together with gradient spectra, parameter spectra, and per-layer forward-change statistics. To keep the AdamW and Muon diagnostics directly comparable, the paper averages the gradient metrics over the seven two-dimensional matrices that are tracked in both runs.

C.4 SKIP-STRENGTH ABLATION SETTING AND METRICS

The skip-strength experiment uses MNIST with a compact pre-norm MLP. Inputs are projected from 784 dimensions to a hidden width of 64. The network contains 8 blocks with branch

$$F_\ell(h) = W_{\ell,2}\,\mathrm{GELU}\left(W_{\ell,1}\,\mathrm{LN}(h)\right),$$

where the internal width is 128. The block output is

$$h_{\ell+1} = \alpha h_\ell + F_\ell(h_\ell)$$

with $\alpha \in \{0, 0.5, 1\}$. All settings use the same branch architecture, initialization rule, classifier head, and optimizer. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and 10 epochs. The fixed data split contains 20,000 training examples and 2,000 held-out examples.

After training, exact 64×64 Jacobians are computed for 32 held-out examples at every block. For each residual Jacobian $J_{\mathrm{res},\ell}$, the recorded quantities are

$$\mathrm{sr}(J) = \frac{\|J\|_F^2}{\|J\|_2^2},$$

the top singular energy

$$\frac{\sigma_1(J)^2}{\|J\|_F^2},$$

and the spectral effective rank

$$\exp\left(-\sum_i p_i \log p_i\right), \qquad p_i = \frac{\sigma_i(J)^2}{\sum_j \sigma_j(J)^2}.$$

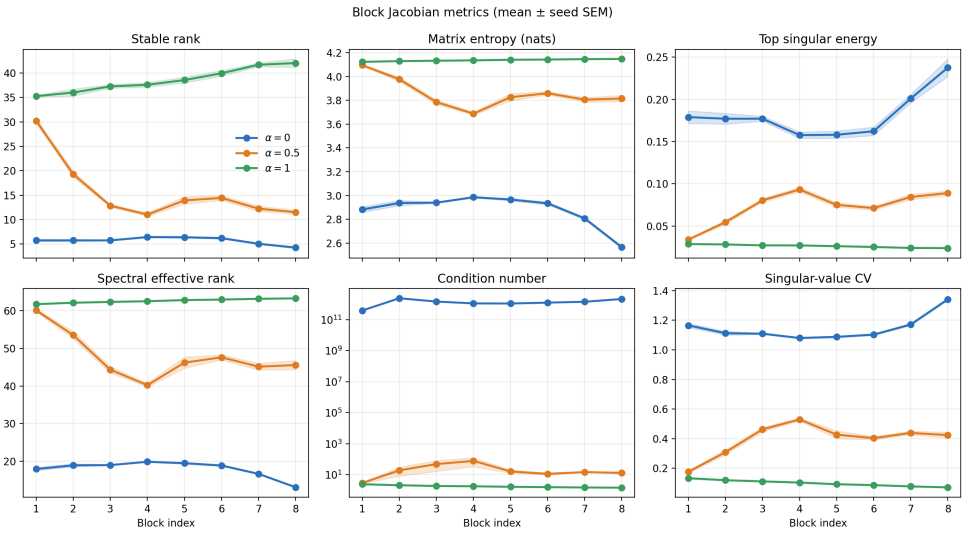


Figure 5: Full-block Jacobian metrics for the skip-strength sweep. The unit-skip model has the broadest spectrum and the smallest condition number and singular-value coefficient of variation. Curves show the mean over three seeds, and bands show seed-level standard errors.

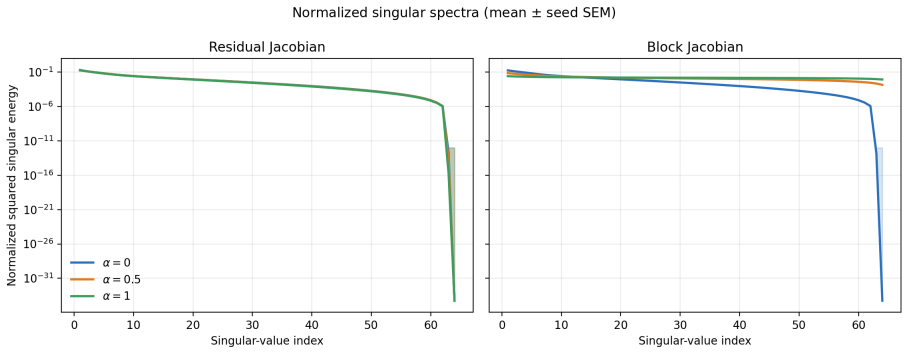


Figure 6: Normalized squared singular-value spectra averaged over layers, held-out examples, and seeds. The unit skip changes the full-block spectrum much more strongly than the residual-branch spectrum, which illustrates how the identity path can mask residual anisotropy.

The same quantities are measured for

$$J_{\mathrm{block},\ell} = \alpha I + J_{\mathrm{res},\ell}.$$

The program first averages the 32 Jacobians within each trained model and then reports means and standard errors over three independent seeds. The analytic Jacobian implementation agrees with automatic differentiation to a maximum absolute error below 4.9×10^{-8} across the nine runs.

C.5 Effective rank of the optimization geometry

We quantify the anisotropy of the local optimization geometry in the diagnostic experiment of a 130M parameter model on C4 dataset. Rather than measuring the spectrum of the raw gradient, which depends jointly on the current residual and the surrounding optimization geometry, we directly characterize the input- and output-side geometry appearing in the local gradient dynamics.

For a linear layer ℓ, let $H_{\mathrm{in},\ell} \succeq 0$ and $H_{\mathrm{out},\ell} \succeq 0$ denote the input- and output-side second-moment operators. The corresponding linear operator acting on a matrix perturbation E_ℓ is

$$\mathcal{G}_\ell := H_{\mathrm{in},\ell}^\top \otimes H_{\mathrm{out},\ell}, \tag{68}$$

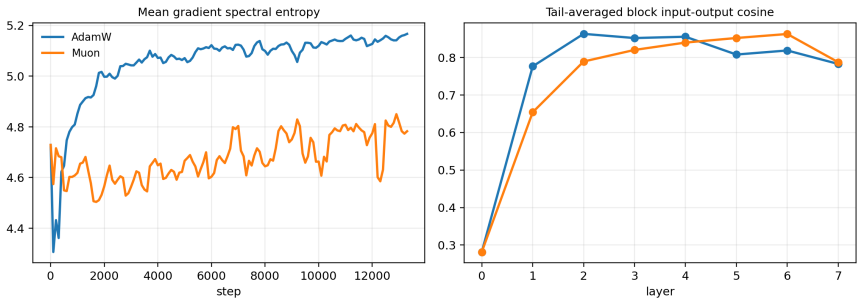


Figure 7: Additional diagnostics for the decoder pretraining run from Section 6. Left: mean gradient spectral entropy. Right: tail-averaged block input-output cosine, showing that later blocks are more aligned with their inputs than the earliest block.

Table 1: Normalized effective rank of the local optimization geometry at the final checkpoint of the 130M-parameter C4 diagnostic. Values are $100\,\rho_{\mathrm{geom},\ell}$; lower values indicate stronger anisotropy.

Optimizer	Layer 0	Layer 5	Layer 11	Mean
AdamW	1.88%	2.39%	3.17%	2.48%
SpecGD	2.98%	3.31%	5.71%	4.00%

since

$$\mathrm{vec}\left(H_{\mathrm{out},\ell} E_\ell H_{\mathrm{in},\ell}\right) = \mathcal{G}_\ell\,\mathrm{vec}(E_\ell). \tag{69}$$

For a positive semidefinite matrix $H \in \mathbb{R}^{d\times d}$, we define its effective rank as

$$r_{\mathrm{eff}}(H) := \frac{\mathrm{tr}(H)}{\lambda_{\max}(H)}. \tag{70}$$

This quantity lies in $[1, d]$ for nonzero H. It equals d when all eigenvalues are equal and becomes smaller as the spectrum becomes more concentrated.

The eigenvalues of the Kronecker operator $\mathcal{G}_\ell$ are all pairwise products of the eigenvalues of $H_{\mathrm{in},\ell}$ and $H_{\mathrm{out},\ell}$. Hence,

$$r_{\mathrm{eff}}(\mathcal{G}_\ell) = r_{\mathrm{eff}}(H_{\mathrm{in},\ell})\, r_{\mathrm{eff}}(H_{\mathrm{out},\ell}). \tag{71}$$

Directly normalizing this quantity by the ambient matrix-space dimension $d_{\mathrm{in},\ell} d_{\mathrm{out},\ell}$ would multiply the normalized effective ranks of the two sides and can therefore produce numerically very small percentages. For a more interpretable dimension-normalized measure, we report their geometric mean,

$$\rho_{\mathrm{geom},\ell} := \sqrt{\frac{r_{\mathrm{eff}}(\mathcal{G}_\ell)}{d_{\mathrm{in},\ell} d_{\mathrm{out},\ell}}} = \sqrt{\frac{r_{\mathrm{eff}}(H_{\mathrm{in},\ell})}{d_{\mathrm{in},\ell}} \frac{r_{\mathrm{eff}}(H_{\mathrm{out},\ell})}{d_{\mathrm{out},\ell}}}. \tag{72}$$

We report $100\,\rho_{\mathrm{geom},\ell}$ as a percentage. A value of 100% corresponds to isotropic input and output geometry up to an overall scale, whereas a small value indicates substantial spectral concentration and hence strong anisotropy. Importantly, $\rho_{\mathrm{geom},\ell}$ is a normalized effective-dimension measure rather than the fraction of algebraically active parameter directions.

We evaluate this quantity at layers 0, 5, and 11 using the same matrix type at each depth. Since the architecture and corresponding matrix dimensions are identical across the three probed layers, the aggregate value reported in the main text is the arithmetic mean

$$\overline{\rho}_{\mathrm{geom}} = \frac{1}{3}\left(\rho_{\mathrm{geom},0} + \rho_{\mathrm{geom},5} + \rho_{\mathrm{geom},11}\right). \tag{73}$$