OPTIMIZER GEOMETRY SETS THE PACE: SPECTRAL LEARNING DYNAMICS IN MATRIX FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent successes of matrix- and curvature-based optimizers have renewed interest in how update geometry shapes learning. These methods normalize or precondition updates, changing how different components progress during training. In deep matrix factorization, the geometry that slows gradient descent (GD) favors low-rank solutions by delaying the emergence of small singular modes. This raises the question of what remains of that spectral bias when normalization or curvature correction weakens or removes the slowdown. We study how optimizer geometry and depth jointly govern this behavior through a common framework for singular mode dynamics. Under explicit balance and alignment assumptions, we derive birth, saturation, and decay laws for Euclidean GD, coordinate-wise updates of SignGD and an instantaneous Adam approximation, spectral updates of Muon and cumulative Shampoo, and a block curvature model of K-FAC. The resulting picture is not a simple ordering from stronger to weaker low-rank bias. To highlight, SignGD and ideal Muon eliminate the divergent birth barrier and drive unsupported modes to zero in finite time. Cumulative Shampoo initially retains GD's depth-dependent barrier, then accumulated gradients produce a catch-up phase while making previously active modes increasingly persistent. Undamped K-FAC cancels the factorization-induced slowdown while preserving the ordering of the target singular values, whereas positive damping introduces a spectral threshold below which the slow GD phase laws reappear. These results give normalization, accumulated state, damping, and depth a direct interpretation as controls determining when modes emerge, persist, and disappear during training.

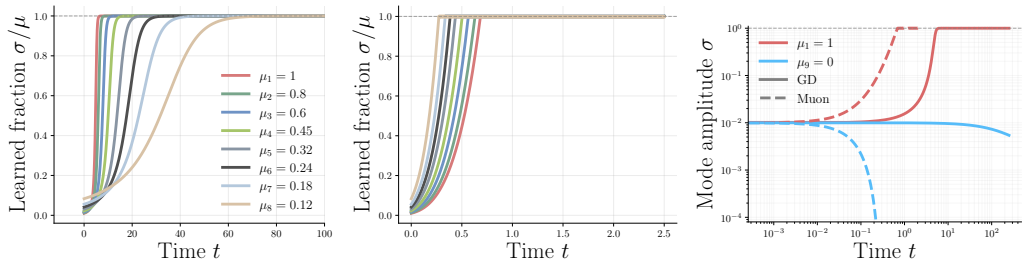
1 INTRODUCTION

For much of modern deep learning, large scale optimization was dominated by SGD and Adam-like coordinate-wise methods, while matrix preconditioning and approximate curvature methods remained comparatively limited in practice. Their additional memory, numerical, and systems requirements made their advantages difficult to realize at scale (Martens & Grosse, 2015; Gupta et al., 2018). This landscape is beginning to change. Recent foundation models with multi-billion and trillion parameters (Team et al., 2025; Liu et al., 2025; Khona et al., 2026; Ji et al., 2026) are effectively trained with spectral optimizers such as Muon (Jordan et al., 2024) and SOAP (Vyas et al., 2025) at substantially large model, data, and batch scales. This empirical shift raises a theoretical question. The implicit bias analysis of gradient descent (GD) shows that Euclidean geometry slows the progress of weak singular modes and thereby promotes low-rank solutions in deep matrix factorization (Arora et al., 2019; Li et al., 2021). Whereas spectral normalization and curvature correction are designed to reduce precisely this kind of anisotropy, in idealized limits they can weaken or even cancel the singular value dependent slowdown. Their recent success therefore raises the question of whether GD's spectral bias induced by depth is lost, preserved in a weaker form, or replaced by a qualitatively different mechanism of mode selection. More broadly, we ask:

> *When do modern optimizers merely accelerate a common learning trajectory, and when do they change which spectral components emerge, persist, or disappear?*

We investigate these questions in deep linear networks, where depth changes the parameterization and optimization dynamics without changing the end-to-end linear functions. This makes it a particularly clean setting for separating optimizer-induced dynamics from parameterization-induced bias.

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.



(a) GD (b) Muon (c) GD vs Muon: Birth and decay

Figure 1: **Optimizer geometry sets the pace of spectral learning.** In balanced aligned deep matrix factorization, GD and Muon produce distinct spectral learning trajectories (experimental details in Sec. F). **(a)** Under GD, larger target modes are learned earlier, while smaller modes remain suppressed for longer. **(b)** Under Muon, factor amplitudes grow at a common speed, allowing smaller targets to reach full recovery earlier. **(c)** From the same initial amplitude 0.01, Muon accelerates the growth of a positive target mode (red) and, importantly, removes the zero target mode in finite time (blue). GD instead exhibits delayed growth and slow decay. Thus, faster emergence of weak modes does not imply weaker removal of unsupported modes. Optimizer geometry determines which spectral components emerge, persist, and disappear. Other optimizers are discussed in later sections.

Our theoretical analysis focuses on deep matrix factorization, a canonical special case in which the evolution of singular modes can be characterized explicitly. Specifically, in full observation regression, it exposes transient singular mode learning (Saxe et al., 2013; Soudry et al., 2018), while in underdetermined matrix sensing and completion, it permits direct study of which interpolating matrix is selected (Arora et al., 2019; Li et al., 2021; Rhee et al., 2026; Parviz et al., 2026).

In contrast to the Euclidean geometry of GD, non-Euclidean optimizers transform the factorization-induced geometry in qualitatively different ways. SignGD and Adam apply coordinate-wise adaptive scaling (Bernstein et al., 2018; Kingma & Ba, 2014); Muon acts directly on the spectrum of matrix updates (Jordan et al., 2024) while Shampoo introduces history-dependent matrix preconditioning (Gupta et al., 2018); and K-FAC attempts to compensate for the parameterization through curvature using approximate natural gradients (Martens & Grosse, 2015). In our study, we categorize these geometries based on their updates into *Euclidean* (GD), *coordinate-wise* (SignGD and Adam), *spectral* (Muon and Shampoo) and *curvature* based (K-FAC). These methods can reach nearly identical training losses while producing weights with different spectra, representations, and performance. Recent studies have begun to identify optimizer-dependent feature and simplicity biases, including differences between Adam and GD in nonlinear models (Cattaneo et al., 2023; Zhang et al., 2024; Xie & Li, 2024) and between spectral and Euclidean updates in deep linear networks (Dragutinović & Ranganath, 2026). These findings further motivate the central question we posed, and answering this requires distinguishing how quickly modes emerge from how they settle or disappear, since an optimizer can affect these phases differently.

Our approach separates the geometry induced by the factorization from the update rule and state maintained by the optimizer. The product dynamics first express how each layer updates combine into the dynamics of the effective matrix. Under explicit balance, target alignment, and state compatibility assumptions, this description reduces to a common equation for each singular mode. The factorization determines how factor dynamics changes the mode amplitude, while the optimizer determines the update through its normalization, accumulated statistics, and damping. This gives a controlled setting in which these mechanisms can be compared without attributing differences to changing singular directions. Our analysis separates three aspects of spectral learning that are often conflated: the *birth* of initially small nonzero modes to reach a prescribed amplitude, *saturation* near their targets, and *decay* of unsupported modes whose target is zero. These phases need not respond to an optimizer in the same way. A method can permit rapid emergence while retaining a slow decay tail, and accumulated gradient history can change the rate at which an already learned component settles. Consequently, the emergence of small modes alone does not describe the spectral bias of an entire trajectory as studied in prior works. Our results make these distinctions explicit as shown in Fig. 1. Moreover, it also brings out how optimizer state parameters, including adaptive stabilization and curvature damping, interpolate between distinct rank selection regimes, and verifies that these

differences persist in underdetermined matrix factorization and regression. **Our contributions** are:
(**1**) **A common framework for spectral learning.** We express effective matrix dynamics through the layer updates and derive an exact mode reduction under explicit balance and alignment assumptions. The reduction separates the contribution of the factorization from the optimizer update states.

(**2**) **Phase laws revealing how optimizer settings shape the bias of depth.** We derive birth, saturation, and decay laws that distinguish normalization, accumulated history, and damping. GD and SignGD dynamics establish contrasting reference behaviors. Instantaneous Adam identifies how a positive stabilizer restores slow dynamics at small gradient magnitudes. Ideal Muon accelerates both emergence and removal of modes in finite time. For cumulative Shampoo, an exact accumulator identity quantifies the transition from initialization dominated to history dominated dynamics. For the specified K-FAC block model, an exact birth time decomposition separates an undamped contribution from a damping-weighted GD contribution. Together, these results identify when optimizer geometry preserves, modifies, or cancels the suppression of small modes induced by depth.

(**3**) **Numerical validation of phase laws and regime transitions.** Controlled experiments in deep matrix factorization and regression verify the derived phase laws. These experiments demonstrate that settings commonly associated with numerical stability can change relative spectral progress, determining which components emerge, persist, and disappear during training.

2 Related Work

Optimization geometry and solution bias. Optimization geometry influences which solution is selected among those that fit the training data (Wilson et al., 2017; Gunasekar et al., 2018). For Adam, implicit bias analyses establish margin characterizations in separable classification under specified optimizer and learning rate assumptions (Zhang et al., 2024; Gronich & Vardi, 2026). Natural gradient provides another illustration of the role of geometry: under suitable rank conditions, its effective matrix dynamics can eliminate the dependence on depth that arises under Euclidean gradient flow (Kerekes et al., 2021; Bernacchia et al., 2018). These results motivate examining both the symmetries respected by an optimizer and the trajectory it induces. Our framework complements this perspective by expressing changes in unobserved matrix directions through accumulated optimizer induced motion, while using mode dynamics to isolate the effects of different optimizers.

Training dynamics in matrix factorization. Analyses of gradient flow and gradient descent connect spectral dynamics to low rank recovery and incremental rank growth under specified initialization and measurement assumptions (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2021; Jin et al., 2023). While depth strengthens the tendency of low rank bias, it is not always equivalent to minimizing a single Schatten norm, indicating that trajectory-level geometry is essential (Razin & Cohen, 2020). Recent Muon analyses characterize how polar normalization changes spectral learning and simplicity bias in deep linear networks and linear attention (Dragutinović & Ranganath, 2026), and establish approximately uniform singular value growth under assumptions on LoRA factorization (Kang et al., 2026). For fully observed bilinear matrix factorization, Rhee et al. (2026) derive uniform factor singular value growth after alignment, and analyze oscillations caused by finite optimization steps. Complementary empirical comparisons show that Muon's relative performance depends on the factorization problem and hyperparameter tuning (Parviz et al., 2026). Shampoo has been interpreted as an adapted version of spectral descent that retains information about stochasticity and the parameter trajectory (Eschenhagen et al., 2026). Building on these results, we compare mode birth, decay, and saturation within a common balanced aligned model. While most analyses study one optimizer in isolation, our analysis includes SignGD and simplified Adam, derives an exact accumulator identity for cumulative Shampoo, and quantifies the effect of additive damping in K-FAC. We give a broader discussion on related works in Sec. A.

3 Preliminaries

Notations. For matrices X and Y of the same dimensions, $\langle X, Y\rangle_F = \mathrm{tr}(X^\top Y)$ denotes the Frobenius inner product. We use $\|\cdot\|_F$ for the Frobenius norm of a matrix, which reduces to the Euclidean norm for a vector, and matrix I_d denotes the identity matrix of size $d \times d$. The notation $\nabla_W f(W)$ denotes the gradient of f with respect to W. The symbols $\odot$ and $\oslash$ denote entrywise multiplication and division, respectively, with $H^{\odot 2} = H \odot H$. In entrywise expressions, square roots are also applied entrywise. A dot denotes differentiation with respect to time, so $\dot{a} = da/dt$.

Optimizer	Z_ℓ	Update category
GD	H_ℓ	Euclidean
SignGD (Bernstein et al., 2018)	$\mathrm{sign}(H_\ell)$	Coordinate-wise
Adam (Kingma & Ba, 2014)	$m_\ell \oslash (\sqrt{v_\ell} + \epsilon)$ with $\tau_m \dot{m}_\ell = H_\ell - m_\ell,\ \tau_v \dot{v}_\ell = H_\ell^{\odot 2} - v_\ell$	Coordinate-wise (adaptive)
Ideal Muon (Jordan et al., 2024)	$U_\ell V_\ell^\top$ where $H_\ell = U_\ell \Sigma_\ell V_\ell^\top$	Spectral
Shampoo (Gupta et al., 2018)	$C_\ell^{-1/4} H_\ell D_\ell^{-1/4}$ with $\lambda_{\mathrm{Sh}} > 0$, $C_\ell(t) = \lambda_{\mathrm{Sh}} I + \int_0^t H_\ell(u) H_\ell(u)^\top du$, $D_\ell(t) = \lambda_{\mathrm{Sh}} I + \int_0^t H_\ell(u)^\top H_\ell(u)\, du$	Spectral
K-FAC (Martens & Grosse, 2015)	$\mathrm{vec}(Z_\ell) = (\mathcal{F}_\ell + \lambda_{\mathrm{KF}} I)^{-1} \mathrm{vec}(H_\ell)$, $\mathcal{F}_\ell = (A_\ell A_\ell^\top) \otimes (B_\ell^\top B_\ell)$, $\lambda_{\mathrm{KF}} \geq 0$	Curvature based

Table 1: Layer update directions in the flow $\dot{W}_\ell = -Z_\ell$, where $H_\ell = B_\ell^\top G A_\ell^\top$. SignGD and Adam applies coordinate-wise updates, Muon uses polar normalization, Shampoo uses cumulative statistics, and K-FAC uses additive block damping. Detailed account of these optimizers is in Sec. B.

Deep matrix factorization. Let $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ denote the parameters of layer ℓ in a depth L linear network with $L \geq 2$. The effective matrix is

$$W = W_L W_{L-1} \cdots W_2 W_1. \tag{1}$$

We assume $d_\ell \geq \min(d_0, d_L)$ for every hidden layer, so that the architecture does not impose a rank bottleneck. Let $S_1, \ldots, S_m \in \mathbb{R}^{d_L \times d_0}$ be known measurement matrices and let $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$ be the observed values. We define $\mathcal{A}(W) = (\langle S_j, W \rangle_F)_{j=1}^m$, and

$$\mathcal{L}(W) = \frac{1}{2} \sum_{j=1}^m ((\mathcal{A}(W))_j - y_j)^2, \qquad G = \nabla_W \mathcal{L}(W) = \sum_{j=1}^m ((\mathcal{A}(W))_j - y_j) S_j.$$

The model is trained by minimizing the mean squared loss $\mathcal{L}(W)$. In noiseless recovery, $M \in \mathbb{R}^{d_L \times d_0}$ denotes the unknown target matrix and $y_j = \langle S_j, M \rangle_F$. Matrix completion corresponds to selecting individual entries of M through the measurement matrices. Full observation gives $\mathcal{L}(W) = \|W - M\|_F^2/2$. The gradient with respect to the effective matrix is G. In particular, $G = W - M$ under full observation. Define the prefix and suffix products

$$A_\ell = W_{\ell-1} \cdots W_1, \qquad B_\ell = W_L \cdots W_{\ell+1}, \qquad A_1 = I_{d_0}, \quad B_L = I_{d_L}. \tag{2}$$

The chain rule gives the layer gradient

$$H_\ell = \nabla_{W_\ell} \mathcal{L}(W) = B_\ell^\top G A_\ell^\top. \tag{3}$$

For full observation, $\mathcal{L}(W) = \|W - M\|_F^2/2$ has the unique minimizer M for any depth. Consequently, different spectra at finite time describe transient learning and cannot establish different interpolating solutions. However, the training trajectories differ across optimizers. Therefore, we focus on how depth and optmizer geometry affect the spectrum before convergence in the full observation setting and distinguish finite time spectral bias from asymptotic solution selection throughout.

Optimizer conventions. We study deterministic flows driven by the full gradient, without weight decay or explicit regularization. For each layer, the continuous time dynamics of the matrix W_ℓ is

$$\dot{W}_\ell(t) = -Z_\ell(t),$$

where t is continuous time, a dot denotes its derivative, The flow uses unit learning rate, with any constant common learning rate absorbed into time and the stated memory times. All moment matrices have the same dimensions as W_ℓ. The equations for Z_ℓ for all the optimizers analyzed in this paper are given in Table 1 and each optimizer is discussed in detail in Sec. B. We present the exact effective matrix dynamics in Proposition 1 (proof in Sec. C) which is the main object in our analysis.

Proposition 1 (Exact effective matrix W dynamics)**.** *Using the definitions of A_ℓ and B_ℓ from Eq.* (2)*, for every layer trajectory satisfying $\dot{W}_\ell = -Z_\ell$, the effective matrix satisfies*

$$\dot{W} = -\sum_{\ell=1}^L B_\ell Z_\ell A_\ell.$$

4 A COMMON FRAMEWORK FOR OPTIMIZER GEOMETRY

We develop a common framework to compare how optimizer geometry controls the emergence, growth, and decay of singular modes. To derive explicit phase laws, we adopt the training dynamics analysis that connects low rank learning to the evolution of the effective matrix spectrum and consider a balanced, aligned setting in which singular directions remain fixed (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2021). This construction follows the exact mode analysis of deep linear networks (Saxe et al., 2013) and isolates the effects of different optimizer geometries.

We consider full observation with $\mathcal{L}(W) = \frac{1}{2}\|W_L \cdots W_1 - M\|_F^2$, and the gradient $G = W - M$. While the minimizer is unique, this setting fixes the target of every mode and allows its learning time to be compared across optimizers. We make the following assumptions.

Assumption 1 (Balanced initialization)**.** *The factors are balanced at initialization:* $W_\ell(0)W_\ell(0)^\top = W_{\ell+1}(0)^\top W_{\ell+1}(0)$, *for* $\ell = 1, \ldots, L-1$.

Balanced initialization is a standard condition in the analysis of deep matrix factorization (Arora et al., 2019; Rhee et al., 2026). For Euclidean gradient flow, $W_\ell(t)W_\ell(t)^\top - W_{\ell+1}(t)^\top W_{\ell+1}(t)$ is conserved for all t, so initial balance persists throughout training. For the other optimizers, preservation follows within the aligned construction from the compatibility condition stated below.

Assumption 2 (Alignment with the target)**.** *There exist orthogonal matrices* $Q_0, \ldots, Q_L$ *with* $Q_\ell \in \mathbb{R}^{d_\ell \times d}$ *where* $d = \min(d_0, d_L)$ *such that*

$$M = Q_L \operatorname{diag}(\mu_1, \ldots, \mu_d) Q_0^\top, \qquad \textit{and} \qquad W_\ell(0) = Q_\ell \operatorname{diag}(a_{\ell 1}(0), \ldots, a_{\ell d}(0)) Q_{\ell-1}^\top.$$

where $\mu_i \geq 0$ *and* $a_{\ell i}(0) > 0$. $Q_0, \ldots, Q_L$ *remain the reference bases for the mode analysis.*

Together, Assumptions 1 and 2 imply $a_{\ell i}(0)^2 = a_{\ell+1,i}(0)^2$, and positivity therefore gives a common initial amplitude $a_{\ell i}(0) = a_i(0)$ across layers. While equal layer strengths and decoupled singular directions underlie the explicit learning curves of Saxe et al. (2013), fixing the directions gives every optimizer the same initial spectral configuration, so differences in mode evolution can be attributed to their update geometry and state.

Assumption 3 (Compatible optimizer state and updates)**.** *All layers use the same optimizer hyperparameters and have identical initial states when expressed in their respective mode bases. The joint evolution of factors* W_ℓ *and optimizer states preserves the aligned form. In particular, whenever the factor amplitudes and mode states agree across layers, the update directions satisfy*

$$Z_\ell = Q_\ell \operatorname{diag}(z_1, \ldots, z_d) Q_{\ell-1}^\top, \qquad \ell = 1, \ldots, L,$$

This condition makes the reduction a property of a specified optimizer flow. It is satisfied by GD, Muon, and cumulative Shampoo in the aligned setting. For Adam and SignGD, diagonality must hold in the actual parameter coordinates, which can be imposed by taking $Q_\ell = I$ for every ℓ. For K-FAC, we use the full observation block curvature model with common additive damping. These choices retain the moment and preconditioner states responsible for the phase behavior studied below and shown in Sec. D. Under Assumption 3, equal initial amplitudes have equal updates across layers. Consequently, the factors remain balanced and admit the representation

$$W_\ell(t) = Q_\ell \operatorname{diag}(a_1(t), \ldots, a_d(t)) Q_{\ell-1}^\top.$$

The following reduction proved in Sec. D applies on intervals where the relevant amplitudes remain positive. At a zero crossing, any continuation is specified in the corresponding optimizer result.

Theorem 1 (Exact dynamics of aligned modes)**.** *Under Assumptions 1 to 3, define* $\sigma_i = a_i^L, q = 1 - \dfrac{1}{L}, p = 2q = 2 - \dfrac{2}{L}, r_i = \sigma_i - \mu_i$. *Then* W *and every layer gradient* H_ℓ *has the form*

$$W = Q_L \operatorname{diag}(\sigma_1, \ldots, \sigma_d) Q_0^\top, \qquad H_\ell = Q_\ell \operatorname{diag}(g_1, \ldots, g_d) Q_{\ell-1}^\top, \quad \textit{where } g_i = r_i \sigma_i^q.$$

The effective singular amplitudes satisfy $\dot{\sigma}_i = -L\sigma_i^q z_i$. *Here,* i *labels a fixed mode and does not require the amplitudes to remain sorted by magnitude.*

The proof follows by substituting the aligned factors into the layer gradient and differentiating $\sigma_i = a_i^L$, as detailed in Sec. D. The coefficient $L\sigma_i^q$ is common to every optimizer and comes

from the factorization. The optimizer determines z_i through the current gradient and its evolving state. For example, GD uses $z_i = g_i$ and yields $\dot{\sigma}_i = -L\sigma_i^p(\sigma_i - \mu_i)$ as also shown in Arora et al. (2019), whereas normalization and preconditioning modify this dependence.

Comparisons across depth. Unless stated otherwise, we compare by fixing the target matrix M and initial effective mode amplitudes $\sigma_i(0) = s_i$, implying balanced mode amplitudes $a_i(0) = s_i^{1/L}$. Optimizer settings and initial states are specified separately for each comparison. A common rescaling of time is used only to distinguish uniform acceleration from changes in relative mode evolution.

5 Optimizer Specific Results

In this section, we ignore mode index i for better readability. For a positive target μ, *birth* is the first passage from $\sigma(0) = s > 0$ to a fixed threshold $b \in (s, \mu)$. We denote it by $T_{\text{opt}}^{\text{birth}}(s, b)$. *Saturation* is the time required for the residual $r(t) = |\sigma(t) - \mu|$ to enter and remain below a tolerance $\tau > 0$. For an unsupported mode $\mu = 0$, *decay* is the passage from $\sigma(0) = s > 0$ to a tolerance $\delta \in (0, s)$, and is denoted by $T_{\text{opt}}^{\text{decay}}(s, \delta)$. Exact extinction means that the mode reaches $\sigma = 0$ in finite time and remains there. Birth always refers to escape from a positive seed, rather than spontaneous activation of an exactly zero factor mode. From Theorem 1, for $s < b$, time is $T_{\text{opt}}(s, b) = -\frac{1}{L}\int_s^b \frac{x^{-q}}{z}\, dx$.

Birth and decay exponents. Suppose the mode dynamics satisfies $\dot{\sigma} = c\sigma^\alpha$, where $c > 0$. Then $T(a, b) = \frac{1}{c}\int_a^b x^{-\alpha}\, dx$ with $0 < a < b$. Consequently,

$$T(a,b) = \frac{1}{c}\log\frac{|b|}{|a|} \text{ for } \alpha = 1, \qquad \text{and} \qquad T(a,b) = \frac{b^{1-\alpha} - a^{1-\alpha}}{c(1-\alpha)} \text{ for } \alpha \neq 1. \tag{4}$$

All the optimizers birth dynamics near $\sigma(0) = s = 0$ satisfies $\dot{\sigma} \sim c_{\text{b}}\sigma^{\alpha_{\text{b}}}$, with $c_{\text{b}} > 0$ for a target mode $\mu > 0$. Thus, $T^{\text{birth}}(s, b)$ from Eq. (4) as $s \to 0$, $\alpha_{\text{b}} < 1$ gives finite escape from a vanishing positive seed, $\alpha_{\text{b}} = 1$ gives a logarithmic birth barrier, and $\alpha_{\text{b}} > 1$ gives a polynomial birth barrier. Similarly, the decay dynamics for an unsupported mode $\mu = 0$ satisfies $\dot{\sigma} \sim -c_{\text{d}}\sigma^{\alpha_{\text{d}}}$, with $c_{\text{d}} > 0$. Thus, $T^{\text{decay}}(s, \delta)$ can be similarly computed. It gives that $\alpha_{\text{d}} < 1$ reaches zero in finite time, $\alpha_{\text{d}} = 1$ has exponential decay time, $\alpha_{\text{d}} > 1$ has algebraic decay and zero is reached only asymptotically.

Saturation exponent. Suppose that near a nonzero target, the residual is $r_0 = |\sigma(t_0) - \mu|$ at time t_0, satisfies $\dot{r} = -\beta r^\gamma$ with $\beta > 0$. For $\gamma = 0$, $r(t) = \max\{0, (r_0 - \beta t)\}$, thus saturates (decreases) linearly at rate β. For $\gamma = 1$, $r(t) = \exp(-\beta t) r_0$ saturates (decreases) exponentially at rate β.

We derive the phase laws for the optimizers and present them using these exponents in Theorem 2. All proofs are provided in Sec. E.

Theorem 2 (Optimizer phase laws)**.** *Based on the layer updates Z_ℓ in Table 1, the singular mode dynamics $\dot{\sigma}$ for each optimizer are derived, and the following table summarizes them with the phase laws expressing the birth, decay, and saturation exponents σ_{b}, σ_{d}, and γ, respectively.*

Optimizer	***Dynamics***	***Birth*** $\dot{\sigma} \sim c_{\text{b}}\sigma^{\alpha_{\text{b}}}$	***Decay*** $\dot{\sigma} \sim -c_{\text{d}}\sigma^{\alpha_{\text{d}}}$	***Saturation*** $\dot{r} \sim -\beta r^\gamma$
GD	$\dot{\sigma} = L\sigma^p(\mu - \sigma)$	$\alpha_{\text{b}} = p$ $c_{\text{b}} = L\mu$ ***Time:*** *Finite for $L = 1$, Asymptotic for $L > 1$*	$\alpha_{\text{d}} = p + 1$ $c_{\text{d}} = L$ ***Time:*** *Asymptotic*	$\gamma = 1$ $\beta = L\mu^p$ *Exponential*
SignGD & Muon	$\dot{\sigma} = L\sigma^q \operatorname{sign}(\mu - \sigma)$	$\alpha_{\text{b}} = q$ $c_{\text{b}} = L$ ***Time:*** *Finite*	$\alpha_{\text{d}} = q$ $c_{\text{d}} = L$ ***Time:*** *Finite*	$\gamma = 0$ $\beta = L\mu^q$ *Linear*
Adam	$\dot{\sigma} = \frac{L(\mu - \sigma)\sigma^p}{\lvert\mu - \sigma\rvert\sigma^q + \epsilon}$ $\kappa = \mu^{q+1}\frac{q^q}{(q+1)^{q+1}}$ $h(\sigma) = (\mu - \sigma)\sigma^q$ $\sigma_{\text{A}}^{\text{d}} = \epsilon^{1/(q+1)}$	*If* $h(\sigma)/\epsilon \to 0$ $\alpha_{\text{b}} = p$ $c_{\text{b}} = L\mu/\epsilon$ ***Time:*** *GD-like* *If* $\epsilon < \kappa$ & $\epsilon/h(\sigma) \to 0$ $\alpha_{\text{b}} = q$ $c_{\text{b}} = L$ ***Time:*** *SignGD-like*	*If* $\sigma(t) \ll \sigma_{\text{A}}^{\text{d}}$, $\alpha_{\text{d}} = p + 1$ $c_{\text{d}} = L/\epsilon$ ***Time:*** *GD-like* *If* $\sigma(t) \gg \sigma_{\text{A}}^{\text{d}}$ $\alpha_{\text{d}} = q$ $c_{\text{d}} = L$ ***Time:*** *SignGD-like*	*If* $r\mu^q \ll \epsilon$, $\gamma = 1$ $\beta = L\mu^p/\epsilon$ *GD-like* *Else* $\gamma = 0$ $\beta = L\mu^q$ *SignGD-like*

Shampoo	$\dot{\sigma} = \frac{L\sigma^p(\mu-\sigma)}{d}$ $d(t) = \sqrt{\lambda_{\text{Sh}} + h(t)}$ $h(t) = \int_0^t g(u)^2 du$ $h(0) = 0, d(0) = \sqrt{\lambda_{\text{Sh}}}$ $\nu(\sigma) = \frac{(\sigma-s)(2\mu-\sigma-s)}{4L}$ $d_0 = d(0)$ $d(\sigma) = d_0 + \nu(\sigma)$	Initial phase: $\nu(\sigma) \ll d_0$ $\alpha_{\text{b}} = p$ $c_{\text{b}} = \frac{L\mu}{d_0}$ ***Time:*** *GD-like* Accumulator phase: $s \ll \sigma \ll \mu$ & $2Ld_0 \ll \mu\sigma$ $\alpha_{\text{b}} = p - 1\ (< q)$ $c_{\text{b}} = 2L^2$ ***Time:*** *Finite across this interval*	Final decay: $\alpha_{\text{d}} = p + 1$ $c_{\text{d}} = \frac{4L^2}{4Ld_0 + s^2}$ ***Time:*** *GD-like*	$\gamma = 1$ $\beta = \frac{L\mu^p}{d_0 + \frac{(\mu-s)^2}{4L}}$ *GD-like* *If* $s \ll \mu$ & $d_0 \ll \frac{(\mu-s)^2}{4L}$ $\beta \approx 4L^2\mu^{p-2}$ *GD-like*, $\beta \propto \mu^{p-2}$: *order can reverse*
K-FAC	$\dot{\sigma} = \frac{L\sigma^p(\mu-\sigma)}{\sigma^p + \lambda_{\text{KF}}}$	Undamped: *When* $\lambda_{\text{KF}} = 0$ *or when* $\lambda_{\text{KF}} > 0$ & *if* $\sigma^p \gg \lambda_{\text{KF}}$ $\alpha_{\text{b}} = 0$ $c_{\text{b}} = L\mu$ ***Time:*** *Finite* Damped: *When* $\lambda_{\text{KF}} > 0$ & *if* $\sigma^p \ll \lambda_{\text{KF}}$ $\alpha_{\text{b}} = p$ $c_{\text{b}} = L\mu/\lambda_{\text{KF}}$ ***Time:*** *GD-like*	$\alpha_{\text{d}} = 1$ $c_{\text{d}} = L$ ***Time:*** *Asymptotic* $\alpha_{\text{d}} = p + 1$ $c_{\text{d}} = L/\lambda_{\text{KF}}$ ***Time:*** *GD-like*	$\gamma = 1$ $\beta = L$ *GD-like* $\gamma = 1$ $\beta = \frac{L\mu^p}{\mu^p + \lambda_{\text{KF}}}$ *GD-like*

Interpretation of Theorem 2 and insights. Let's start with **Euclidean GD**. The birth time $T_{\text{GD}}^{\text{birth}}(s,b)$ for $0 < s < b < \mu$, and decay time $T_{\text{GD}}^{\text{decay}}(s,\delta)$ for $\mu = 0$ are

$$T_{\text{GD}}^{\text{birth}}(s,b) = \begin{cases} \log\frac{\mu-s}{\mu-b}, & L=1, \\ \frac{1}{2\mu}\log\frac{b(\mu-s)}{s(\mu-b)}, & L=2, \\ \frac{s^{1-p}-b^{1-p}}{L\mu(p-1)} + O(1), & L>2. \end{cases} \qquad T_{\text{GD}}^{\text{decay}}(s,\delta) = \begin{cases} \log\frac{s}{\delta}, & L=1, \\ \frac{\delta^{-p}-s^{-p}}{Lp}, & L>1. \end{cases}$$

Consequently, as $s \to 0$, birth time $T_{\text{GD}}^{\text{birth}}(s,b)$ diverges for $L \geq 2$, and as $\delta \to 0$, decay time $T_{\text{GD}}^{\text{decay}}(s,\delta)$ diverges for all L. $T_{\text{GD}}^{\text{birth}}(s,b)$ clearly shows that $L=1$ has no initialization-dependent birth barrier, whereas $L=2$ introduces a logarithmic delay, while $L>2$ a polynomial delay. From the decay time, decay is logarithmic for $L=1$, and algebraic for $L>1$, with the polynomial tail becoming slower as depth increases. No unsupported mode reaches zero in finite time. The exponential saturation implies that modes with smaller target singular values converge slowly.

> *GD delays the growth of small positive modes while it removes zero target modes only asymptotically. Depth promotes low effective rank primarily by suppressing the birth of new modes, not by rapidly pruning active ones.*

For **SignGD**, the birth time $T_{\text{SignGD}}^{\text{birth}}(s,b)$ for $\mu > 0$, and decay time $T_{\text{SignGD}}^{\text{decay}}(s,\delta)$ for $\mu = 0$ are

$$T_{\text{SignGD}}^{\text{birth}}(s,b) = b^{1/L} - s^{1/L} \qquad T_{\text{SignGD}}^{\text{decay}}(s,\delta) = s^{1/L} - \delta^{1/L}.$$

Since the time to reach any target μ is $|s^{1/L} - \mu^{1/L}|$, birth, saturation and decay occur in *finite* time. For $\mu = 0$, the mode becomes exactly zero at time $s^{1/L}$. The birth time remains bounded as $s \to 0$.

Adam updates are considered under a quasi-steady approximation to isolate the effect of coordinate normalization, with $m = g$ and $v = g^2$ giving the instantaneous state approximation. If $|\sigma - \mu|\sigma^q \gg \epsilon$ then $\dot{\sigma} = -L\sigma^q \operatorname{sign}(\sigma - \mu)$, thus behaving like SignGD. If $|\sigma - \mu|\sigma^q \ll \epsilon$ then $\dot{\sigma} = -L\sigma^p(\sigma-\mu)/\epsilon$, thus behaving like GD upto $1/\epsilon$ factor. This approximation explains a possible transition between normalized and GD-type motion, but it is not used as an exact theorem for Adam.

> *SignGD exhibits birth, saturation and decay of all modes in finite time for every depth. Instantaneous Adam inherits this behavior only when the gradient magnitude dominates ϵ. For sufficiently small birth modes, unsupported modes and final saturation, it behaves like GD.*

Muon in the aligned model gives $\dot{\sigma} = -L\sigma^q \operatorname{sign}(\sigma-\mu)$, reducing it to a scalar sign rule as in SignGD updates. The disappearance of slow sequential learning under this idealization has been studied in Dragutinović & Ranganath (2026); Rhee et al. (2026). However, our phase laws highlight that the same normalization that accelerates weak mode emergence also removes zero target modes in finite time, so faster learning of small components can coexist with faster removal of unsupported ones.

For **cumulative Shampoo**, based on the common scalar gradient accumulator $h(t)$, and $d(t)$ defined in Theorem 2, along a nonstationary aligned mode starting at $\sigma(0) = s > 0$ with $d(0) > 0$, we have

$$d(\sigma) = d(0) + \frac{\mu(\sigma - s) - (\sigma^2 - s^2)/2}{2L} = d_0 + \nu(\sigma).$$

Shampoo's dynamics result in two distinct regimes: **Initial phase:** When the accumulator has barely changed, $d_0 \gg \nu(\sigma)$, its dynamics satisfy $\dot{\sigma} \approx L\sigma^p(\mu - \sigma)/d(0)$, recovering GD dynamics with $1/d_0$ factor. **Accumulator phase:** An intermediate regime arises on an interval $\sigma_a \le \sigma \le b$ satisfying $s \ll \sigma_a$, $b \ll \mu$, and $2Ld_0 \ll \mu\sigma_a$. Throughout this interval, accumulated history dominates and $\dot{\sigma} \approx 2L^2\sigma^{p-1}$, $T^{\mathrm{birth}}_{\mathrm{Sh}}(\sigma_a, b) \approx (b^{2/L} - \sigma_a^{2/L})/(4L)$. In this regime, the growth law is independent of μ, although total birth times retain target dependent ordering. Ideal Muon instead has target independent birth times to a common absolute threshold from the outset. Saturation near a positive target follows a separate local law, with $\gamma = 1, \beta = \frac{L\mu^p}{d_0 + (\mu - s)^2/(4L)}$. If $s \ll \mu$ and $4Ld_0 \ll (\mu - s)^2$, then $\beta \approx 4L^2\mu^{-2/L}$, so larger targets can have smaller local saturation rates. For zero target modes, the decay dynamics in this phase is $\dot{\sigma} = -\frac{L}{d(\sigma)}\sigma^{p+1}$, which is dominated by the $p + 1$ exponent as shown in Sec. E, similar to GD. Thus unsupported modes approach zero only asymptotically. At fixed initial amplitude and decay tolerance, a larger inherited denominator d_0 increases the decay passage time. Ideal Muon instead removes unsupported aligned modes in finite time.

Cumulative Shampoo has a two phase emergence behavior: GD like target sensitive birth first, followed by history adapted target independent catch-up phase. It has increasing historical inertia, making larger modes harder to remove, whereas Muon has finite time decay.

K-FAC updates each layer with the per-layer Fisher curvature block using a damping parameter λ_{KF}. From the dynamics, we get two cases: **Undamped case ($\lambda_{\mathrm{KF}} = 0$):** the dynamics near zero initialization reduces to $\dot{\sigma} = L(\mu - \sigma)$, thus $T^{\mathrm{birth}}_{\mathrm{KFAC}}(s, b) = \frac{1}{L}\log\frac{\mu - s}{\mu - b}$, resulting in finite barrier escape time with target-dependent mode ordering. Whereas Muon removes both the factorization barrier and target-amplitude ordering during aligned birth, distinguishing it from K-FAC. **Damped case ($\lambda_{\mathrm{KF}} > 0$):** When the damped factor dominates, $\lambda_{\mathrm{KF}} \gg \sigma^p$, $\dot{\sigma} = \frac{L\sigma^p(\mu - \sigma)}{\lambda_{\mathrm{KF}}}$, thus recovering GD-like behavior with $1/\lambda_{\mathrm{KF}}$ factor. Therefore, λ_{KF} defines a spectral mobility threshold. For a zero target, undamped K-FAC removes unwanted modes exponentially, but not in finite time, whereas damped K-FAC does it significantly slower, at only a polynomial rate.

Undamped K-FAC updates cancel the depth-dependent mobility barrier, while positive additive damping restores the GD birth asymptotic and a polynomial zero target tail.

6 Experiments

We numerically verify the phase laws in Theorem 2 and examine how optimizer-state parameters determine the active spectral regime in deep matrix factorization. Unless varied explicitly, we use the same balanced, aligned depth $L = 4$ factorization, effective initialization, and target modes across optimizers. Note that $p = 1.5$ and $q = 0.75$. We estimate the local birth and decay exponents from the trajectory slopes $\alpha_{\mathrm{b}}(\sigma) = \frac{d\log\dot{\sigma}}{d\log\sigma}$, $\alpha_{\mathrm{d}}(\sigma) = \frac{d\log(-\dot{\sigma})}{d\log\sigma}$. For saturation, we fit the local residual law $\dot{r} = -\beta r^{\gamma}$, $r = |\sigma - \mu|$, where β is an exponential rate when $\gamma = 1$ and a constant residual speed when $\gamma = 0$. Implementation details, hyperparameters and additional experiments are in Sec. F and G.

Verification of the phase laws in Theorem 2. Fig. 2 compares quantities measured from the trajectories with the theoretical predictions. For birth, GD retains exponent p, ideal Muon follows q, instantaneous Adam interpolates between the Euclidean and normalized regimes, cumulative Shampoo moves from p toward its history-adapted exponent $p - 1$, and undamped K-FAC removes the singular-magnitude dependence with $\alpha_{\mathrm{b}} = 0$. The decay curves distinguish finite extinction under ideal Muon, exponential decay under undamped K-FAC, and the polynomial tails of GD and cumulative Shampoo. The saturation experiment further reveals optimizer-dependent mode ordering: GD and Adam converge faster for larger targets; undamped K-FAC is approximately target independent while Muon shows dependence; history-dominated Shampoo reverses the GD ordering. Thus the optimizers change relative spectral progress rather than merely rescaling a common trajectory.

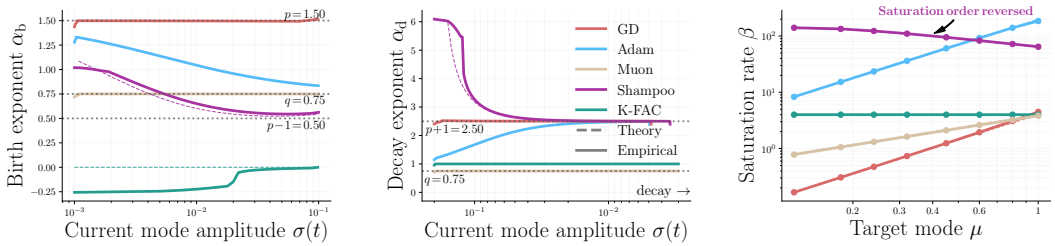
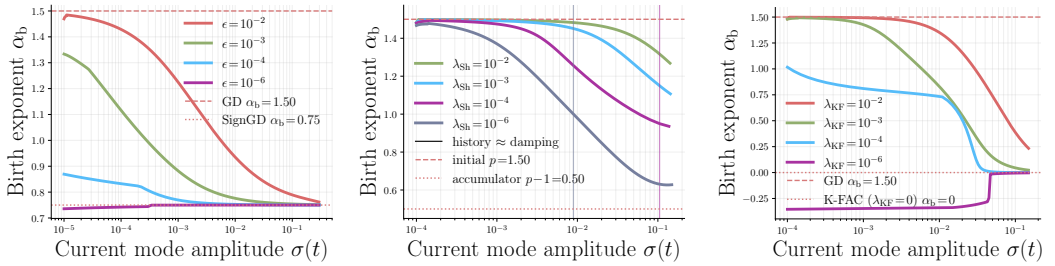


Figure 2: **Verification of the phase laws in Theorem 2 at depth $L = 4$. (a)** The birth exponent α_b separates Euclidean, normalized, history adapted, and curvature corrected dynamics. **(b)** The decay exponent α_d distinguishes finite extinction of Muon ($\alpha_d = q < 1$) from the other optimizers. **(c)** The saturation coefficient β reveals target dependent mode ordering.



(a) Adam ϵ controls the regime

(b) Two phases of Shampoo

(c) K-FAC $\lambda_{KF} > 0$ restores barrier

Figure 3: **Optimizer state parameters control the local birth regime.** Adam's stability parameter ϵ shifts the GD-to-SignGD crossover; Shampoo damping delays the transition from p to $p-1$; and K-FAC damping restores the GD exponent below its curvature threshold.

Optimizer state controls spectral crossovers. Fig. 3 varies the state parameters determined in Theorem 2 that separate the local regimes. For instantaneous Adam, decreasing ϵ moves the transition from the GD exponent p to the SignGD exponent q toward smaller modes. Large ϵ retains the Euclidean exponent over a wider range of mode amplitudes, whereas small ϵ yields the normalized exponent q almost immediately. For cumulative Shampoo, decreasing λ_{Sh} causes accumulated history to dominate earlier, advancing the transition from p to $p-1$. For K-FAC, positive damping restores the GD exponent when $\sigma^p \ll \lambda_{KF}$, whereas curvature correction removes the barrier when $\sigma^p \gg \lambda_{KF}$. As $\lambda_{KF} \to 0$, this transition moves toward smaller amplitudes, recovering the undamped K-FAC regime. These experiments confirm that stabilization, accumulated state, and damping determine which phase law governs a mode at a given spectral scale.

7 Conclusion

Our analysis shows that the spectral bias of depth depends on optimizer geometry and the stage of training. A surprising consequence is that rapid learning of weak components can coexist with rapid removal of unsupported ones, as illustrated by the Muon analysis. Conversely, accumulated history can accelerate the emergence of a mode while prolonging its persistence. These effects make a single ordering of optimizers by stronger or weaker low-rank bias inadequate.

A further implication is that parameters associated with numerical stability also shape learning bias. In the analyzed models, Adam's stabilizer, Shampoo's accumulator initialization, and K-FAC's damping determine where normalization, history, or curvature correction governs spectral progress. They can therefore change how strongly depth suppresses small modes, even within the same optimizer. Although the phase laws rely on balance and alignment assumptions, they identify concrete mechanisms to understand optimizer behavior. This perspective motivates evaluating optimizers through which components they accelerate, suppress, or preserve alongside the training loss, since comparable fit can conceal differences in their dynamics. The resulting framework provides a basis for understanding and testing how faster optimizers change the complexity of intermediate models.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, related work discovery, and selected coding and debugging tasks. They were not used to draft complete sections or to generate mathematical proofs. All proofs were developed by the authors, although AI tools were occasionally used to check arguments and suggest clearer formulations. Example use cases include code for plotting, logging, and refactoring the functions which were all verified to do what was intended. All references, code, and final text were independently verified and thoroughly reviewed by the authors, who take full responsibility for the content.

ETHICS STATEMENT

This work is theoretical and empirical research on optimization dynamics in deep matrix factorization. It does not involve human participants, personal or sensitive data, or deployment in high-stakes decision-making settings. The experiments primarily use controlled synthetic matrix factorization and regression problems designed to derive deeper insights about the optimizer behavior. Thus, we believe the potential benefits of this research outweigh its risks. Additionally, we do not see any immediate risk stemming from our work.

REPRODUCIBILITY STATEMENT

We state all assumptions, optimizer conventions, continuous-time update rules, and initialization conditions used in the theoretical analysis. Complete derivations and proofs are provided in the appendix, including the effective-matrix dynamics, aligned-mode reduction, and birth, decay, and saturation laws. For the experiments, we specify the data, depth, balanced initialization, optimizer hyperparameters, damping and stabilization parameters, stopping criteria, and procedures used to estimate the local phase exponents in Sec. F and G. We provide the source code and experiment configurations needed to reproduce the reported figures and tables in the supplementary material as a zip file.

REFERENCES

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pp. 244–253. PMLR, 2018.

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

Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the adam algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021.

Alberto Bernacchia, Máté Lengyel, and Guillaume Hennequin. Exact natural gradient in deep linear networks and its application to the nonlinear case. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Matias D Cattaneo, Jason M Klusowski, and Boris Shigida. On the implicit bias of adam. *arXiv preprint arXiv:2309.00079*, 2023.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Sara Dragutinović and Rajesh Ranganath. To use or not to use muon: How simplicity bias in optimizers matters. *arXiv preprint arXiv:2603.00742*, 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.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. *arXiv preprint arXiv:2602.09314*, 2026.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *Advances in Neural Information Processing Systems*, 38:39622–39669, 2026.

Rhys Gould and Hidenori Tanaka. Continuous-time analysis of adaptive optimization and normalization. *arXiv preprint arXiv:2411.05746*, 2024.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. *arXiv preprint arXiv:2602.16340*, 2026.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Yiping Ji, Mahalakshmi Sabanayagam, Peyman Moghadam, Hemanth Saratchandran, and Simon Lucey. The quantization benefits of residual-free transformers. *arXiv preprint arXiv:2605.25880*, 2026.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pp. 15200–15238. PMLR, 2023.

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

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform spectral growth and convergence of muon in lora-style matrix factorization. *arXiv preprint arXiv:2602.06385*, 2026.

Anna Kerekes, Anna Mészáros, and Ferenc Huszár. Depth without the magic: Inductive bias of natural gradient descent. *arXiv preprint arXiv:2111.11542*, 2021.

Mikail Khona, Aditya Vavre, Boxiang Wang, Deyu Fu, Hao Wu, Mike Chrzanowski, Bryan Catanzaro, Dheevatsa Mudigere, Jeff Pool, Michael Lightstone, et al. Soap, muon, and beyond: Pushing llm pretraining scales. *arXiv preprint arXiv:2607.20548*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AHOs7Sm5H7R.

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.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Ali Parviz, Gal Mishne, and Alex Cloninger. Reassessing muon for matrix factorization. *arXiv preprint arXiv:2607.13246*, 2026.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Mark Rhee, Jamie Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. *arXiv preprint arXiv:2606.30509*, 2026.

Mahalakshmi Sabanayagam, Pascal Esser, and Debarghya Ghoshdastidar. Analysis of convolutions, non-linearity and depth in graph neural networks using neural tangent kernel. *arXiv preprint arXiv:2210.09809*, 2022.

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.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Bhavya Vasudeva, Jung Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of adam and sgd. *Advances in neural information processing systems*, 38: 161725–161761, 2026.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pp. 10849–10858. PMLR, 2021.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Shuo Xie and Zhiyuan Li. Implicit bias of adamw: $\ell_i nfty$ norm constrained optimization. *arXiv preprint arXiv:2404.04454*, 2024.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Preliminaries **3**

4 A Common Framework for Optimizer Geometry **5**

5 Optimizer Specific Results **6**

6 Experiments **8**

7 Conclusion **9**

A Extended Related Work **15**

B Optimizers **16**

C Effective Matrix W Dynamics **18**

D Aligned Singular Mode Dynamics **18**

D.1 Balance under Euclidean Gradient Flow 19

D.2 Consequences of Balance and Alignment 19

D.3 The Aligned Manifold Before Imposing Balance 20

D.4 Preservation of Balance for Compatible Optimizer Flows 20

D.5 Why the Alignment Assumption is Appropriate 21

D.6 Proof of Theorem 1 22

E Optimizer Specific Phase Laws **23**

E.1 Birth, Decay, and Saturation times 23

E.2 Birth and Decay Exponents 24

E.3 Saturation Exponent 25

E.4 Euclidean Gradient Descent Phase Laws 26

E.5 SignGD Phase Laws 28

E.6 Instantaneous Adam Phase Laws 29

E.7 Muon Phase Laws 31

E.8 Cumulative Shampoo Phase Laws 32

E.9 K-FAC Phase Laws 36

F Experimental Settings **39**

G Additional Experimental Results **41**

G.1 Matrix Factorization 41
G.1.1 Role of Depth 43
G.1.2 Adam Phase Transitions 46
G.1.3 Shampoo Phase Transitions 46
G.1.4 K-FAC Phase Transitions 47
G.2 Linear Regression 47

A Extended Related Work

Deep linear dynamics and balancedness. Saxe et al. (2013) derive exact learning dynamics for deep linear networks under structured initial conditions, including a decoupled mode representation and equal strengths across layers. Their analysis supplies the direct precedent for our aligned phase model. Arora et al. (2018) show that linear overparameterization changes the effective optimization geometry and can act as a preconditioner. Du et al. (2018) study balancing properties of gradient flow in homogeneous models, while Arora et al. (2019) use balanced initialization to derive effective matrix and singular value dynamics for deep matrix factorization. These GD results motivate our setup, but conservation of the GD balance invariant generally does not hold for other optimizers.

Low-rank bias and the limits of norm descriptions. Gunasekar et al. (2017) investigate implicit regularization in two-layer matrix factorization and conjecture a nuclear-norm characterization in a restricted commuting setting. Arora et al. (2019) extend the analysis to depth and derive singular-value-dependent mobility under balanced gradient flow, showing that depth strengthens the tendency toward low-rank solutions. Razin & Cohen (2020) construct matrix-completion problems in which norms diverge as the loss vanishes, demonstrating the limits of a universal norm-minimization account. Li et al. (2021) instead develop a greedy low-rank learning interpretation under explicit assumptions. A complementary trajectory perspective is provided by Neural Tangent Kernel and lazy-training analyses, which characterize sufficiently wide networks through a linearization around initialization and an approximately fixed kernel governing function-space evolution (Jacot et al., 2018; Lee et al., 2019; Sabanayagam et al., 2022). In contrast, our analysis retains the nonlinear multiplicative parameterization and its evolving, optimizer-dependent spectral mobility, allowing us to compare the birth, saturation, and decay of singular modes before convergence.

Optimization geometry and implicit bias. Wilson et al. (2017) demonstrate that adaptive optimization can select solutions with substantially different generalization from GD even when training fit is good. Gunasekar et al. (2018) characterize implicit bias for several optimization geometries in underdetermined linear regression and separable classification. These results establish the broader principle that optimization geometry affects solution selection. Deep factorization adds a parameter dependent product differential, so the geometry acting on the effective matrix changes with both the factors and optimizer state captured by our general trajectory identity in Proposition 1.

Implicit bias of Adam and related adaptive methods. Cattaneo et al. (2023) use backward error analysis to characterize corrections to Adam and RMSProp dynamics, relating implicit regularization to hyperparameters and training stage. The classification literature builds on the maximum margin characterization of GD for separable data (Soudry et al., 2018). Wang et al. (2021) study limiting parameter directions in homogeneous classification under their adaptive flow assumptions. Zhang et al. (2024) establish a maximum ℓ_∞ margin result for full batch Adam in separable linear classification. Xie & Li (2024) connect AdamW, which includes decoupled weight decay, to ℓ_∞-norm constrained optimization under full-batch assumptions. Vasudeva et al. (2026) analyze a difference in feature learning between Adam and GD in a Gaussian classification model with two layer ReLU networks. These results concern distinct losses, optimizer conventions, or notions of bias. Our squared loss analysis studies finite singular amplitudes and phase times without weight decay, so classification margin results do not directly determine its birth or saturation laws.

Analysis of spectral descent, Shampoo, Muon and K-FAC. Fan et al. (2026) characterize spectral descent and Muon through margin maximization in multiclass linear classification. Chen et al. (2025) connect Muon with decoupled weight decay to spectral-norm-constrained optimization. Eschenhagen et al. (2026) interpret Shampoo as spectral descent adapted to stochasticity and the parameter trajectory. History dependence is therefore an established feature of this geometry. Our contribution is the exact first integral of the specified cumulative accumulator in the aligned factorization model and its consequences for phase times. That identity does not extend unchanged to exponential averages or different inverse powers. Kerekes et al. (2021) explicitly study natural gradient implicit bias in deep linear classification and matrix factorization. Under their Gaussian model and Jacobian rank condition, natural gradient flow for matrix completion matches direct optimization of the effective matrix, leaving unobserved entries unchanged. This is a direct precedent for connecting geometry to motion outside the observed coordinates. Our block curvature analysis

quantifies the effect of additive damping on mode birth and decay, and distinguishes it from both full curvature and factored damping. These works provide important geometry and momentum precedents, while our phase laws concern squared loss, explicit stabilization or damping factors, and finite targets in factorized models.

Analysis of Muon in matrix factorization. Kang et al. (2026) analyze spectral gradient flow in a simplified LoRA factorization setting, proving approximately equal spectral growth rates and convergence results under stated conditions. Dragutinović & Ranganath (2026) derive dynamics for simplified Muon in deep linear networks and linear attention and investigate how avoiding slow sequential learning changes simplicity bias. Rhee et al. (2026) analyze bilinear factorization with full observation, including uniform factor growth after alignment, a conserved quantity distinct from GD's, and finite step oscillations near solutions. They also study alignment in simplified settings. These results already establish much of the ideal polar growth baseline. Our treatment places it alongside memory dependent adaptive flows and damped curvature flows, with separate statements for birth, saturation, and zero target decay. Parviz et al. (2026) report controlled optimizer comparisons showing that Muon's relative performance in factorization depends on the problem and hyperparameter tuning. This motivates separating exact polar flow predictions from practical implementation effects and comparing recovery at a common training fit.

B Optimizers

As we detailed in Sec. 3, we study deterministic flows driven by the full gradient, without weight decay or explicit regularization. For each layer, the continuous time dynamics of the matrix W_ℓ is

$$\dot{W}_\ell(t) = -Z_\ell(t), \qquad W_{\ell,k+1} = W_{\ell,k} - \eta Z_{\ell,k},$$

where t is continuous time, a dot denotes its derivative, k is a discrete iteration index, and $\eta > 0$ is the step size. The second expression is the corresponding factor update once the optimizer states have been specified. The flow uses a unit learning rate, with any constant common learning rate absorbed into time and the stated memory times. All moment matrices have the same dimensions as W_ℓ. The equations for Z_ℓ for all the optimizers analyzed in this paper are discussed in detail below.

Euclidean GD. From our setting, recollect that H_ℓ is the layer gradient $H_\ell = \nabla_{W_\ell}\mathcal{L}(W)$. Gradient descent uses the layer gradient directly, giving

$$Z_\ell = H_\ell, \qquad \dot{W}_\ell = -B_\ell^\top G A_\ell^\top.$$

This gradient flow is the baseline for the established analysis of deep matrix factorization (Arora et al., 2019; Li et al., 2021).

SignGD. SignGD applies sign function entrywise to the layer gradient (Bernstein et al., 2018). Therefore,

$$\dot{W}_\ell = -Z_\ell = -\operatorname{sign}(H_\ell).$$

Adam. Adam maintains entrywise exponential averages of the first moment m_ℓ and second moment v_ℓ (Kingma & Ba, 2014). In discrete time, the decay coefficients $\beta_1, \beta_2 \in [0, 1)$ specify the fraction of the previous moment retained:

$$m_{\ell,k+1} = \beta_1 m_{\ell,k} + (1-\beta_1) H_{\ell,k}, \qquad v_{\ell,k+1} = \beta_2 v_{\ell,k} + (1-\beta_2) H_{\ell,k}^{\odot 2}.$$

Both states start at zero. The continuous dynamics of Adam with its normalization and moment states are established in Barakat & Bianchi (2021); Gould & Tanaka (2024), which are described below.

$$\tau_m \dot{m}_\ell = H_\ell - m_\ell, \qquad \tau_v \dot{v}_\ell = H_\ell^{\odot 2} - v_\ell, \qquad \dot{W}_\ell = -Z_\ell = -m_\ell \oslash (\sqrt{v_\ell} + \epsilon).$$

The constant $\epsilon > 0$ in the denominator is the stabilizer. At $t = 0$, for the uncorrected moment flow, take $m_\ell(0) = 0$ and $v_\ell(0) = 0$. The displayed differential equations describe these uncorrected moments. The phase laws below instead concern the separately specified instantaneous approximation $m_\ell = H_\ell$ and $v_\ell = H_\ell^{\odot 2}$. The finite memory limit scales the discrete decay coefficients with the step size, and is distinct from a limit holding those coefficients fixed.

Muon. Muon orthogonalizes a momentum matrix before applying the update (Jordan et al., 2024). We use exact polar normalization: if $X = U_X \Sigma_X V_X^\top$ is a compact singular value decomposition retaining only positive singular values, define $\mathrm{polar}(X) = U_X V_X^\top$ and $\mathrm{polar}(0) = 0$. We consider an ideal setting without momentum, which reduces to the following update.

$$\dot{W}_\ell = -Z_\ell = -\,\mathrm{polar}(H_\ell),$$

The variant with momentum replaces $\mathrm{polar}(H_\ell)$ with $\mathrm{polar}(m_\ell)$. Approximate polar computation, Nesterov momentum, and shape dependent update scaling are additional choices in practical Muon implementations.

Shampoo. For matrix parameters, the original Shampoo geometry applies inverse fourth roots of left and right gradient statistics (Gupta et al., 2018). Our cumulative flow is

$$\dot{C}_\ell = H_\ell H_\ell^\top, \qquad \dot{D}_\ell = H_\ell^\top H_\ell, \qquad \dot{W}_\ell = -Z_\ell = -C_\ell^{-1/4} H_\ell D_\ell^{-1/4},$$

with $C_\ell(0) = \lambda_{\mathrm{Sh}} I_{d_\ell}$, $D_\ell(0) = \lambda_{\mathrm{Sh}} I_{d_{\ell-1}}$, and $\lambda_{\mathrm{Sh}} > 0$. Here $C_\ell \in \mathbb{R}^{d_\ell \times d_\ell}$ and $D_\ell \in \mathbb{R}^{d_{\ell-1} \times d_{\ell-1}}$ accumulate gradient history, and λ_{Sh} specifies their initial regularization. From the state equations, we get

$$C_\ell(t) = \lambda_{\mathrm{Sh}} I_{d_\ell} + \int_0^t H_\ell(u) H_\ell(u)^\top \, du, \quad \text{and} \quad D_\ell(t) = \lambda_{\mathrm{Sh}} I_{d_{\ell-1}} + \int_0^t H_\ell(u)^\top H_\ell(u) \, du.$$

K-FAC. K-FAC approximates layer curvature by a Kronecker product of activation and backward signal statistics (Martens & Grosse, 2015). First, to get the Kronecker product form of H_ℓ, let's remind the vectorization identity $\mathrm{vec}(AXB) = (B^\top \otimes A)\,\mathrm{vec}(X)$, using which we obtain

$$\mathrm{vec}(H_\ell) = (A_\ell \otimes B_\ell^\top)\,\mathrm{vec}(G) = \mathcal{J}_\ell^\top \,\mathrm{vec}(G).$$

Note that $\mathcal{J}_\ell$ is the Jacobian of the end to end matrix with respect to layer ℓ. For squared loss, the Hessian with respect to the end to end prediction is the identity. Therefore, the generalized Gauss Newton block associated with layer ℓ is

$$\begin{aligned} \mathcal{F}_\ell^{\mathrm{GN}} &= \mathcal{J}_\ell^\top \mathcal{J}_\ell \\ &= (A_\ell \otimes B_\ell^\top)(A_\ell^\top \otimes B_\ell) \\ &= (A_\ell A_\ell^\top) \otimes (B_\ell^\top B_\ell). \end{aligned} \tag{5}$$

The block curvature is therefore already in Kronecker factored form in the full observation deep linear model.

Note that the same block is obtained from the expected Fisher under an isotropic Gaussian observation model. Briefly, let's consider y given x and parameters Θ is isotropic Gaussian, that is, $y \mid x, \Theta \sim \mathcal{N}\left(Wx, \tau^2 I\right)$. Let $h_{\ell-1}(x) = A_\ell x$ denote the input activation to the layer ℓ. The Jacobian of the output with respect to W_ℓ is $J_\ell(x) = h_{\ell-1}(x)^\top \otimes B_\ell$. The expected Fisher block is

$$\begin{aligned} \mathcal{F}_\ell^{\mathrm{F}} &= \frac{1}{\tau^2} \mathbb{E}_x \left[J_\ell(x)^\top J_\ell(x) \right] \\ &= \frac{1}{\tau^2} \mathbb{E}_x \left[h_{\ell-1}(x) h_{\ell-1}(x)^\top \right] \otimes (B_\ell^\top B_\ell) \\ &= \frac{1}{\tau^2} (A_\ell \mathbb{E}[x x^\top] A_\ell^\top) \otimes (B_\ell^\top B_\ell) \quad \text{; whitened full observations } \mathbb{E}[x x^\top] = I \\ &= \frac{1}{\tau^2} (A_\ell A_\ell^\top) \otimes (B_\ell^\top B_\ell). \end{aligned}$$

After absorbing the constant τ^{-2} into the learning rate or damping, $\mathcal{F}_\ell = \mathcal{F}_\ell^{\mathrm{F}} = \mathcal{F}_\ell^{\mathrm{GN}}$. Thus the generalized Gauss Newton block and expected Fisher block coincide under the observation model used here. We use additive damping on the complete per layer curvature block:

$$\mathrm{vec}(Z_\ell) = (\mathcal{F}_\ell + \lambda_{\mathrm{KF}} I)^{-1} \,\mathrm{vec}(H_\ell).$$

The continuous update is $\dot{W}_\ell = -Z_\ell$. When $\lambda_{\mathrm{KF}} = 0$ and the factors are invertible, the Kronecker identity gives

$$Z_\ell = (B_\ell^\top B_\ell)^{-1} H_\ell (A_\ell A_\ell^\top)^{-1}.$$

For positive damping, it is preferable to retain the block form $(\mathcal{F}_\ell + \lambda_{\mathrm{KF}} I)^{-1}$, because additive damping of the full block is not generally equal to damping the two Kronecker factors separately. The undamped case is considered only where the inverse is defined, or with an explicitly specified pseudoinverse. For general matrix sensing, the measurement loss does not automatically yield these regression curvature factors, so a Kronecker approximation must be specified separately.

C Effective Matrix W Dynamics

Consider the setting described in Sec. 3, where the deep linear network is

$$W = W_L W_{L-1} \cdots W_1.$$

Here W is the effective matrix and each $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$. Loss for matrix factorization:

$$\mathcal{L}(W) = \frac{1}{2}\|W - M\|_F^2$$

where M is the matrix whose entries are known and to be factorized. Gradient of the loss $\mathcal{L}(W)$ with respect to W is

$$G = \nabla_W \mathcal{L}(W) = W - M.$$

Layer gradients H_ℓ. We compute the layer gradients H_ℓ, which are the gradients of the loss with respect to layer ℓ. That is $H_\ell = \nabla_{W_\ell} \mathcal{L}(W)$. From the definition of prefix and suffix products, we have

$$A_\ell = W_{\ell-1} \cdots W_1, \qquad A_1 = I_{d_0}$$
$$B_\ell = W_L \cdots W_{\ell+1}, \qquad B_L = I_{d_L}$$

This implies,

$$W = B_\ell W_\ell A_\ell$$

Therefore, applying the chain rule, we get

$$H_\ell = \nabla_{W_\ell} \mathcal{L}(W) = B_\ell^\top G A_\ell^\top$$

Proof of Proposition 1: Effective matrix dynamics.

Proof. Since the effective matrix is the product of all W_ℓ, we derive its continuous time dynamics using the above.

$$\begin{aligned} W &= W_L W_{L-1} \cdots W_1 \\ \dot{W} &= \sum_{\ell=1}^{L} W_L \cdots \dot{W}_\ell W_1 \\ &= \sum_{\ell=1}^{L} B_\ell \dot{W}_\ell A_\ell \end{aligned} \tag{6}$$

$\dot{W}_\ell$ in Eq. (6) changes depending on the optimizer. For each optimizer, we give the updates Z_ℓ in Sec. B and summarize it in Table 1. Thus, eventually $\dot{W}_\ell = -Z_\ell$. □

D Aligned Singular Mode Dynamics

In this section, we prove Theorem 1 by restricting these dynamics to the balanced and aligned mode manifold introduced in the main text. We first explain when balance is preserved, then show how balance and alignment produce a common layer amplitude, and finally derive the effective singular-mode dynamics. Throughout this section, let $d \le \min_{\ell=0,\dots,L} d_\ell$ denote the active mode dimension. For each layer, let there be an orthogonal matrix $Q_\ell \in \mathbb{R}^{d_\ell \times d}$, $Q_\ell^\top Q_\ell = I_d$. All matrix identities below are understood on these active subspaces. Components in the orthogonal complements are initialized at zero and remain absent under the compatible aligned updates considered here.

D.1 Balance under Euclidean Gradient Flow

We first recall why balancedness is an exact invariant of Euclidean gradient flow. For adjacent layers, define the imbalance matrix

$$\mathcal{D}_\ell = W_\ell W_\ell^\top - W_{\ell+1}^\top W_{\ell+1}, \qquad \ell = 1, \dots, L-1.$$

Lemma 1 (Balancedness invariant for Euclidean gradient flow)**.** *Let the loss be any differentiable function of the end-to-end product $W = W_L \cdots W_1$. Under Euclidean gradient flow $\dot{W}_\ell = -H_\ell$, the imbalance matrices satisfy*

$$\frac{d}{dt}\mathcal{D}_\ell = 0.$$

Consequently, $\mathcal{D}_\ell(t) = \mathcal{D}_\ell(0)$ for every time at which the flow exists.

Proof. Fix $\ell \in \{1, \dots, L-1\}$ and define $R_\ell = B_{\ell+1}^\top G A_\ell^\top$ with $G = \nabla_W \mathcal{L}(W)$. Since

$$B_\ell = B_{\ell+1} W_{\ell+1}, \qquad \text{and} \qquad A_{\ell+1} = W_\ell A_\ell,$$

the adjacent layer gradients can be written as

$$H_\ell = W_{\ell+1}^\top R_\ell, \qquad \text{and} \qquad H_{\ell+1} = R_\ell W_\ell^\top.$$

Under Euclidean gradient flow, $\dot{W}_\ell = -H_\ell$, $\dot{W}_{\ell+1} = -H_{\ell+1}$. Therefore,

$$\begin{aligned}\frac{d}{dt}\left(W_\ell W_\ell^\top\right) &= -H_\ell W_\ell^\top - W_\ell H_\ell^\top \\ &= -W_{\ell+1}^\top R_\ell W_\ell^\top - W_\ell R_\ell^\top W_{\ell+1}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{d}{dt}\left(W_{\ell+1}^\top W_{\ell+1}\right) &= -H_{\ell+1}^\top W_{\ell+1} - W_{\ell+1}^\top H_{\ell+1} \\ &= -W_\ell R_\ell^\top W_{\ell+1} - W_{\ell+1}^\top R_\ell W_\ell^\top.\end{aligned}$$

The two derivatives are identical. Hence

$$\frac{d}{dt}\left(W_\ell W_\ell^\top - W_{\ell+1}^\top W_{\ell+1}\right) = 0.$$

□

Lemma 1 holds without alignment and without specializing the loss to full observation. It is an exact property of continuous time Euclidean gradient flow. This conservation is also showed in (Saxe et al., 2013; Gunasekar et al., 2017; Arora et al., 2019). For finite-step gradient descent, the first-order terms still cancel, but an imbalance of order equal to the square of the step size can be introduced at each update.

For non-Euclidean optimizers, the matrix identity $\frac{d}{dt}\mathcal{D}_\ell = 0$ does not hold in general. Balance is instead preserved in our analysis because Assumptions 1 to 3 define an invariant aligned manifold on which all layers have equal mode amplitudes, equal optimizer states, and equal mode updates. We establish this property below.

D.2 Consequences of Balance and Alignment

Under Assumption 2, the target and the layer factors at initialization have the forms

$$M = Q_L \operatorname{diag}(\mu_1, \dots, \mu_d) Q_0^\top, \qquad \mu_i \geq 0, \text{and}$$

$$W_\ell(0) = Q_\ell \operatorname{diag}\left(a_{\ell 1}(0), \dots, a_{\ell d}(0)\right) Q_{\ell-1}^\top, \qquad a_{\ell i}(0) > 0.$$

The corresponding adjacent Gram matrices are

$$W_\ell(0) W_\ell(0)^\top = Q_\ell \operatorname{diag}\left(a_{\ell 1}(0)^2, \dots, a_{\ell d}(0)^2\right) Q_\ell^\top, \qquad \text{and}$$

$$W_{\ell+1}(0)^\top W_{\ell+1}(0) = Q_\ell \operatorname{diag}\left(a_{\ell+1,1}(0)^2, \dots, a_{\ell+1,d}(0)^2\right) Q_\ell^\top.$$

Assumption 1 states that these matrices are equal. Multiplying this equality by $Q_\ell^\top$ on the left and by Q_ℓ on the right gives

$$a_{\ell i}(0)^2 = a_{\ell+1,i}(0)^2$$

for every mode i and every adjacent pair of layers. The positivity assumption removes the sign ambiguity, yielding $a_{\ell i}(0) = a_{\ell+1,i}(0)$. Applying this relation across the network gives a common initial mode amplitude

$$a_{\ell i}(0) = a_i(0) \qquad \text{for every } \ell.$$

Thus Assumptions 1 and 2 together imply more than equal layer norms. They imply equality of the corresponding aligned mode amplitudes across all layers.

D.3 The Aligned Manifold Before Imposing Balance

It is useful first to compute the layer gradients for aligned factors without assuming that their amplitudes are equal. Suppose that at some time t,

$$W_\ell = Q_\ell \operatorname{diag}(a_{\ell 1}, \dots, a_{\ell d})\, Q_{\ell-1}^\top.$$

Then the end-to-end matrix W is

$$W = Q_L \operatorname{diag}(\sigma_1, \dots, \sigma_d)\, Q_0^\top \qquad \text{where} \qquad \sigma_i = \prod_{k=1}^{L} a_{ki}.$$

As a result, the effective gradient is

$$G = W - M = Q_L \operatorname{diag}(r_1, \dots, r_d)\, Q_0^\top, \qquad \text{where} \qquad r_i = \sigma_i - \mu_i.$$

For layer ℓ, define the prefix and suffix mode products

$$\pi_{\ell i}^- = \prod_{k=1}^{\ell-1} a_{ki}, \qquad \text{and} \qquad \pi_{\ell i}^+ = \prod_{k=\ell+1}^{L} a_{ki}.$$

The empty products are equal to one. On the active subspaces,

$$A_\ell = Q_{\ell-1} \operatorname{diag}\left(\pi_{\ell 1}^-, \dots, \pi_{\ell d}^-\right) Q_0^\top, \qquad \text{and} \qquad B_\ell = Q_L \operatorname{diag}\left(\pi_{\ell 1}^+, \dots, \pi_{\ell d}^+\right) Q_\ell^\top.$$

Using the layer gradient formula $H_\ell = B_\ell^\top G A_\ell^\top$, we obtain

$$\begin{aligned} H_\ell &= Q_\ell \operatorname{diag}\left(\pi_{\ell 1}^+, \dots, \pi_{\ell d}^+\right) Q_L^\top Q_L \operatorname{diag}(r_1, \dots, r_d)\, Q_0^\top Q_0 \operatorname{diag}\left(\pi_{\ell 1}^-, \dots, \pi_{\ell d}^-\right) Q_{\ell-1}^\top \\ &= Q_\ell \operatorname{diag}(g_{\ell 1}, \dots, g_{\ell d})\, Q_{\ell-1}^\top, \end{aligned}$$

where

$$g_{\ell i} = r_i \pi_{\ell i}^- \pi_{\ell i}^+ = r_i \prod_{k \neq \ell} a_{ki}.$$

Thus alignment of the factors with the target implies alignment of every layer gradient with the corresponding pair of layer bases. This statement does not yet require balance.

D.4 Preservation of Balance for Compatible Optimizer Flows

We now explain the role of Assumption 3. Suppose that, at some time, the amplitudes of mode i agree across layers: $a_{\ell i} = a_i$ for every ℓ. Then

$$\sigma_i = a_i^L, \qquad \text{and} \qquad g_{\ell i} = r_i a_i^{L-1},$$

which is independent of ℓ. Hence all layers receive the same scalar gradient for mode i. Assumption 3 requires that the optimizer use identical hyperparameters and identical initial mode states in every layer. It also requires that equal mode gradients and equal mode states produce equal mode updates and equal state derivatives. Consequently, whenever $a_{\ell i} = a_i$ and the corresponding optimizer states agree across layers, the update has the form

$$Z_\ell = Q_\ell \operatorname{diag}(z_1, \dots, z_d)\, Q_{\ell-1}^\top,$$

with the same coefficient z_i in every layer. Since $\dot{W}_\ell = -Z_\ell$, we have $\dot{a}_{\ell i} = -z_i$ for every layer. Therefore

$$\frac{d}{dt}\left(a_{\ell i} - a_{\ell+1,i}\right) = 0.$$

Equal amplitudes remain equal, and the aligned factorization remains balanced: $W_\ell W_\ell^\top = W_{\ell+1}^\top W_{\ell+1}$. The same argument applies to the optimizer states. Equal states receive equal state derivatives because the gradients and hyperparameters are equal. Thus equality of both amplitudes and states is preserved by the joint optimizer flow. Equivalently, the set

$$\mathcal{M}_{\mathrm{bal,align}} = \left\{ W_\ell = Q_\ell \operatorname{diag}(a_1, \ldots, a_d) Q_{\ell-1}^\top \text{ for every } \ell \right\},$$

together with equal mode states across layers, is an invariant manifold of every optimizer satisfying Assumption 3. This conclusion holds on intervals where the optimizer vector field is well defined and the relevant amplitudes remain positive. For nonsmooth updates at zero, the continuation is specified separately in the corresponding optimizer result.

Compatibility of the optimizers considered in this work. The compatibility condition is exact for the idealized optimizer flows studied in the main text.

For Euclidean GD, $z_i = g_i$, so equal layer gradients immediately give equal updates.

For coordinatewise SignGD, $z_i = \operatorname{sign}(g_i)$. Because elementwise sign is not invariant under general orthogonal changes of basis, the aligned bases must coincide with the actual parameter coordinates. We therefore take $Q_\ell = I$ for SignGD, up to signed coordinate permutations. The same restriction applies to coordinatewise Adam and to the instantaneous Adam approximation. With $Q_\ell = I$, diagonal gradients preserve diagonal first- and second-moment states. Identical moment initialization and common hyperparameters then give identical states and updates across layers.

For ideal Muon without momentum, the layer gradient in the aligned model $H_\ell = Q_\ell \operatorname{diag}(g_1, \ldots, g_d) Q_{\ell-1}^\top$ gives $\operatorname{polar}(H_\ell) = Q_\ell \big(\operatorname{sign}(g_1), \ldots, \operatorname{sign}(g_d)\big) Q_{\ell-1}^\top$ where $\operatorname{sign}(0) = 0$. Therefore, it preserves equality of the scalar states g_i across layers.

For cumulative Shampoo, let the left and right accumulators be initialized isotropically. An aligned gradient $H_\ell = Q_\ell \operatorname{diag}(g_1, \ldots, g_d) Q_{\ell-1}^\top$ gives

$$H_\ell H_\ell^\top = Q_\ell \operatorname{diag}(g_1^2, \ldots, g_d^2) Q_\ell^\top, \qquad \text{and} \qquad H_\ell^\top H_\ell = Q_{\ell-1} \operatorname{diag}(g_1^2, \ldots, g_d^2) Q_{\ell-1}^\top.$$

The accumulator eigenvalues therefore remain aligned and equal across layers. The preconditioned update consequently has the compatible form $Z_\ell = Q_\ell \operatorname{diag}(z_1, \ldots, z_d) Q_{\ell-1}^\top$.

For the block-curvature model of K-FAC, the curvature eigenvalue of mode i in every layer is $\sigma_i^{2-2/L}$ under balance. Common additive damping therefore produces the same preconditioned scalar update in every layer.

These statements concern the idealized flows specified in the main text. They do not imply that arbitrary practical implementations preserve balance away from the aligned manifold. In particular, coordinatewise methods are generally basis dependent, and practical K-FAC damping applied separately to the two Kronecker factors need not preserve the same layer symmetry.

D.5 Why the Alignment Assumption is Appropriate

Assumption 2 does not claim that arbitrary training trajectories become aligned with the target. It restricts the analysis to a canonical invariant subspace in which exact mode dynamics can be derived. This restriction is standard in analyses of deep linear networks and deep matrix factorization (Saxe et al., 2013; Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2021). It has three useful properties. First, full-observation regression fixes a target basis through the singular directions of M. If the factors and optimizer states are initialized in compatible chain bases, the layer gradients remain in those bases. Second, fixing the same initial spectral configuration for every optimizer ensures that differences in mode evolution arise from the optimizer update and its state, rather than from different initial singular directions. Third, the aligned model retains the nonlinear multiplicative dependence of the effective modes on the factors. It therefore preserves the depth-dependent mobility responsible for the phase laws while removing singular-vector rotation from the comparison.

For spectral and curvature-based optimizers, the aligned subspace is preserved by their matrix-equivariant updates. For coordinatewise methods, the same construction is exact when the aligned basis is the parameter coordinate basis. The general matrix dynamics outside this invariant subspace are given in the previous appendix section.

If the target contains repeated singular values, the singular basis is not unique within the corresponding degenerate subspace. Any fixed orthonormal basis in that subspace may be chosen as the reference basis. The theorem tracks modes in these fixed reference bases and does not require their amplitudes to remain sorted by magnitude.

D.6 Proof of Theorem 1

Proof. By Assumptions 1 and 2, the layer amplitudes agree at initialization: $a_{\ell i}(0) = a_i(0)$. By Assumption 3 and the invariant-manifold argument above, this equality persists throughout training: $a_{\ell i}(t) = a_i(t)$ for every layer ℓ and every time before a mode reaches zero. Hence

$$W_\ell(t) = Q_\ell \operatorname{diag}\left(a_1(t), \ldots, a_d(t)\right) Q_{\ell-1}^\top.$$

Multiplying the factors gives

$$\begin{aligned} W &= W_L \cdots W_1 \\ &= Q_L \operatorname{diag}(a_1, \ldots, a_d) Q_{L-1}^\top Q_{L-1} \operatorname{diag}(a_1, \ldots, a_d) Q_{L-2}^\top \cdots Q_1 \operatorname{diag}(a_1, \ldots, a_d) Q_0^\top \\ &= Q_L \operatorname{diag}\left(a_1^L, \ldots, a_d^L\right) Q_0^\top. \end{aligned}$$

We can rewrite it as

$$W = Q_L \operatorname{diag}\left(\sigma_1, \ldots, \sigma_d\right) Q_0^\top, \qquad \text{where} \qquad \sigma_i = a_i^L.$$

For the full-observation loss,

$$G = W - M = Q_L \operatorname{diag}\left(r_1, \ldots, r_d\right) Q_0^\top, \qquad \text{where} \qquad r_i = \sigma_i - \mu_i.$$

Under the common layer amplitudes,

$$A_\ell = Q_{\ell-1} \operatorname{diag}\left(a_1^{\ell-1}, \ldots, a_d^{\ell-1}\right) Q_0^\top, \qquad \text{and} \qquad B_\ell = Q_L \operatorname{diag}\left(a_1^{L-\ell}, \ldots, a_d^{L-\ell}\right) Q_\ell^\top.$$

Using $H_\ell = B_\ell^\top G A_\ell^\top$, we obtain

$$\begin{aligned} H_\ell &= Q_\ell \operatorname{diag}\left(a_1^{L-\ell}, \ldots, a_d^{L-\ell}\right) \operatorname{diag}\left(r_1, \ldots, r_d\right) \operatorname{diag}\left(a_1^{\ell-1}, \ldots, a_d^{\ell-1}\right) Q_{\ell-1}^\top \\ &= Q_\ell \operatorname{diag}\left(r_1 a_1^{L-1}, \ldots, r_d a_d^{L-1}\right) Q_{\ell-1}^\top. \end{aligned}$$

Let $q = 1 - \frac{1}{L}$, then $a_i^{L-1} = \left(a_i^L\right)^{(L-1)/L} = \sigma_i^q$, and the layer gradient has the form

$$H_\ell = Q_\ell \operatorname{diag}\left(g_1, \ldots, g_d\right) Q_{\ell-1}^\top, \qquad \text{where} \qquad g_i = r_i \sigma_i^q. \tag{7}$$

The coefficient g_i is independent of the layer index ℓ.

By Assumption 3, the optimizer update is $Z_\ell = Q_\ell \operatorname{diag}\left(z_1, \ldots, z_d\right) Q_{\ell-1}^\top$, with the same z_i in every layer. Since $\dot{W}_\ell = -Z_\ell$, comparison of the aligned coefficients gives

$$\dot{a}_i = -z_i.$$

Differentiate $\sigma_i = a_i^L$. This yields

$$\begin{aligned} \dot{\sigma}_i &= L a_i^{L-1} \dot{a}_i \\ &= -L a_i^{L-1} z_i \\ &= -L \sigma_i^q z_i. \end{aligned}$$

Therefore $\dot{\sigma}_i = -L \sigma_i^q z_i$. The same result follows directly from the effective-matrix dynamics derived in Sec. C:

$$\begin{aligned} \dot{W} &= -\sum_{\ell=1}^{L} B_\ell Z_\ell A_\ell \\ &= -\sum_{\ell=1}^{L} Q_L \operatorname{diag}\left(a_1^{L-\ell}, \ldots, a_d^{L-\ell}\right) \operatorname{diag}\left(z_1, \ldots, z_d\right) \operatorname{diag}\left(a_1^{\ell-1}, \ldots, a_d^{\ell-1}\right) Q_0^\top \\ &= -L Q_L \operatorname{diag}\left(a_1^{L-1} z_1, \ldots, a_d^{L-1} z_d\right) Q_0^\top \\ &= Q_L \operatorname{diag}\left(\dot{\sigma}_1, \ldots, \dot{\sigma}_d\right) Q_0^\top. \end{aligned}$$

Finally, define $p = 2q = 2 - \frac{2}{L}$. This exponent appears whenever the optimizer update contains one additional factor of the layer gradient $g_i = r_i \sigma_i^q$. For example, Euclidean GD has $z_i = g_i$, and therefore $\dot{\sigma}_i = -L\sigma_i^q g_i = -L\sigma_i^p(\sigma_i - \mu_i)$. □

E Optimizer Specific Phase Laws

In this section, we prove the phase laws for all the optimizers discussed in Sec. 5. Theorem 1 reduces every compatible optimizer to a scalar evolution for each aligned mode. Suppressing the mode index, the effective singular amplitude satisfies

$$\dot{\sigma} = -L\sigma^q z(\sigma, S(\sigma)), \qquad q = 1 - \frac{1}{L},$$

where z is the optimizer update coefficient and S denotes any associated optimizer state. The state $S(\sigma)$ may include momentum, adaptive moments, accumulated gradient covariances, or curvature estimates. Before deriving the individual optimizer results, we establish a common passage time calculus for birth, decay, and saturation. Each optimizer specific phase law then follows by identifying the local form of z, or equivalently the local power law of $\dot{\sigma}$. This separates the universal passage time argument from the optimizer specific evolution of its internal state.

We use asymptotic notations $o(.)$ and $O(.)$ when taking limits in the asymptotic analysis.

E.1 Birth, Decay, and Saturation times

As discussed in Sec. 5, for a positive target $\mu > 0$, let a mode begin at $\sigma(0) = s > 0$. For a fixed threshold $s < b < \mu$, the birth time is the first passage time

$$T_{\text{opt}}^{\text{birth}}(s, b) = \inf\{t \geq 0 : \sigma(t) = b\}.$$

Birth always refers to escape from a positive seed $s \to 0^+$. It does not refer to spontaneous activation of an exactly zero factor mode. For an unsupported mode with $\mu = 0$, and a tolerance $0 < \delta < s$, the decay time is

$$T_{\text{opt}}^{\text{decay}}(s, \delta) = \inf\{t \geq 0 : \sigma(t) = \delta\}.$$

For a nonzero target, define the residual $r(t) = |\sigma(t) - \mu|$. Given a tolerance $\tau > 0$ and $r(t_0) = r_0$, the saturation time is

$$T_{\text{opt}}^{\text{sat}}(r_0, \tau) = \inf\{u \geq 0 : r(t_0 + v) \leq \tau \text{ for every } v \geq u\}.$$

The requirement that the mode remain inside the tolerance is relevant for optimizers with momentum or relay dynamics, which may cross the target and subsequently leave the tolerance region. For a monotone first order flow, this definition reduces to the first time at which $r(t) = \tau$.

We first state the passage time formula directly in terms of the optimizer update coefficient.

Proposition 2 (Passage times from the common mode dynamics)**.** *Suppose that a positive target mode increases monotonically from s to b and satisfies $z(\sigma, S(\sigma)) < 0$ for $\sigma \in [s, b]$. Then*

$$T_{\text{opt}}^{\text{birth}}(s, b) = -\frac{1}{L}\int_s^b \frac{x^{-q}}{z(x, S(x))}\, dx.$$

Suppose that an unsupported mode decreases monotonically from s to δ and satisfies $z(\sigma, S(\sigma)) > 0$ for $\sigma \in [\delta, s]$. Then

$$T_{\text{opt}}^{\text{decay}}(s, \delta) = \frac{1}{L}\int_\delta^s \frac{x^{-q}}{z(x, S(x))}\, dx.$$

If the residual decreases monotonically and satisfies $\dot{r} = -\Psi(r)$, and $\Psi(r) > 0$, then the time required to reduce the residual from r_0 to τ is

$$T_{\text{opt}}^{\text{sat}}(r_0, \tau) = \int_\tau^{r_0} \frac{dr}{\Psi(r)}.$$

Proof. During birth, $\dot{\sigma} = -L\sigma^q z(\sigma, S) > 0$. Since the trajectory is monotone, the optimizer state can be viewed locally as a function of σ. Separating variables gives

$$dt = -\frac{1}{L}\frac{\sigma^{-q}}{z(\sigma, S(\sigma))}\, d\sigma.$$

Integrating from $\sigma = s$ to $\sigma = b$ proves the birth formula. During decay, $\dot{\sigma} = -L\sigma^q z(\sigma, S) < 0$. Thus

$$dt = -\frac{1}{L}\frac{\sigma^{-q}}{z(\sigma, S(\sigma))}\, d\sigma.$$

Integrating from $\sigma = s$ down to $\sigma = \delta$ and reversing the integral limits proves the decay formula. The saturation formula follows similarly from

$$dt = -\frac{dr}{\Psi(r)}$$

and integration from r_0 to τ. □

For a stateful optimizer, Proposition 2 is exact but need not be closed. The notation $S(x)$ denotes the optimizer state along the particular monotone trajectory. A closed phase law requires either solving the coupled state equations or eliminating the state through an invariant. Cumulative Shampoo admits such an elimination through its accumulator identity. Full Adam and momentum Muon generally retain coupled state dynamics.

E.2 Birth and Decay Exponents

We now classify passage times using the local power of the complete singular mode dynamics. The exponent is defined from $\dot{\sigma}$, rather than from the factor update z alone.

Definition 1 (Birth and decay exponents)**.** *A positive target mode has birth exponent α_{b} and birth coefficient $c_{\mathrm{b}} > 0$ if*

$$\dot{\sigma} = c_{\mathrm{b}}\sigma^{\alpha_{\mathrm{b}}}\,(1 + o(1)) \qquad \text{as } \sigma \to 0^+,$$

where $o(1)$ denotes a quantity tending to zero in the stated limit. An unsupported mode has decay exponent α_{d} and decay coefficient $c_{\mathrm{d}} > 0$ if

$$\dot{\sigma} = -c_{\mathrm{d}}\sigma^{\alpha_{\mathrm{d}}}\,(1 + o(1)) \qquad \text{as } \sigma \to 0^+.$$

The coefficients c_{b} and c_{d} may depend on the target, depth, optimizer parameters, and limiting optimizer state. The exponent records only the leading dependence on the current singular amplitude. A stateful optimizer may have different exponents in different phases. The following elementary calculation determines whether the corresponding passage time remains finite or diverges.

Proposition 3 (Birth and decay times)**.** *Suppose that an increasing mode satisfies $\dot{\sigma} = c\sigma^{\alpha}$, with $c > 0$. Then the time required to move from s to $b > s$ is*

$$T^{\mathrm{birth}}(s, b) = \begin{cases} \dfrac{1}{c}\log\dfrac{b}{s}, & \alpha = 1, \\ \dfrac{b^{1-\alpha} - s^{1-\alpha}}{c(1-\alpha)}, & \alpha \neq 1. \end{cases}$$

Suppose that a decreasing mode satisfies $\dot{\sigma} = -c\sigma^{\alpha}$, with $c > 0$. Then the time required to move from s to $\delta < s$ is

$$T^{\mathrm{decay}}(s, \delta) = \begin{cases} \dfrac{1}{c}\log\dfrac{s}{\delta}, & \alpha = 1, \\ \dfrac{s^{1-\alpha} - \delta^{1-\alpha}}{c(1-\alpha)}, & \alpha \neq 1. \end{cases}$$

Proof. For birth, $dt = \frac{1}{c}\sigma^{-\alpha} d\sigma$. Therefore

$$T^{\mathrm{birth}}(s, b) = \frac{1}{c}\int_s^b x^{-\alpha}\, dx.$$

The integral is logarithmic when $\alpha = 1$ and gives the stated power law when $\alpha \neq 1$. For decay, $dt = -\frac{1}{c}\sigma^{-\alpha}d\sigma$. Integrating from s down to δ gives

$$T^{\text{decay}}(s, \delta) = \frac{1}{c}\int_{\delta}^{s} x^{-\alpha}\, dx,$$

which yields the stated formula. □

The same classifications hold when the dynamics are only asymptotically proportional to a power.

Corollary 1 (Birth time classification). *Suppose $\dot{\sigma} = c_{\text{b}}\sigma^{\alpha_{\text{b}}}(1 + o(1))$ as $\sigma \to 0^+$. For fixed $b > 0$, the birth time satisfies*

$$T^{\text{birth}}(s, b) = \begin{cases} T^{\text{birth}}(0, b) - \dfrac{s^{1-\alpha_{\text{b}}}}{c_{\text{b}}(1-\alpha_{\text{b}})}(1 + o(1)), & \alpha_{\text{b}} < 1, \\ \dfrac{1}{c_{\text{b}}}\log\dfrac{1}{s}(1 + o(1)), & \alpha_{\text{b}} = 1, \\ \dfrac{s^{1-\alpha_{\text{b}}}}{c_{\text{b}}(\alpha_{\text{b}}-1)}(1 + o(1)), & \alpha_{\text{b}} > 1. \end{cases}$$

If $\alpha_{\text{b}} < 1$ gives finite escape from a vanishing positive seed.

If $\alpha_{\text{b}} = 1$ gives a logarithmic birth barrier.

If $\alpha_{\text{b}} > 1$ gives a polynomial birth barrier.

Finite escape as $s \to 0^+$ does not imply spontaneous activation from $\sigma(0) = 0$. If the vector field vanishes at zero and the optimizer takes zero to be absorbing, an exactly zero mode remains zero even when $\alpha_{\text{b}} < 1$. The birth exponent concerns the limit from positive initialization.

Corollary 2 (Decay time classification). *Suppose $\dot{\sigma} = -c_{\text{d}}\sigma^{\alpha_{\text{d}}}(1 + o(1))$ as $\sigma \to 0^+$. For fixed $s > 0$, the decay time satisfies*

$$T^{\text{decay}}(s, \delta) = \begin{cases} T^{\text{decay}}(s, 0) - \dfrac{\delta^{1-\alpha_{\text{d}}}}{c_{\text{d}}(1-\alpha_{\text{d}})}(1 + o(1)), & \alpha_{\text{d}} < 1, \\ \dfrac{1}{c_{\text{d}}}\log\dfrac{1}{\delta}(1 + o(1)), & \alpha_{\text{d}} = 1, \\ \dfrac{\delta^{1-\alpha_{\text{d}}}}{c_{\text{d}}(\alpha_{\text{d}}-1)}(1 + o(1)), & \alpha_{\text{d}} > 1. \end{cases}$$

If $\alpha_{\text{d}} < 1$, the mode reaches zero in finite time under the absorbing continuation.

If $\alpha_{\text{d}} = 1$, the mode decays exponentially: $\sigma(t) = \exp(-c_{\text{d}}t + o(t))$

If $\alpha_{\text{d}} > 1$, the mode decays algebraically: $\sigma(t) \sim (c_{\text{d}}(\alpha_{\text{d}} - 1)t)^{-1/(\alpha_{\text{d}}-1)}$.

In the final two cases, zero is reached only asymptotically.

The terms finite, logarithmic, and polynomial refer to the dependence of the passage time on s or δ. The terms exponential and algebraic refer to the trajectory as a function of training time. For example, $\alpha_{\text{d}} = 1$ gives logarithmic tolerance time and exponential trajectory decay.

E.3 Saturation Exponent

Birth and decay describe behavior near $\sigma = 0$. Saturation instead describes the local approach to a nonzero target μ.

Definition 2 (Saturation exponent). *Let $r(t) = |\sigma(t) - \mu|$. The mode has saturation exponent γ and coefficient $\beta > 0$ if*

$$\dot{r} = -\beta r^{\gamma}(1 + o(1)) \qquad \text{as } r \to 0^+.$$

Proposition 4 (Saturation times). *Suppose that the residual satisfies* $\dot{r} = -\beta r^\gamma, \qquad \beta > 0$. *The time required to reduce the residual from* r_0 *to* $\tau \in (0, r_0)$ *is*

$$T^{\mathrm{sat}}(r_0, \tau) = \begin{cases} \dfrac{1}{\beta} \log \dfrac{r_0}{\tau}, & \gamma = 1, \\ \dfrac{r_0^{1-\gamma} - \tau^{1-\gamma}}{\beta(1-\gamma)}, & \gamma \neq 1. \end{cases}$$

If $\gamma < 1$, *the target is reached in finite time.*

If $\gamma = 0$, *the residual decreases linearly:* $r(t) = (r_0 - \beta t)_+$, *and* β *is the constant residual speed.*

If $\gamma = 1$, *the residual decreases exponentially:* $r(t) = r_0 e^{-\beta t}$, *and* β *is the exponential saturation rate.*

If $\gamma > 1$, *the residual has an algebraic tail:* $r(t) \sim (\beta(\gamma-1)t)^{-1/(\gamma-1)}$.

Proof. Separating variables gives $dt = -\frac{1}{\beta} r^{-\gamma} dr$. Integrating from r_0 down to τ yields

$$T^{\mathrm{sat}}(r_0, \tau) = \frac{1}{\beta} \int_\tau^{r_0} r^{-\gamma}\, dr.$$

Evaluating the integral gives the passage time formula. Solving the same separated equation for $r(t)$ yields the stated trajectory laws. □

For $\gamma = 0$, the coefficient β is a speed rather than an exponential rate. For $\gamma = 1$, it is the usual local exponential rate. For general γ, we refer to β as the saturation coefficient. The scalar saturation exponent is most useful for monotone first order flows. If the optimizer state introduces overshoot or oscillation, the residual may not satisfy a closed first order equation. In that case, saturation must be analyzed through the full state dynamics, and the enter and remain definition distinguishes true saturation from a temporary crossing of the tolerance band.

The optimizer specific proofs that follow therefore have two steps. First, we derive or eliminate the relevant optimizer state to identify z along the phase. Second, we apply Propositions 3 and 4 to obtain the corresponding birth, decay, and saturation times.

E.4 Euclidean Gradient Descent Phase Laws

The update step of Euclidean GD is $Z_\ell = H_\ell$, and so in the aligned model $z_i = g_i = r_i \sigma^q$ from Eq. (7). Therefore, each mode evolves as

$$\dot{\sigma} = L\sigma^p(\mu - \sigma).$$

Note that this is the established GD baseline (Arora et al., 2019).

Proposition 5 (Euclidean GD phase laws). *The birth time* $T^{\mathrm{birth}}_{\mathrm{GD}}(s, b)$ *for* $0 < s < b < \mu$ *with* b *and* μ *fixed is*

$$T^{\mathrm{birth}}_{\mathrm{GD}}(s, b) = \begin{cases} \log\left(\dfrac{\mu - s}{\mu - b}\right), & L = 1, \\ \dfrac{1}{2\mu} \log \dfrac{b(\mu - s)}{s(\mu - b)}, & L = 2, \\ \dfrac{s^{-(L-2)/L} - b^{-(L-2)/L}}{\mu(L-2)} + O(1), & L > 2. \end{cases}$$

For $\mu = 0$, *decay time* $T^{\mathrm{decay}}_{\mathrm{GD}}(s, \delta)$ *with* s *fixed is*

$$T^{\mathrm{decay}}_{\mathrm{GD}}(s, \delta) = \begin{cases} \log \dfrac{s}{\delta}, & L = 1, \\ \dfrac{\delta^{-p} - s^{-p}}{Lp}, & L > 1. \end{cases}$$

Near a nonzero target μ,

$$T_{\mathrm{GD}}^{\mathrm{sat}}(r_0, \delta) = \frac{1}{L\mu^p} \log \frac{r_0}{\delta} + O(1).$$

For nonzero target modes $\mu > 0$, the birth exponent $\alpha_{\mathrm{b}} = p$ and coefficient $c_{\mathrm{b}} = L\mu$ with exponential local saturation $\gamma = 1$ at rate $\beta = L\mu^p$ near target μ. As $s \to 0$, birth time $T_{\mathrm{GD}}^{\mathrm{birth}}(s, b)$ diverges for $L \geq 2$. For zero target modes $\mu = 0$, the decay exponent $\alpha_{\mathrm{d}} = p + 1$ and coefficient $c_{\mathrm{d}} = L$. As $\delta \to 0$, decay time $T_{\mathrm{GD}}^{\mathrm{decay}}(s, \delta)$ diverges for all L.

Proof.

Birth and decay times. From the mode dynamics $\dot{\sigma} = L\sigma^p(\mu - \sigma)$, the exact birth and decay times are

$$T_{\mathrm{GD}}^{\mathrm{birth}}(s, b) = \frac{1}{L} \int_s^b \frac{x^{-p}}{\mu - x}\, dx, \qquad T_{\mathrm{GD}}^{\mathrm{decay}}(s, \delta) = -\frac{1}{L} \int_s^\delta x^{-(p+1)}\, dx.$$

For birth time, especially for $L > 2$, we can decompose it to,

$$\begin{aligned} T_{\mathrm{GD}}^{\mathrm{birth}}(s, b) &= \frac{1}{L} \int_s^b \frac{x^{-p}}{\mu - x}\, dx \\ &= \frac{1}{L} \int_s^b \frac{\mu x^{-p} - x^{1-p} + x^{1-p}}{\mu(\mu - x)}\, dx \\ &= \frac{1}{L\mu} \int_s^b x^{-p}\, dx + \frac{1}{L\mu} \int_s^b \frac{x^{1-p}}{\mu - x}\, dx \\ &= \frac{1}{L\mu} \int_s^b x^{-p}\, dx + O(1) \end{aligned}$$

Integrating the birth and decay times using Proposition 3 give the times in Proposition 5. From the dynamics and Corollaries 1 and 2, we get the birth and decay exponents: $\alpha_{\mathrm{b}} = p$ with $c_{\mathrm{b}} = L\mu$ and $\alpha_{\mathrm{d}} = p + 1$ with $c_{\mathrm{d}} = L$.

From birth time $T_{\mathrm{GD}}^{\mathrm{birth}}$ and Corollary 1, as $s \to 0^+$, $L = 1$ has no initialization-dependent birth barrier and escapes in finite time as $p = 2 - 2/L = 0$. Thus the mode can escape from an arbitrarily small initialization in a uniformly bounded amount of time. In the corresponding birth approximation, $\dot{\sigma} \approx \mu_i$, so the mode grows linearly: $\sigma(t) \approx \epsilon + \mu_i t$. On the other hand, $L = 2$ introduces a logarithmic delay as $p = 1$. The corresponding mode trajectory is exponential: $\dot{\sigma} \approx 2\mu\sigma$ and hence $\sigma(t) \approx \epsilon e^{2\mu_i t}$, therefore the singular mode itself grows exponentially in time. Similarly, $L > 2$ produces a polynomial delay as $p > 1$.

As $\delta \to 0^+$, $T_{\mathrm{GD}}^{\mathrm{decay}}$ diverges from Corollary 2, since $\alpha_{\mathrm{d}} = p + 1 \geq 1$. Equality holds for $L = 1$, while $\alpha_d > 1$ for $L > 1$. Thus GD does not kill zero-target singular values in finite time.
For $L = 1$, $\alpha_{\mathrm{d}} = 1$, therefore the corresponding $\sigma(t)$ trajectory is exponential and the decay time is logarithmic in $1/\delta$. The state decays exponentially in time. It never reaches zero at a finite time.

For $L > 1$, it is a polynomial divergence in $1/\tau$. The exact trajectory is $\sigma(t) = \left(\sigma_0^{-p} + Lp\, t\right)^{-1/p}$. Therefore, at late times, $\sigma(t) \sim (Lp\, t)^{-1/p}$. So decay is algebraic rather than exponential.

Saturation near target. Suppose $\sigma(t) \approx \mu > 0$. Let $r(t) = \mu - \sigma(t)$, therefore $\dot{r} = -\dot{\sigma}$. Then, from the dynamics

$$\dot{r} = -L\sigma^p r.$$

For approach to a positive target from below and above, respectively,

$$T_{\mathrm{GD},-}^{\mathrm{sat}}(r_0, \delta) = \frac{1}{L} \int_\delta^{r_0} \frac{dr}{(\mu - r)^p r}, \qquad \text{and} \qquad T_{\mathrm{GD},+}^{\mathrm{sat}}(r_0, \delta) = \frac{1}{L} \int_\delta^{r_0} \frac{dr}{(\mu + r)^p r}.$$

The first formula requires $0 < \delta < r_0 < \mu$. For fixed r_0,

$$T_{\mathrm{GD},\pm}^{\mathrm{sat}}(r_0, \delta) = \frac{1}{L\mu^p} \log \frac{r_0}{\delta} + O(1) \qquad \text{as } \delta \to 0.$$

Thus, the convergence is exponential with rate $L\mu^p$ from Proposition 4. The larger μ gives a larger local saturation rate only for $L > 1$. At $L = 1$, the rate is one for every target. The exponential saturation implies that modes with smaller target singular values also converge slowly.

□

Role of depth. Depth strongly suppresses the birth of new modes, strengthening the initialization barrier. For $L > 2$, the divergence of birth time is polynomial. Thus an infinitesimally initialized singular mode takes polynomially longer to become macroscopically active. This is the primary mechanism behind sequential low-rank learning.

Depth does not rapidly eliminate existing modes. Decay time diverges for all depth. Indeed: $L = 1$ gives an exponential tail; $L > 1$ gives the slower polynomial tail. Thus Euclidean GD does not produce exact finite-time rank reduction. Its low-rank bias comes mainly from preventing weak modes from being born, rather than from rapidly pruning modes that are already active.

E.5 SignGD Phase Laws

SignGD updates each factor using $Z_\ell = \operatorname{sign}(H_\ell)$. Therefore, in the balanced aligned model, this gives $z_i = \operatorname{sign}(g_i) = \operatorname{sign}(r_i\sigma^q) = \operatorname{sign}(r_i)$ since σ is positive. Therefore, for each mode

$$\dot{\sigma} = -L\sigma^q \operatorname{sign}(\sigma - \mu).$$

Thus, each factor amplitude moves toward its target at unit speed, independently of the gradient magnitude.

Proposition 6 (SignGD phase laws)**.** *The birth time $T^{\text{birth}}_{\text{SignGD}}(s, b)$ for $0 < s < b < \mu$, and decay time $T^{\text{decay}}_{\text{SignGD}}(s, \delta)$ for $\mu = 0$ are*

$$T^{\text{birth}}_{\text{SignGD}}(s, b) = b^{1/L} - s^{1/L}, \qquad T^{\text{decay}}_{\text{SignGD}}(s, \delta) = s^{1/L} - \delta^{1/L}.$$

If a mode approaches a nonzero target from above and below and its residual decreases from r_0 to δ, then

$$T^{\text{sat}}_{\text{SignGD},-}(r_0, \delta) = (\mu-\delta)^{1/L} - (\mu-r_0)^{1/L} \quad \textit{and} \quad T^{\text{sat}}_{\text{SignGD},+}(r_0, \delta) = (\mu+r_0)^{1/L} - (\mu+\delta)^{1/L}.$$

In particular, $T^{\text{sat}}_{\text{SignGD},\pm}(r_0, 0) < \infty$. The same finite-time conclusion holds when the target is approached from above. The birth and decay exponents are $\alpha_{\text{b}} = \alpha_{\text{d}} = q$ and coefficients $c_{\text{b}} = c_{\text{d}} = L$. The saturation is linear $\gamma = 0$, with rate $\beta = L\mu^q$.

Proof.

Birth, saturation and decay times. From the dynamics, $\dot{\sigma} = -L\sigma^q \operatorname{sign}(\sigma - \mu)$, when $0 < s < b < \mu$, $\operatorname{sign}(\sigma - \mu) = -1$. Therefore, $\dot{\sigma} = L\sigma^q$. When $\mu = 0$, $\operatorname{sign}(\sigma - \mu) = 1$ since we consider infinitesimal positive initialization. Therefore, $\dot{\sigma} = -L\sigma^q$. Thus

$$T^{\text{birth}}_{\text{SignGD}}(s, b) = \frac{1}{L}\int_s^b x^{-q}\, dx \qquad T^{\text{decay}}_{\text{SignGD}}(s, \delta) = -\frac{1}{L}\int_s^\delta x^{-q}\, dx.$$

Integrating these give the times in Proposition 6. From the dynamics and Corollaries 1 and 2, we get the birth and decay exponents: $\alpha_{\text{b}} = q$ with $c_{\text{b}} = L$ and $\alpha_{\text{d}} = q$ with $c_{\text{d}} = L$. Note that the coefficients are target independent.

Since the exponent $q < 1$ and the time to reach any target μ is $|s^{1/L} - \mu^{1/L}|$, birth, saturation and decay occur in *finite* time. For $\mu = 0$, the mode becomes exactly zero at time $s^{1/L}$. The birth time remains bounded as $s \to 0$. Therefore, SignGD weakens the birth barrier for nonzero modes, but it can be more aggressively sparsifying for modes whose target is exactly zero. It still slows small singular values through $\sigma^{1-1/L}$ but much less than GD.

The exact saturation times from below and above are $\dfrac{1}{L}\displaystyle\int_\delta^{r_0} \frac{dr}{(\mu \pm r)^q}$.

$$T^{\text{sat}}_{\text{SignGD},-}(r_0, \delta) = (\mu-\delta)^{1/L} - (\mu-r_0)^{1/L} \quad \text{and} \quad T^{\text{sat}}_{\text{SignGD},+}(r_0, \delta) = (\mu+r_0)^{1/L} - (\mu+\delta)^{1/L}.$$

Locally, $\dot{r} = -L\mu^q + O(r)$. Thus $\gamma = 0$ and $\beta = L\mu^q$. For $L > 1$, the residual is only asymptotically linear near the hitting time from Proposition 4.

□

E.6 Instantaneous Adam Phase Laws

Adam updates are considered under a quasi-steady approximation and a diagonal aligned model to isolate the effect of coordinate normalization, with $m = g$ and $v = g^2$ giving the instantaneous state approximation where

$$\dot{\sigma} = -\frac{L\sigma^p(\sigma - \mu)}{|\sigma - \mu|\sigma^q + \epsilon}$$

for each mode. Note that, if $|\sigma - \mu|\sigma^q \gg \epsilon$ then $\dot{\sigma} = -L\sigma^q \operatorname{sign}(\sigma - \mu)$, thus behaving like SignGD. If $|\sigma - \mu|\sigma^q \ll \epsilon$ then $\dot{\sigma} = -L\sigma^p(\sigma - \mu)/\epsilon$, thus behaving like GD upto $1/\epsilon$ factor.

Proposition 7 (Instantaneous Adam phase laws)**.** *Let $0 < s < b < \mu$. The birth time is*

$$T^{\text{birth}}_{\text{IAdam}}(s, b) = b^{1/L} - s^{1/L} + \epsilon\, T^{\text{birth}}_{\text{GD}}(s, b).$$

Let $h_\mu(\sigma) = (\mu - \sigma)\sigma^q$, when $\epsilon < \mu^{q+1}\dfrac{q^q}{(q+1)^{q+1}}$ and $h_\mu(\sigma)/\epsilon \to 0$, Adam behaves GD-like, and $\epsilon/h_\mu(\sigma) \to 0$ it behaves like SignGD-like.

For an unsupported mode with $\mu = 0$, the decay time from s to $\delta \in (0, s)$ is

$$T^{\text{decay}}_{\text{IAdam}}(s, \delta) = s^{1/L} - \delta^{1/L} + \epsilon\, T^{\text{decay}}_{\text{GD}}(s, \delta).$$

Let $\sigma^{\text{d}}_{\text{A}} \asymp \epsilon^{1/(q+1)}$, the final decay satisfies $\sigma(t) < \sigma^{\text{d}}_{\text{A}}$ and behaves like GD.

Suppose a nonzero mode approaches its target from below and its residual decreases from r_0 to δ, where $0 < \delta < r_0 < \mu$. Then

$$T^{\text{sat}}_{\text{IAdam}}(r_0, \delta) = (\mu - \delta)^{1/L} - (\mu - r_0)^{1/L} + \frac{\epsilon}{L}\int_{\mu - r_0}^{\mu - \delta} \frac{x^{-p}}{\mu - x}\, dx.$$

An analogous expression holds when the mode approaches the target from above. Near saturation, it behaves like GD with rate $L\mu^p/\epsilon$.

Proof.

Birth of a positive mode. From the dynamics, $\dot{\sigma} = -\dfrac{L\sigma^p(\sigma - \mu)}{|\sigma - \mu|\sigma^q + \epsilon}$, the birth time for $0 < s < b < \mu$ is

$$\begin{aligned} T^{\text{birth}}_{\text{IAdam}}(s, b) &= \frac{1}{L}\int_s^b x^{-q}\, dx + \frac{\epsilon}{L}\int_s^b \frac{x^{-p}}{\mu - x}\, dx \\ &= T^{\text{birth}}_{\text{SignGD}}(s, b) + \epsilon T^{\text{birth}}_{\text{GD}}(s, b) \end{aligned}$$

Therefore, birth time is the sum of a finite SignGD contribution and an ϵ-dependent Euclidean contribution. For fixed $\epsilon > 0$, b, and μ, as $s \to 0^+$,

$$T^{\text{birth}}_{\text{IAdam}}(s, b) = \begin{cases} \Theta(1), & L = 1, \\ \dfrac{\epsilon}{2\mu}\log\dfrac{1}{s} + O(1), & L = 2, \\ \dfrac{\epsilon}{\mu(L-2)}s^{-(L-2)/L} + O(1), & L > 2. \end{cases}$$

Thus a fixed positive ϵ restores the Euclidean GD birth classification in the extreme small-initialization regime: birth remains finite for $L = 1$, is logarithmically delayed for $L = 2$, and is polynomially delayed for $L > 2$.

For $L > 1$, for $0 < \sigma < \mu$ define

$$h_\mu(\sigma) = (\mu - \sigma)\sigma^q, \qquad h_{\max} = \mu^{q+1}\frac{q^q}{(q+1)^{q+1}}.$$

If $0 < \epsilon < h_{\max}$, the equation $h_\mu(\sigma) = \epsilon$ has two roots

$$0 < \sigma_- < \frac{q\mu}{q+1} < \sigma_+ < \mu.$$

Then $h_\mu(\sigma) > \epsilon$ exactly on (σ_-, σ_+). An accurate approximation with SignGD requires the stronger condition $h_\mu(\sigma) \gg \epsilon$.

If $\epsilon \geq h_{\max}$, no interval satisfies $h_\mu(\sigma) > \epsilon$. If $\epsilon \gg h_{\max}$, the entire birth trajectory is uniformly approximated by GD divided by ϵ.

As $\epsilon/\mu^{q+1} \to 0$, $\sigma_- = (\epsilon/\mu)^{1/q}(1 + o(1))$, $\mu - \sigma_+ = (\epsilon/\mu^q)(1 + o(1))$.

For $L = 1$, $h_\mu(\sigma) = \mu - \sigma$, so the two root description does not apply.

The early-growth crossover is approximately

$$\sigma_{\mathrm{A}}^{\mathrm{b}} \asymp \left(\frac{\epsilon}{\mu}\right)^{1/q},$$

while the crossover near saturation satisfies $\mu - \sigma_{\mathrm{A}}^{\mathrm{s}} \asymp \dfrac{\epsilon}{\mu^q}$.

Modes far below $\sigma_{\mathrm{A}}^{\mathrm{b}}$ initially follow the Euclidean birth law. Modes between the two crossovers move approximately according to SignGD. Modes sufficiently close to the target return to an ϵ-dominated regime.

Decay of an unsupported mode. For $\mu = 0$, the mode dynamics equation becomes $\dot\sigma = -L\dfrac{\sigma^{p+1}}{\sigma^{q+1} + \epsilon}$. When $\sigma^{q+1} \gg \epsilon$, the mode follows the SignGD-like law $\dot\sigma \approx -L\sigma^q$. When $\sigma^{q+1} \ll \epsilon$, the final decay satisfies $\dot\sigma \approx -\dfrac{L}{\epsilon}\sigma^{p+1}$. The crossover scale is therefore

$$\sigma_{\mathrm{A}}^{\mathrm{d}} \asymp \epsilon^{1/(q+1)}.$$

For $\mu = 0$, similar to birth time,

$$T^{\mathrm{decay}}_{\mathrm{IAdam}}(s, \delta) = T^{\mathrm{decay}}_{\mathrm{SignGD}}(s, \delta) + \epsilon T^{\mathrm{decay}}_{\mathrm{GD}}(s, \delta).$$

Therefore,

$$T^{\mathrm{decay}}_{\mathrm{IAdam}}(s, \delta) = \begin{cases} s - \delta + \epsilon \log \dfrac{s}{\delta}, & L = 1, \\ s^{1/L} - \delta^{1/L} + \dfrac{\epsilon}{Lp}\left(\delta^{-p} - s^{-p}\right), & L > 1. \end{cases}$$

For fixed $\epsilon > 0$ and fixed initial value s, as $\delta \to 0^+$,

$$T^{\mathrm{decay}}_{\mathrm{IAdam}}(s, \delta) = \begin{cases} \epsilon \log \dfrac{1}{\delta} + O(1), & L = 1, \\ \dfrac{\epsilon}{Lp}\delta^{-p} + O(1), & L > 1. \end{cases}$$

Thus instantaneous Adam does not remove an unsupported mode in finite time when $\epsilon > 0$. For $L = 1$, the final trajectory decays exponentially: $\sigma(t) \asymp e^{-t/\epsilon}$. For $L > 1$, it has the algebraic tail $\sigma(t) \asymp \left(\frac{Lp}{\epsilon}t\right)^{-1/p}$. The SignGD finite-time decay is therefore replaced by a logarithmic tolerance time at depth one and a polynomial tolerance time at greater depth.

Saturation at a nonzero target. Following the above analysis, for approach to a positive target from either side,

$$T^{\mathrm{sat}}_{\mathrm{IAdam},\pm}(r_0,\delta) = T^{\mathrm{sat}}_{\mathrm{Sign},\pm}(r_0,\delta) + \epsilon T^{\mathrm{sat}}_{\mathrm{GD},\pm}(r_0,\delta).$$

These identities are exact. For fixed r_0,

$$T^{\mathrm{sat}}_{\mathrm{IAdam},\pm}(r_0,\delta) = \frac{\epsilon}{L\mu^p}\log\frac{r_0}{\delta} + O(1).$$

Thus the outer approach is approximately linear, but the final approach is exponential with rate $\beta = \frac{L\mu^p}{\epsilon}$. Unlike ideal SignGD, instantaneous Adam reaches a nonzero target only asymptotically when $\epsilon > 0$.

□

Depth and the order of limits. The two exponents q and p play different roles. The exponent q controls the SignGD-like motion in the factor-amplitude coordinate and the locations of the crossover scales. The exponent p controls the ϵ-dominated birth barrier, the final decay tail, and the local exponential saturation rate.

For fixed $\epsilon > 0$, increasing depth changes the birth asymptotics from logarithmic at $L = 2$ to polynomial at $L > 2$, with exponent $p - 1 = \frac{L-2}{L}$. The unsupported-mode decay tail has exponent $p = 2 - \frac{2}{L}$, which also increases with depth. Thus deeper factorizations make the extreme small-gradient regimes increasingly resistant both to the emergence of infinitesimal positive modes and to the complete removal of already active modes.

For $L \geq 2$, the limits $\epsilon \to 0$ and $s \to 0$ do not commute. For every fixed $s > 0$,

$$\lim_{\epsilon\to 0} T^{\mathrm{birth}}_{\mathrm{IAdam}}(s,b) = T^{\mathrm{birth}}_{\mathrm{SignGD}}(s,b).$$

In contrast, for every fixed $\epsilon > 0$, the limit $s \to 0$ reveals the Euclidean GD birth barrier. Therefore an Adam-like update may appear SignGD-like at practical mode scales while remaining mathematically GD-like in the infinitesimal limit. For $L = 1$, both iterated limits equal b.

Takeaway. *Ideal SignGD makes $\sigma^{1/L}$ a constant-speed coordinate: positive modes are born, unsupported modes disappear, and nonzero targets are reached in finite time for every depth. The instantaneous Adam approximation inherits this behavior only over the range in which the gradient magnitude dominates ϵ. For fixed $\epsilon > 0$, sufficiently small birth modes recover GD's finite–logarithmic–polynomial depth transition, unsupported modes develop logarithmic or polynomial decay tails, and final saturation becomes exponential. Thus coordinate normalization controls the phase crossovers.*

E.7 Muon Phase Laws

Muon updates each factor by orthogonalizing matrix updates, that is $Z_\ell = U_\ell V_\ell^\top$. In the diagonal aligned singular mode model, $G = \mathrm{diag}(g_i)$. Therefore, $z_i = \mathrm{sign}(g_i)$. Thus, for each mode

$$\begin{aligned}\dot\sigma &= -L\sigma^q \,\mathrm{sign}(g)\\ &= -L\sigma^q \,\mathrm{sign}((\sigma-\mu)\sigma^q)\\ &= -L\sigma^q \,\mathrm{sign}(\sigma-\mu)\end{aligned}$$

reducing it to a scalar sign rule as in SignGD updates. Therefore, Proposition 6 holds for Muon as well.

Note that this is a spectral update in matrix space even though the aligned model reduces it to a scalar sign rule as in SignGD updates, and this does not mean Muon has the same global bias as coordinate SignGD. Coordinate SignGD is coordinate-basis dependent. Muon is matrix-basis / spectral-geometry dependent. But in an aligned diagonal model, both reduce to sign dynamics on singular modes. Therefore, Muon weakens low rank bias in the diagonal mode. It controls spectral geometry, not rank directly. A spectral-norm ball constrains the largest singular value, but it does not penalize the number of nonzero singular values. So Muon should not be expected to behave like nuclear-norm minimization.

E.8 Cumulative Shampoo Phase Laws

Shampoo updates each factor using $Z_\ell = C_\ell^{-1/4} H_\ell D_\ell^{-1/4}$ where C_ℓ and D_ℓ are the accumulated left and right gradient-covariance matrices defined in Table 1. In the balanced aligned model, the two covariance factors reduce to the same scalar accumulator. In the aligned diagonal model, $H = \mathrm{diag}(g_i)$, therefore $HH^\top = H^\top H = \mathrm{diag}(g_i^2)$. For the common scalar accumulator, let

$$h(t) = \int_0^t g(u)^2\, du, \qquad \text{and} \qquad d(t) = \sqrt{\lambda_{\mathrm{Sh}} + h(t)},\ d(0) = \sqrt{\lambda_{\mathrm{Sh}}}.$$

For a fresh isotropic Shampoo state, $d_0 = d(0) = \sqrt{\lambda_{\mathrm{Sh}}}$, although the analysis below also permits d_0 to include gradient history accumulated before the phase under consideration. The scalar factor update $C_\ell^{-1/4} H_\ell D_\ell^{-1/4} = \mathrm{diag}(g_i/d)$. Thus

$$\dot{a} = -\frac{g}{d}.$$

Since $\dot{\sigma} = L\sigma^q \dot{a}$, each singular mode satisfies

$$\dot{\sigma} = -L\frac{(\sigma-\mu)\sigma^p}{d} = L\frac{(\mu-\sigma)\sigma^p}{d}.$$

The accumulator obeys

$$\dot{h} = g^2 = (\sigma-\mu)^2\sigma^p.$$

Therefore

$$\dot{d} = \frac{\dot{h}}{2d} = \frac{(\sigma-\mu)^2\sigma^p}{2d}.$$

An exact first integral. Along any nonstationary monotone segment, divide $\dot{d}$ by $\dot{\sigma}$:

$$\frac{dd}{d\sigma} = \frac{(\sigma-\mu)^2\sigma^p/(2d)}{-L(\sigma-\mu)\sigma^p/d} = -\frac{\sigma-\mu}{2L}.$$

Integrating from an initial effective singular value $\sigma(0) = s$ gives

$$d(\sigma) = d_0 + \frac{\mu(\sigma - s) - \frac{1}{2}(\sigma^2 - s^2)}{2L}. \tag{8}$$

This relation exactly eliminates the accumulator from the aligned scalar dynamics. In particular, cumulative Shampoo is not simply an instantaneous approximation to SignGD or Muon: its current mobility is determined by the distance already travelled by the mode.

Proposition 8 (Cumulative Shampoo phase laws)**.** *Let a nonstationary aligned mode begin at $\sigma(0) = s > 0$ with denominator $d_0 > 0$. Consider $d(\sigma)$ in Eq. (8). For $0 < s < b < \mu$, the birth time is exactly*

$$T_{\mathrm{Sh}}^{\mathrm{birth}}(s, b) = \frac{1}{L}\int_s^b \frac{d(x)}{x^p(\mu - x)}\, dx.$$

For an unsupported mode with $\mu = 0$, define $c_s = d_0 + \dfrac{s^2}{4L}$. The decay time from s to $\delta \in (0, s)$ is

$$T_{\mathrm{Sh}}^{\mathrm{decay}}(s, \delta) = \begin{cases} c_s \log\dfrac{s}{\delta} - \dfrac{s^2 - \delta^2}{8} & L = 1, \\ \dfrac{c_s}{Lp}\left(\delta^{-p} - s^{-p}\right) - \dfrac{s^{2-p} - \delta^{2-p}}{4L^2(2-p)} & L > 1. \end{cases}$$

For fixed $d_0 > 0$, the birth asymptotics from a vanishing positive seed have exponent $\alpha_{\mathrm{b}} = p$ and coefficient $c_{\mathrm{b}} = L\mu/d_0$. On an intermediate interval $\sigma_a \le \sigma \le b$ satisfying $s \ll \sigma_a$, $b \ll \mu$, and $2Ld_0 \ll \mu\sigma_a$, the leading birth law instead has exponent $\alpha_{\mathrm{b}} = p - 1$ and coefficient $c_{\mathrm{b}} = 2L^2$. The uniform approximation and its error bound are stated below. For an unsupported mode with fixed $s > 0$ and $d_0 > 0$, the final decay as $\sigma \to 0$ has exponent $\alpha_{\mathrm{d}} = p + 1$ and coefficient $c_{\mathrm{d}} = L/(d_0 + s^2/(4L))$.

For a nonzero target approached monotonically from above and below, let the residual move from r_0 to δ. Then local saturation is exponential at rate $\beta = \dfrac{L\mu^p}{d_\star}$ with

$$T^{\mathrm{sat}}_{\mathrm{Sh},\pm}(r_0,\delta) = \frac{d_\star}{L\mu^p}\log\frac{r_0}{\delta} + O(1), \qquad \textit{where} \qquad d_\star = d(\mu) = d_0 + \frac{(\mu-s)^2}{4L}.$$

Proof.

Birth of a positive mode. The exact first integral $d(\sigma)$ reveals two distinct birth regimes. For fixed positive damping, every mode initially experiences a GD-like phase. After sufficient gradient history has accumulated, the dynamics can enter a history-adapted phase with a different singular-value exponent.

Initial damping dominated phase. Let's define

$$\nu_s(x) = \frac{(x-s)(2\mu-x-s)}{4L}.$$

The exact birth time decomposition is

$$T^{\mathrm{birth}}_{\mathrm{Sh}}(s,b) = d_0 T^{\mathrm{birth}}_{\mathrm{GD}}(s,b) + R_s(b), \qquad \text{where} \qquad R_s(b) = \frac{1}{L}\int_s^b \frac{\nu_s(x)}{x^p(\mu-x)}\,dx.$$

Since $0 \le \nu_s(x) \le \mu x/(2L)$ and $p < 2$, $0 \le R_s(b) \le \dfrac{\mu}{2L^2}\displaystyle\int_0^b \frac{x^{1-p}}{\mu-x}\,dx < \infty$. Thus, for fixed $d_0 > 0$ and $b < \mu$,

$$T^{\mathrm{birth}}_{\mathrm{Sh}}(s,b) = \begin{cases} T^{\mathrm{birth}}_{\mathrm{Sh}}(0,b) + o(1), & L = 1, \\ \dfrac{d_0}{2\mu}\log(1/s) + O(1), & L = 2, \\ \dfrac{d_0 s^{1-p}}{L\mu(p-1)} + O(1), & L > 2. \end{cases}$$

The depth one limit is

$$T^{\mathrm{birth}}_{\mathrm{Sh}}(0,b) = \left(d_0 + \frac{\mu^2}{4}\right)\log\frac{\mu}{\mu-b} - \frac{\mu b}{4} + \frac{b^2}{8}.$$

Thus the initial Shampoo birth exponent $\alpha_{\mathrm{b}} = p$, which is exactly the Euclidean GD exponent and its coefficient $c_{\mathrm{b}} = L\mu/d_0$. Therefore, cumulative Shampoo with fixed positive damping retains the Euclidean GD birth classification: the birth time is finite at depth one, logarithmic at depth two, and polynomial for $L > 2$. This conclusion is important because the later adapted phase does not remove the infinitesimal barrier created by the initial denominator. The adapted phase changes how a mode grows after sufficient history has accumulated, but a mode starting arbitrarily close to zero must first traverse the damping-dominated GD-like regime.

Crossover to the history adapted phase. The accumulator becomes comparable to its initial value when

$$d_0 \asymp \frac{\mu(\sigma-s) - \frac{1}{2}(\sigma^2-s^2)}{2L}.$$

An interior point satisfying $\nu_s(\sigma_c) = d_0$ exists if and only if $4Ld_0 < (\mu-s)^2$. It is $\sigma^{\mathrm{b}}_{\mathrm{Sh}} = \mu - \sqrt{(\mu-s)^2 - 4Ld_0}$. If $4Ld_0/(\mu-s)^2 \to 0$, then

$$\sigma^{\mathrm{b}}_{\mathrm{Sh}} - s = \frac{2Ld_0}{\mu-s}\left(1 + O\left(\frac{Ld_0}{(\mu-s)^2}\right)\right).$$

The approximation $\sigma^{\mathrm{b}}_{\mathrm{Sh}} \approx s + 2Ld_0/\mu$ additionally requires $s \ll \mu$.

A substantial history adapted birth interval exists only when this crossover occurs well before the target. Equivalently, the total history accumulated by the target must be large relative to the initial denominator, that is, $d_0 \ll \frac{(\mu - s)^2}{4L}$. If this condition fails, the mode remains predominantly in the GD-like regime throughout its approach to μ.

At fixed d_0 and μ, the crossover scale increases with depth. Thus deeper factorizations require a larger effective singular value before accumulated Shampoo history dominates the initial damping.

History-adapted accumulator phase. Consider an interval $\sigma_a \le \sigma \le b < \mu$ with $s < \sigma_a$. Define

$$\mathcal{E} = \frac{s}{\sigma_a} + \frac{b}{\mu} + \frac{2Ld_0}{\mu\sigma_a}.$$

For fixed L and $\mathcal{E} \to 0$, uniformly on this interval, $d(\sigma) = \frac{\mu\sigma}{2L}(1 + O(\mathcal{E}))$, that is accumulated history is dominating. Then,

$$\dot{\sigma} = 2L^2\sigma^{p-1}(1 + O(\mathcal{E})).$$

Hence the adapted Shampoo birth exponent in this phase $\alpha_{\mathrm{b}} = p - 1 = 1 - \frac{2}{L}$.

If the adapted phase begins at σ_a and ends at $b \ll \mu$, its duration is

$$T_{\mathrm{Sh}}^{\mathrm{birth}}(\sigma_a, b) = \frac{1 + O(\mathcal{E})}{2L^2} \int_{\sigma_a}^{b} x^{-(p-1)}\, dx = \frac{b^{2/L} - \sigma_a^{2/L}}{4L}\,(1 + O(\mathcal{E}))$$

from Proposition 3 and corollary 1. This intermediate approximation cannot be extended to $\sigma_a = 0$ at fixed $d_0 > 0$. The target amplitude μ has disappeared from this law. Thus, after enough gradient history has accumulated, Shampoo no longer orders growing modes by their target amplitudes.

The initial and adapted Shampoo exponents lie on opposite sides of Muon's exponent $q = 1 - 1/L$. For $L > 1$, $p > q$, so initial Shampoo suppresses small modes more strongly than Muon. In contrast, $p - 1 = q - 1/L < q$, so the adapted phase has weaker singular magnitude dependence than Muon. Cumulative Shampoo therefore changes from a GD-like, target sensitive birth law to a history-adapted, target independent catch-up law. This catch-up law is reached only after the accumulator crosses the damping controlled threshold.

Decay of an unsupported mode. For $\mu = 0$, the first integral reduces to

$$d(\sigma) = d_0 + \frac{s^2 - \sigma^2}{4L} = c_s - \frac{\sigma^2}{4L}, \qquad \text{where} \qquad c_s = d_0 + \frac{s^2}{4L}.$$

The decay dynamics are $\dot{\sigma} = -L\frac{\sigma^{p+1}}{d(\sigma)}$. Therefore,

$$\begin{aligned} T_{\mathrm{Sh}}^{\mathrm{decay}}(s, \delta) &= -\frac{c_s}{L} \int_s^{\delta} x^{-(p+1)}\, dx + \frac{1}{4L^2} \int_s^{\delta} x^{-(p-1)}\, dx \\ &= c_s T_{\mathrm{GD}}^{\mathrm{decay}}(s, \delta) + \frac{1}{4L^2} \int_s^{\delta} x^{-(p-1)}\, dx \end{aligned}$$

Note that $p - 1 = 1 - 2/L$ is less than 1, thus the second term results in finite limit as $\delta \to 0$ from Proposition 3. So, the decay is determined by the first term, and so the decay exponent is $\alpha_{\mathrm{d}} = p+1$, which is the same as Euclidean GD, and $c_{\mathrm{d}} = L/c_s$. The full solution of the decay time is derived from Corollary 2 and proposition 3:

$$T_{\mathrm{Sh}}^{\mathrm{decay}}(s, \delta) = \begin{cases} c_s \log \dfrac{s}{\delta} - \dfrac{s^2 - \delta^2}{8} & L = 1, \\ \dfrac{c_s}{Lp}\,(\delta^{-p} - s^{-p}) - \dfrac{s^{2-p} - \delta^{2-p}}{4L^2(2-p)} & L > 1. \end{cases}$$

The mode consequently decays exponentially for $L = 1$: $\sigma(t) \asymp e^{-t/c_s}$, and algebraic for $L > 1$: $\sigma(t) \sim \left(\frac{Lp}{c_s} t\right)^{-1/p}$.

Thus cumulative Shampoo never removes an unsupported mode in finite time. Its decay has the same logarithmic-versus-polynomial classification as Euclidean GD, but its coefficient retains the mode's accumulated gradient history. This produces a historical inertia effect. At fixed s, δ, and L,

$$\frac{\partial}{\partial d_0} T_{\mathrm{Sh}}^{\mathrm{decay}}(s,\delta) = T_{\mathrm{GD}}^{\mathrm{decay}}(s,\delta) > 0.$$

Thus a larger inherited denominator increases every decay passage time when the initial amplitude is held fixed. Therefore, a mode that has previously received larger gradients begins the decay phase with a larger denominator d_0, and therefore a larger limiting constant c_s. Such a mode takes proportionally longer to remove. Muon has no corresponding cumulative denominator and can drive an unsupported aligned mode to zero in finite time. Cumulative Shampoo instead makes previously active modes increasingly persistent.

Depth strengthens the late decay tail. For $L > 1$, $p = 2 - \dfrac{2}{L}$ increases toward two as L increases. Accordingly, the tolerance time diverges as δ^{-p}, and the trajectory decays as $t^{-1/p}$, whose exponent decreases toward $1/2$ with depth.

Saturation at a nonzero target. At the target $\sigma = \mu$,

$$d_\star = d(\mu) = d_0 + \frac{(\mu - s)^2}{4L}.$$

Near the target, for a residual interval traversed by the trajectory,

$$T_{\mathrm{Sh},-}^{\mathrm{sat}}(r_0,\delta) = \frac{1}{L}\int_\delta^{r_0} \frac{d_\star - r^2/(4L)}{(\mu - r)^p r}\, dr, \qquad \text{and} \qquad T_{\mathrm{Sh},+}^{\mathrm{sat}}(r_0,\delta) = \frac{1}{L}\int_\delta^{r_0} \frac{d_\star - r^2/(4L)}{(\mu + r)^p r}\, dr.$$

Consequently,

$$T_{\mathrm{Sh},\pm}^{\mathrm{sat}}(r_0,\delta) = \frac{d_\star}{L\mu^p}\log\frac{r_0}{\delta} + O(1).$$

From Proposition 4, Cumulative Shampoo consequently saturates monotonically and exponentially $\gamma = 1$ with local rate $\beta = \dfrac{L\mu^p}{d_\star}$. Thus Shampoo approaches a nonzero target only asymptotically. Unlike Muon or SignGD in their ideal memoryless forms, it does not reach the target in finite time.

The dependence of the saturation rate on the target amplitude changes with accumulated history. If the initial denominator dominates, $d_0 \gg \dfrac{(\mu - s)^2}{4L}$, then $\beta \approx \dfrac{L\mu^p}{d_0}$. This has the same target ordering as GD: larger target singular values saturate faster.

If accumulated history dominates, $d_0 \ll \dfrac{(\mu - s)^2}{4L}$, then $\beta \approx \dfrac{4L^2\mu^p}{(\mu - s)^2}$. When $s \ll \mu$, this simplifies to

$$\beta \approx 4L^2\mu^{p-2} = 4L^2\mu^{-2/L}.$$

For modes compared at common s and d_0 with $\mu > s$,

$$\frac{d}{d\mu}\log\beta = \frac{p}{\mu} - \frac{2(\mu - s)}{4Ld_0 + (\mu - s)^2}.$$

Thus the rate now decreases with the target singular value μ precisely when $4Lp\, d_0 < (\mu - s)\big((2 - p)\mu + ps\big)$. Thus sufficiently adapted Shampoo can reverse the GD saturation ordering: larger modes, which generated larger cumulative preconditioners, can converge more slowly near their targets.

As depth increases, the exponent $-\frac{2}{L}$ approaches zero. Hence the history dominated saturation rate becomes less sensitive to the target amplitude, even though its overall depth-dependent prefactor remains. □

Depth dependent phase structure. Cumulative Shampoo has no single birth exponent valid throughout training. At fixed positive damping, an infinitesimal mode first follows $\dot{\sigma} \propto \mu\sigma^p$, with the GD exponent $p = 2 - 2/L$. After the cumulative state dominates, it follows $\dot{\sigma} \propto \sigma^{p-1}$, with adapted exponent $p - 1 = 1 - 2/L$.

For $L = 1$, the adapted exponent is -1. For $L = 2$, it is zero and the adapted singular value grows approximately at constant speed. For $L > 2$, it lies strictly between zero and one, so the adapted phase has a finite hitting time from a vanishing positive mode. Nevertheless, for any fixed $d_0 > 0$, the complete birth time as $s \to 0^+$ remains governed by the preceding GD-like phase.

The decay exponent remains $p + 1$ in the mode equation and produces a polynomial tolerance time of order δ^{-p} for $L > 1$. The saturation rate is $\beta = \dfrac{L\mu^p}{d_\star}$, which is approximately $4L^2\mu^{-2/L}$ at saturation. Thus, it can make larger singular modes saturate more slowly, reversing the order during saturation. Saturation depends jointly on depth, target amplitude, damping, and the mode's accumulated history.

Takeaway. *Cumulative Shampoo is a two-stage spectral optimizer rather than a fixed interpolation between GD and Muon. A newly emerging mode first obeys the GD birth law and therefore retains the finite–logarithmic–polynomial depth barrier under fixed damping. After sufficient gradient history has accumulated, the mode enters a target-independent catch-up phase with exponent $p - 1$, weaker singular-magnitude dependence than Muon, and finite adapted-phase hitting time. The same accumulated state acts as inertia during decay: unsupported modes vanish only asymptotically, and previously active modes become harder to remove. Near a nonzero target, Shampoo saturates monotonically and exponentially, but strong accumulation can reverse GD's mode ordering by making larger target modes converge more slowly.*

E.9 K-FAC Phase Laws

K-FAC updates each layer with the per-layer Fisher or Gauss Newton curvature block. As detailed in Sec. B from Eq. (5), $\mathcal{F}_\ell = (A_\ell A_\ell^\top) \otimes (B_\ell^\top B_\ell)$ where $A_\ell = W_{\ell-1} \cdots W_1$ and $B_\ell = W_L \cdots W_{\ell+1}$. Therefore, from Sec. D.2 based on the balanced aligned model, the corresponding eigenvalue of the layerwise Gauss Newton block is

$$\kappa_\ell = \left(a^{\ell-1}\right)^2 \left(a^{L-\ell}\right)^2 = a^{2L-2} = \sigma^p.$$

The curvature eigenvalue is independent of the layer index ℓ. Consequently, every layer receives the same scalar update, and balancedness is preserved within this aligned idealization.

The damped layer update is

$$\dot{a}_\ell = -\frac{g_\ell}{\kappa_\ell + \lambda_{\mathrm{KF}}} = -\frac{(\sigma - \mu)\sigma^q}{\sigma^p + \lambda_{\mathrm{KF}}}.$$

Recall that $\dot{\sigma} = L\sigma^q \dot{a}$. Hence, the singular value dynamics in the aligned model for each mode is

$$\dot{\sigma} = \frac{L\sigma^p(\mu - \sigma)}{\sigma^p + \lambda_{\mathrm{KF}}}.$$

Relation to exact joint Gauss Newton. The preceding derivation uses a block diagonal curvature model that inverts each layer block separately. It should be distinguished from exact Gauss Newton over all layers jointly. For one balanced scalar mode, the full product Jacobian with respect to the L layer amplitudes is

$$J = a^{L-1}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}.$$

Hence the predictor space curvature is $JJ^\top = La^{2L-2} = L\sigma^p$. Exact damped Gauss Newton therefore gives

$$\dot{\sigma}_{\mathrm{GN}} = \eta \frac{L\sigma^p}{L\sigma^p + \lambda_{\mathrm{GN}}}(\mu - \sigma).$$

Comparing with the K-FAC dynamics derived above, we observe that when damping is zero, both cancel the factorization-induced power σ^p. They differ only by a common factor L. With positive damping, the two filters are not identical unless their damping parameters are scaled consistently.

K-FAC phase laws. From the dynamics $\dot{\sigma} = \dfrac{L\sigma^p(\mu-\sigma)}{\sigma^p+\lambda_{\mathrm{KF}}}$, the factor $\dfrac{\sigma^p}{\sigma^p+\lambda_{\mathrm{KF}}}$ determines how much of the depth-dependent factorization mobility is cancelled.

When $\lambda_{\mathrm{KF}} = 0$ and $\sigma > 0$, the mobility factor cancels: $\dot{\sigma} = L(\mu-\sigma)$. Thus undamped K-FAC behaves like direct gradient flow on the effective singular value, up to the common factor L. When $\lambda_{\mathrm{KF}} > 0$, the cancellation is incomplete for sufficiently small modes.

Proposition 9 (K-FAC phase laws)**.** *Let $0 < s < b < \mu$. The time required for a positive mode initialized at $\sigma(0) = s$ to reach b is*

$$T^{\mathrm{birth}}_{\mathrm{KFAC}}(s,b) = \frac{1}{L}\log\frac{\mu-s}{\mu-b} + \lambda_{\mathrm{KF}} T^{\mathrm{birth}}_{\mathrm{GD}}(s,b)$$

For an unsupported mode with $\mu = 0$, the time required to decay from s to $\delta \in (0,s)$ is

$$T^{\mathrm{decay}}_{\mathrm{KFAC}}(s,\delta) = \begin{cases} (1+\lambda_{\mathrm{KF}})\log\dfrac{s}{\delta}, & L = 1, \\ \dfrac{1}{L}\log\dfrac{s}{\delta} + \dfrac{\lambda_{\mathrm{KF}}}{Lp}\left(\delta^{-p} - s^{-p}\right), & L > 1. \end{cases}$$

For a nonzero target, the time required to reduce the residual from a fixed r_0 to δ satisfies from either side satisfies,

$$T^{\mathrm{sat}}_{\mathrm{KF},\pm}(r_0,\delta) = \frac{1}{L}\log\frac{r_0}{\delta} + \lambda_{\mathrm{KF}} T^{\mathrm{sat}}_{\mathrm{GD},\pm}(r_0,\delta).$$

Consequently,

$$T^{\mathrm{sat}}_{\mathrm{KF},\pm}(r_0,\delta) = \frac{\mu^p+\lambda_{\mathrm{KF}}}{L\mu^p}\log\frac{r_0}{\delta} + O(1).$$

Proof.

Birth of a positive mode. From the dynamics, we get

$$\begin{aligned} T^{\mathrm{birth}}_{\mathrm{KFAC}}(s,b) &= \frac{1}{L}\int_s^b (\mu-x)^{-1}\,dx + \frac{\lambda_{\mathrm{KF}}}{L}\int_s^b \frac{\sigma^{-p}}{(\mu-x)}\,dx \\ &= \frac{1}{L}\log\frac{\mu-s}{\mu-b} + \lambda_{\mathrm{KF}} T^{\mathrm{birth}}_{\mathrm{GD}}(s,b). \end{aligned}$$

Thus, the K-FAC birth time decomposes into two terms.

Undamped K-FAC. When $\lambda_{\mathrm{KF}} = 0$, the mode dynamics reduce to $\dot{\sigma} = L(\mu-\sigma)$. The birth time is $T^{\mathrm{birth}}_{\mathrm{KFAC}}(s,b) = \dfrac{1}{L}\log\dfrac{\mu-s}{\mu-b}$, thus finite time emergence. The exact trajectory is $\sigma(t) = \mu - (\mu-s)e^{-Lt}$.

The near-zero birth dynamics are $\dot{\sigma} \approx L\mu$. Hence the undamped K-FAC birth exponent α_{b} is zero. The initial velocity does not vanish with σ. Undamped K-FAC therefore removes the depth-dependent small-mode barrier.

For a common absolute threshold b and seed s, ideal Muon has target independent birth time whenever $\mu > b$. Undamped curvature flow instead has birth time

$$\frac{1}{L}\log\frac{\mu-s}{\mu-b},$$

which decreases with μ.

The target amplitude nevertheless remains in the velocity. This distinguishes undamped K-FAC from ideal Muon. Muon removes both the factorization barrier and target-amplitude ordering during aligned birth, whereas undamped K-FAC removes the barrier but preserves target ordering.

The cancellation at exactly $\sigma = 0$ requires care. At an exactly zero mode, the undamped Fisher block is singular, so the update depends on the inverse or pseudoinverse convention. The finite birth statement above concerns the limit from a positive seed $s \to 0^+$.

Damped K-FAC. For $\lambda_{\mathrm{KF}} > 0$, define the spectral curvature threshold $\sigma_{\mathrm{KF}}^{\mathrm{b}} = \lambda_{\mathrm{KF}}^{1/p},\ L > 1$. This is the scale at which $\sigma^p \asymp \lambda_{\mathrm{KF}}$. When $\sigma \ll \sigma_{\mathrm{KF}}^{\mathrm{b}}$, the preconditioning factor is approximately $\dfrac{\sigma^p}{\lambda_{\mathrm{KF}}}$. For a small positive mode with $\sigma \ll \mu$, $\dot{\sigma} \approx \dfrac{L\mu}{\lambda_{\mathrm{KF}}}\sigma^p$. Thus damped K-FAC recovers the Euclidean GD birth exponent $\alpha_{\mathrm{b}} = p$.

When $\sigma \gg \sigma_{\mathrm{KF}}^{\mathrm{b}}$, the preconditioning factor approaches one. The mode then follows $\dot{\sigma} \approx L(\mu - \sigma)$, and behaves like direct optimization of the effective matrix.

Damped K-FAC therefore has a two-regime birth law: GD-like below $\sigma_{\mathrm{KF}}^{\mathrm{b}}$, and curvature-corrected above $\sigma_{\mathrm{KF}}^{\mathrm{b}}$. Thus any fixed positive damping restores the Euclidean finite–logarithmic–polynomial birth classification in the infinitesimal limit.

For $L \geq 2$, the limits $\lambda_{\mathrm{KF}} \to 0$ and $s \to 0$ do not commute. For every fixed $s > 0$,

$$\lim_{\lambda_{\mathrm{KF}} \to 0} T_{\mathrm{KFAC}}^{\mathrm{birth}}(s, b) = \frac{1}{L} \log \frac{\mu - s}{\mu - b}.$$

In contrast, for every fixed $\lambda_{\mathrm{KF}} > 0$, the limit $s \to 0^+$ is logarithmically divergent at $L = 2$ and polynomially divergent for $L > 2$. A numerically small damping can therefore be asymptotically decisive for sufficiently small spectral modes.

For $L = 1$, $\dot{\sigma} = \dfrac{\mu - \sigma}{1 + \lambda_{\mathrm{KF}}}$, so damping only changes the speed.

For $L > 1$, an exactly zero mode satisfies $\dot{\sigma} = 0$ under every fixed positive damping because the numerator contains σ^p. Damped K-FAC therefore also requires a positive seed to initiate growth.

Decay of an unsupported mode. For zero target mode $\mu = 0$, the K-FAC mode dynamics equation becomes $\dot{\sigma} = -L\dfrac{\sigma^{p+1}}{\sigma^p + \lambda_{\mathrm{KF}}}$.

When $\lambda_{\mathrm{KF}} = 0$, this reduces to $\dot{\sigma} = -L\sigma$. Hence

$$T_{\mathrm{KFAC},0}^{\mathrm{decay}}(s, \delta) = \frac{1}{L} \log \frac{s}{\delta}, \qquad \text{and} \qquad \sigma(t) = se^{-Lt}.$$

Undamped K-FAC removes an unsupported mode exponentially in time, but the mode reaches zero only asymptotically. Its tolerance time is logarithmic.

When $\lambda_{\mathrm{KF}} > 0$ and $\sigma \gg \sigma_{\mathrm{KF}}^{\mathrm{b}}$, the decay is still approximately exponential: $\dot{\sigma} \approx -L\sigma$. Once $\sigma \ll \sigma_{\mathrm{KF}}^{\mathrm{b}}$, the damping term dominates and $\dot{\sigma} \approx -\frac{L}{\lambda_{\mathrm{KF}}}\sigma^{p+1}$.

For $L = 1$, $p = 0$ and damping only rescales the exponential rate: $\dot{\sigma} = -\frac{1}{1+\lambda_{\mathrm{KF}}}\sigma$, so

$$T_{\mathrm{KFAC}}^{\mathrm{decay}}(s, \delta) = (1 + \lambda_{\mathrm{KF}}) \log \frac{s}{\delta}.$$

For $L > 1$, the final decay is algebraic as shown in Proposition 9. Equivalently, the singular value evolves as $\sigma(t) \sim \left(\dfrac{Lp}{\lambda_{\mathrm{KF}}} t\right)^{-1/p}$. Thus positive damping converts the undamped exponential decay into a polynomial tail for every depth $L > 1$.

The curvature threshold governs both birth and decay. Above $\sigma_{\mathrm{KF}}^{\mathrm{b}}$, K-FAC behaves like direct effective-matrix optimization. Below it, damping restores the same small-mode mobility that appears under Euclidean GD.

Saturation at a nonzero target. The exact saturation times from below and above are,

$$T_{\mathrm{KF},\pm}^{\mathrm{sat}}(r_0, \delta) = \frac{1}{L} \log \frac{r_0}{\delta} + \lambda_{\mathrm{KF}} T_{\mathrm{GD},\pm}^{\mathrm{sat}}(r_0, \delta).$$

Consequently, $T_{\mathrm{KF},\pm}^{\mathrm{sat}}(r_0, \delta) = \dfrac{\mu^p + \lambda_{\mathrm{KF}}}{L\mu^p} \log \dfrac{r_0}{\delta} + O(1)$.

Therefore, the residual dynamics is $\dot{r} \approx -L\dfrac{\mu^p}{\mu^p + \lambda_{\mathrm{KF}}} r$. This implies K-FAC saturates monotonically and exponentially, with the rate $\beta = L\dfrac{\mu^p}{\mu^p + \lambda_{\mathrm{KF}}}$.

In the undamped case, $\beta = L$. The rate is independent of the target singular value and of the mobility exponent p. Under the present time normalization it is proportional to depth; after rescaling time by the common factor L, every mode has unit saturation rate.

With positive damping, the saturation rate has two regimes. If $\mu^p \gg \lambda_{\mathrm{KF}}$, then $\beta \approx L$, so the mode behaves like undamped K-FAC. If $\mu^p \ll \lambda_{\mathrm{KF}}$, then $\beta \approx \dfrac{L}{\lambda_{\mathrm{KF}}}\mu^p$. Small target modes therefore recover the GD-like target sensitivity and saturate more slowly.

K-FAC always approaches a nonzero target asymptotically in this model. It does not have the finite-time relay behavior of ideal Muon. □

Role of depth. Undamped K-FAC removes the depth-dependent factorization exponent from all three phases. Positive modes have finite birth time, unsupported modes decay exponentially, and nonzero targets saturate exponentially. Depth remains only through the common multiplicative rate L under the present time convention.

Positive damping restores depth dependence through p. For birth, the asymptotic delay changes from finite at $L = 1$ to logarithmic at $L = 2$ and polynomial for $L > 2$, with exponent $p - 1$. For unsupported modes, damping produces a polynomial tolerance time for $L > 1$. For saturation, small target modes have rate $\dfrac{L}{\lambda_{\mathrm{KF}}}\mu^p$, whereas sufficiently large modes approach the undamped rate L.

Thus damping does more than stabilize curvature inversion. It determines which spectral modes experience natural-gradient-like cancellation and which retain the depth-dependent mobility of Euclidean GD.

Takeaway. *Undamped K-FAC cancels the factorization-induced small-mode barrier while preserving target-amplitude ordering: positive modes are born in finite time, unsupported modes decay exponentially, and all nonzero targets share the same normalized saturation rate. Additive damping introduces the spectral threshold $\sigma_{\mathrm{KF}}^{\mathrm{b}} = \lambda_{\mathrm{KF}}^{1/p}$. Above this threshold, K-FAC remains curvature corrected; below it, the Euclidean depth laws return. Consequently, any fixed positive damping restores logarithmic or polynomial birth delays, produces a polynomial zero-target tail for $L > 1$, and slows the saturation of small target modes. Damping is therefore an inductive-bias parameter, not merely a numerical regularizer.*

F Experimental Settings

We use synthetic deep linear models to isolate the effects of optimizer geometry from nonlinear representation effects. Unless stated otherwise, the effective predictor is $W = W_L W_{L-1} \cdots W_1$, and all hidden layers have width equal to the input dimension. We use deterministic full batch optimization and do not use weight decay or explicit regularization. All experiments use double precision. We report continuous optimization time as $t = k\eta$, where k is the optimization step and η is the step size or learning rate.

For comparisons between optimizers, the learning rate is selected separately for each method because the update magnitudes induced by Euclidean, normalized, spectral, and curvature based methods are not directly comparable. For each optimizer, we sweep the learning rate over a logarithmic grid and select the largest value that gives stable monotone progress without numerical divergence or persistent finite step oscillations away from the target, as well as reaches a similar loss within an admissible range for all optimizers. We then adjust the training horizon so that all optimizers reach approximately the same final training loss. In general, we aim to train for 5000 steps or lower, with each optimizer attaining similar losses. This prevents differences in optimization progress from being interpreted as differences in spectral bias. We plot comparisons of spectra, effective rank, and mode trajectories over time.

Spectral diagnostics. Write $W = U\Sigma V^\top$ for a singular value decomposition, with singular values $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$, where $r = \min(d_0, d_L)$. We measure stable rank and effective rank using

$$\mathrm{srank}(W) = \frac{\sum_i \sigma_i^2}{\sigma_1^2}, \qquad \mathrm{erank}(W) = \exp\left(-\sum_k p_k \log p_k\right),\ p_k = \frac{\sigma_k}{\sum_i \sigma_i}.$$

These quantities measure different aspects of spectral concentration and are reported together with the training loss. Thresholds and conventions at $W = 0$ are fixed before comparison.

Matrix Factorization. The first setting directly matches the full observation model used in the phase analysis. The input and output dimensions are both 16, and the network has depth $L = \{2, 4, 8\}$ with hidden width 16. The target matrix is diagonal in the initialization basis,

$$M = \mathrm{diag}\left(1.00, 0.80, 0.60, 0.45, 0.32, 0.24, 0.18, 0.12, 0, \ldots, 0\right),$$

and therefore has rank 8. This choice gives several positive target modes with distinct amplitudes together with unsupported modes whose target is zero.

We use balanced aligned initialization. For the phase law experiments, positive target modes are initialized at effective amplitude 10^{-3} while zero target modes are initialized at 0.2. The corresponding factor amplitudes are initialized as $a_i(0) = s_i^{1/L}$, so that $\sigma_i(0) = a_i(0)^L = s_i$ for every depth. This initialization preserves the aligned construction used in the theoretical analysis and allows birth and decay to be observed within the same run.

The phase statistics are evaluated every optimization step. For birth, we track the local mode dynamics over $10^{-3} \leq \sigma \leq 10^{-1}$. For decay, we follow a zero target mode from its initial active value toward zero. For saturation, we estimate the residual dynamics $r(t) = |\mu - \sigma(t)|$ for every positive target mode. Local exponents are estimated by linear regression over a short sliding window in logarithmic coordinates. The default window contains 21 recorded points.

For instantaneous Adam, we set $\beta_1 = \beta_2 = 0$ so that the update isolates the normalization mechanism studied in the phase analysis. For Muon, we use exact compact polar normalization and no momentum. For cumulative Shampoo, the left and right statistics are accumulated without exponential forgetting, the initial accumulator is $\lambda_{\mathrm{Sh}} I$, and the preconditioner is refreshed at every step. For K-FAC, we use the additive block Gauss Newton model analyzed in the theory. No update norm clipping is applied to K-FAC. The reported learning rates are numerical integration parameters rather than quantities used to establish optimizer rankings. We train all the optimizers for 5000 steps or lower, and the hyperparameters are given in Table 2.

Table 2: Optimizer settings used for the full observation matrix factorization experiments using depths $L = \{2, 4, 8\}$. The learning rates are chosen to keep the discrete trajectories close to the corresponding continuous flows while allowing all methods to reach comparable final training loss. Each optimizer is trained for a maximum of 5000 steps.

Optimizer	Learning rate	Main state parameter	State value	Depth
GD	5×10^{-2}	none	none	$L = \{2, 4\}$
GD	8×10^{-2}	none	none	$L = \{8\}$
Adam	1×10^{-3}	$(\beta_1, \beta_2, \epsilon)$	$(0, 0, 0.02)$	$L = \{2, 4\}$
Adam	2×10^{-3}	$(\beta_1, \beta_2, \epsilon)$	$(0, 0, 0.02)$	$L = \{8\}$
Muon	5×10^{-4}	momentum	0	$L = \{2, 4, 8\}$
Shampoo	2×10^{-4}	λ_{Sh}	10^{-6}	$L = \{2, 4, 8\}$
K-FAC, undamped	5×10^{-4}	λ_{KF}	0	$L = \{2, 4, 8\}$
K-FAC, damped	5×10^{-4}	λ_{KF}	5×10^{-3}	$L = \{4, 8\}$
K-FAC, damped	8×10^{-4}	λ_{KF}	5×10^{-3}	$L = \{2\}$

Linear Regression. The second setting studies optimizer behavior when the observations do not uniquely determine the effective matrix. We generate noiseless multivariate linear regression

Table 3: Optimizer settings used for underdetermined linear regression with depth $L = 4$. The learning rates were selected over a different grid for different optimizers to obtain stable training and comparable final data loss.

Optimizer	Learning rate	Main state parameter	State value
GD	3	none	none
Adam	0.05	$(\beta_1, \beta_2, \epsilon)$	$(0, 0, 0.02)$
Muon	0.00025	momentum	0
Shampoo	0.003	λ_{Sh}	10^{-6}
K-FAC	0.0005	λ_{KF}	10^{-4}

data according to $Y = W_{\mathrm{true}}X$, and train the deep linear predictor by minimizing $\mathcal{L}(W) = \frac{1}{2n}\|WX - Y\|_F^2$.

We use input output data dimensions as 64 and number of training samples 32, so the regression problem is underdetermined. The teacher has rank 8. Its nonzero singular values are the same as M in matrix factorization. The left and right singular directions of the teacher are sampled once from random orthogonal bases and are kept fixed across all optimizer runs.

The entries of $X \in \mathbb{R}^{64\times 32}$ are sampled independently from a standard Gaussian distribution. We do not whiten the data. Consequently, $\Sigma_X = \frac{1}{n}XX^\top$ has rank at most 32 and is not proportional to the identity. The effective gradient is therefore $G = (W - W_{\mathrm{true}})\Sigma_X$, rather than $W - W_{\mathrm{true}}$. This setting intentionally leaves the balanced aligned scalar regime and tests whether the optimizer dependent spectral behavior persists when the data covariance couples target modes.

The network has depth $L = 4$ and hidden width 64. Each layer is initialized with a scaled orthogonal matrix. The initialization scale is 0.1, giving a small but full rank initial effective predictor while avoiding the extremely slow escape observed for much smaller deep products. The same initialization is used for all optimizers.

For this experiment, we use similar optimizers as in the matrix factorization experiment. The specific learning rates are selected for this finite sample problem rather than transferred from the aligned full observation experiments.

Because the optimizers have substantially different effective step sizes, a common iteration count does not imply a common degree of optimization. We therefore use the final training loss as the primary tuning criterion. For each optimizer, the learning rate is selected from a different grid for different optimizers, particularly smaller values for strongly preconditioned methods when necessary. Configurations that diverge or produce persistent numerical instability are discarded. Among the stable configurations, we select the largest learning rate for which the final training loss is within a factor of two of the lowest final loss attained by the compared methods. We obtain the hyperparameters in Table 3 such that the training loss for all optimizers is within a range when trained for 5000 steps or lower.

Unless stated otherwise, all reported curves use the same random data realization and the same model initialization across optimizers.

Hardware and Code. Each experiment is run on 1 NVIDIA GeForce RTX 4090 and takes about 5-10 minutes to complete for each optimizer. A functioning code repository is provided as a zip file in the supplementary material.

G Additional Experimental Results

G.1 Matrix Factorization

We first evaluate the optimizer dependent spectral dynamics in the full observation matrix factorization setting described in Sec. F, where the theoretical phase laws apply most directly.

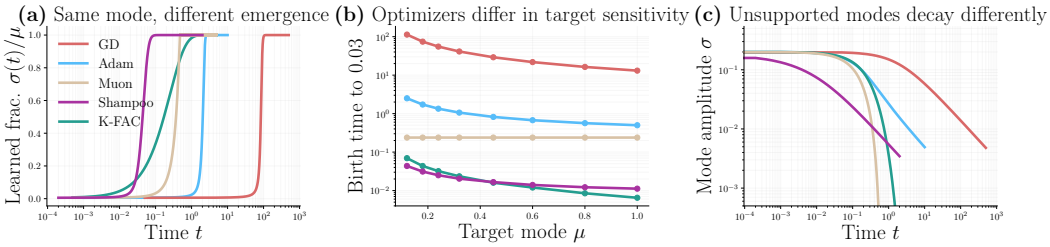


Figure 4: **Optimizer geometry sets the pace of spectral learning evaluation in matrix factorization. (a)** Evolution of a weak target mode with $\mu = 0.18$, shown as the learned fraction $\sigma(t)/\mu$. GD exhibits the factorization induced small mode barrier, whereas normalization and preconditioning substantially accelerate emergence. **(b)** Birth time to the common amplitude $b = 0.03$ as the target mode μ varies. While ideal Muon is independent of the target mode ordering during birth, other optimizers are target sensitive. **(c)** Decay trajectories for an unsupported mode with target $\mu = 0$. Ideal Muon removes the mode in finite time, undamped K-FAC produces exponential decay, while GD, instantaneous Adam with fixed stabilization, and cumulative Shampoo retain increasingly slow tails near zero. Optimizer geometry therefore determines which spectral components emerge, persist, and disappear.

Fig. 4 provides a broader validation of the theory in addition to the theoretical evaluation done in Sec. 6. Fig. 5 compares the optimization and rank dynamics. Although the methods eventually reach similar losses and effective ranks, their trajectories differ substantially. GD learns the target modes sequentially and retains a low effective rank for longer, consistent with the small mode mobility factor σ^p. Adam reduces this separation and activates weaker modes much earlier. Muon produces an even more rapid redistribution of spectral mass, reflecting its normalized spectral update. Cumulative Shampoo initially behaves more like GD but quickly changes regime as its accumulated preconditioner adapts. K-FAC exhibits a smoother evolution in which curvature correction reduces the factorization induced separation between mode time scales. The stable and effective rank curves therefore mainly differ during the transient phase, while their later agreement reflects the unique full observation target rather than identical optimization dynamics.

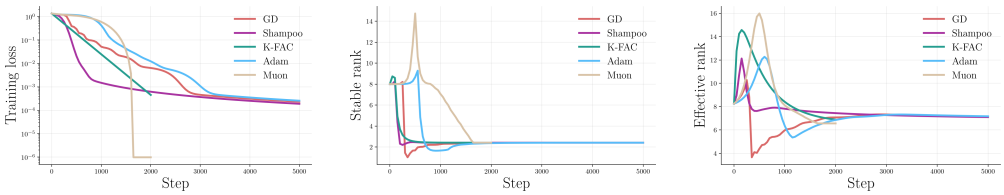
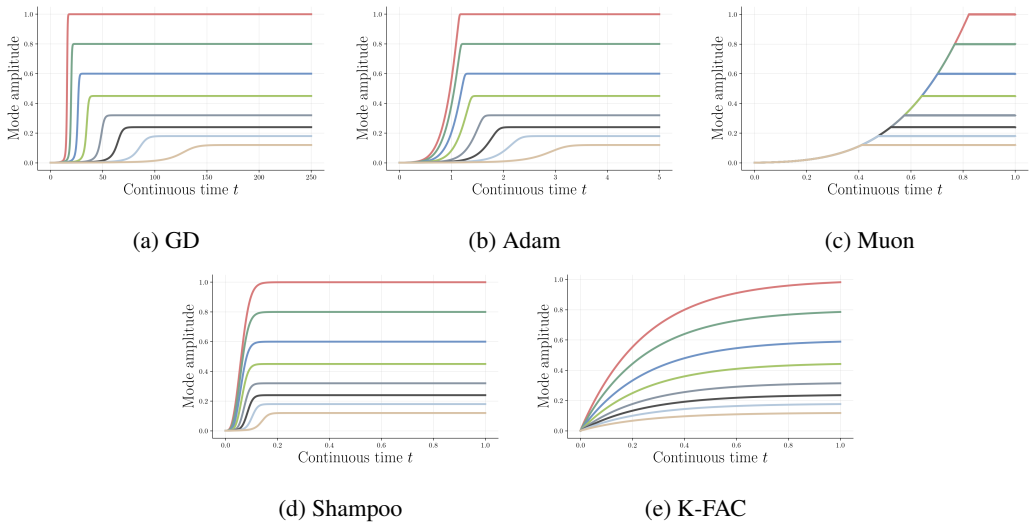


Figure 5: **Optimization and rank dynamics in matrix factorization.** All optimizers train the same depth $L = 4$ deep linear factorization toward the same rank eight target and reach comparable final training loss. **Left:** training loss over optimization steps. **Middle:** stable rank of W. **Right:** effective rank of W. The optimizers exhibit markedly different transient rank dynamics despite converging toward similar final spectra. GD retains a lower rank representation for longer, whereas Adam, Muon, cumulative Shampoo, and K-FAC alter the rate at which spectral mass is distributed across modes.

The singular value trajectories in Fig. 6 make these differences more explicit. Under GD, larger target modes emerge first and weaker modes remain close to initialization for substantially longer times. This sequential emergence is the dynamical mechanism underlying its transient low rank bias. Adam substantially compresses these mode dependent delays, while Muon causes the positive modes to follow nearly the same spectral clock until they encounter their different target amplitudes. Cumulative Shampoo exhibits the predicted two stage behavior: early growth remains separated across modes, after which accumulated gradient history accelerates the weaker modes and reduces target amplitude dependence. K-FAC produces the weakest separation between mode trajectories, consistent with curvature correction compensating for the factorization induced mobility barrier. These trajectories illustrate that optimizers reaching similar final solutions can nevertheless pass through very different spectral representations during training.



(a) GD (b) Adam (c) Muon

(d) Shampoo (e) K-FAC

Figure 6: **Singular mode trajectories reveal optimizer dependent spectral clocks.** GD exhibits strongly sequential learning;Adam reduces these delays, while Muon makes the modes evolve on nearly the same clock until they approach their respective targets. Cumulative Shampoo shows a faster history adapted growth of weaker modes. K-FAC further reduces the factorization induced separation between mode time scales through curvature correction. The common final amplitudes are determined by the unique full observation target, whereas the transient trajectories expose the distinct inductive biases of the optimizers.

G.1.1 Role of Depth

Finally, we vary the factorization depth while keeping the target matrix and effective initialization fixed. Fig. 7 focuses on Adam, cumulative Shampoo, and K-FAC for depths$= \{2, 4, 8\}$ because their state dependent transitions make this interaction particularly explicit. For Adam, increasing depth raises both the GD like and normalized exponents and changes the location of the ϵ controlled crossover. For Shampoo, depth changes both the initial exponent p and the history adapted exponent $p-1$, while also shifting the point at which accumulated history overtakes the initial damping. For damped K-FAC, increasing depth strengthens the GD like small mode regime through p, whereas the curvature corrected regime continues to approach exponent zero. The corresponding depth dependence for GD and Muon follows directly from their fixed exponents p and q and is consistent with the phase laws reported in the main text.

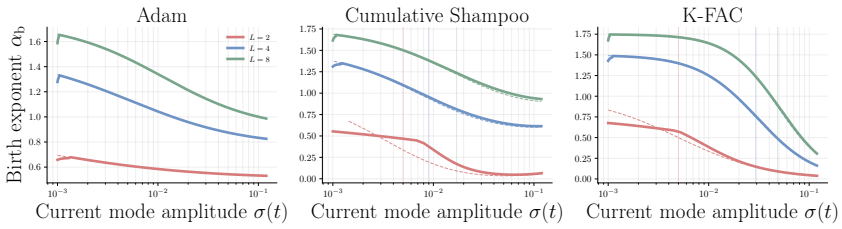


Figure 7: **Depth changes both the spectral exponents and the optimizer state transitions. Left:** for instantaneous Adam, increasing depth changes both the GD like exponent $p = 2 - 2/L$ and the normalized exponent $q = 1 - 1/L$. **Middle:** for cumulative Shampoo, depth changes the initial GD like exponent p, the history adapted exponent $p-1$, and the amplitude at which accumulated history begins to dominate. **Right:** for damped KFAC, deeper factorizations strengthen the small mode GD like barrier while curvature correction still drives the exponent toward zero above the damping controlled mobility threshold.

Fig. 8 studies the decay exponent of unsupported modes and focuses on Shampoo saturation, where mode order reversal can happen theoretically. For decay, increasing depth changes the small mode geometry through $p = 2 - 2/L$. Empirically, GD, Adam, and cumulative Shampoo follow the predicted increase $\alpha_d = p + 1 = 3 - \frac{2}{L}$, so deeper factorizations produce a steeper polynomial tail near zero. In contrast, undamped K-FAC remains close to $\alpha_d = 1$, reflecting its exponential decay law, while Muon exhibits a much weaker depth dependence, consistent with its normalized geometry. Thus depth does not affect all optimizers uniformly: it strengthens the Euclidean style barrier for GD like methods, but has much less impact on normalized or curvature corrected decay.

For saturation, we focus on cumulative Shampoo because it is the optimizer for which mode reversal can happen and depth also affects it. Its asymptotic saturation coefficient depends on both target size and accumulated history, $\beta_{\mathrm{Sh}}(\mu) = \dfrac{L\mu^p}{D_\star(\mu)}$. As a result, the ordering of saturation rates across target modes need not agree with GD. The right panel shows precisely this effect. For $L = 2$, β decreases with μ, so smaller target modes saturate faster than larger ones. For $L = 8$, the trend reverses and larger target modes saturate faster. The case $L = 4$ lies near the transition and becomes weakly nonmonotone. This depth driven change in ordering is the clearest empirical instance of the Shampoo mode reversal predicted by the theory.

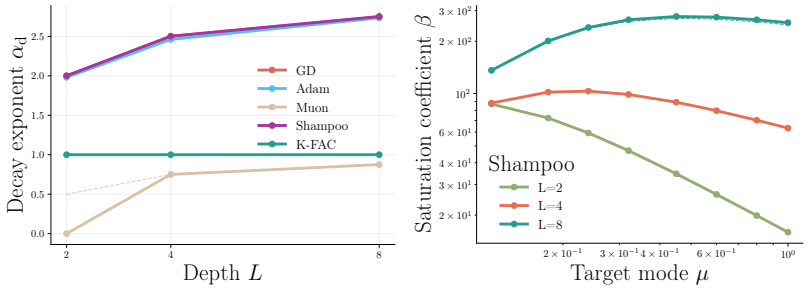


Figure 8: **Effect of depth on decay laws and Shampoo saturation order. Left:** empirical decay exponents α_d for unsupported modes as depth varies. All optimizers closely follow the theoretical predictions in general. **Right:** Shampoo saturation coefficient β as a function of the positive target mode μ for $L \in \{2, 4, 8\}$. Shampoo is singled out here because its saturation phase is where mode order reversal can happen theoretically. At $L = 2$, β decreases with μ, so smaller target modes saturate faster. At $L = 8$, the ordering reverses and larger target modes saturate faster, while $L = 4$ lies near the transition. These results show that depth influences not only the size of the phase law exponents, but also the ordering of mode fitting under optimizer state dynamics.

Figs. 9 and 11 show the evolution of training loss and ranks throughout training for depths $L = 2$ and 8, respectively. In both cases, all optimizers converge to similar rank and loss solutions. But their trajectory shows the difference. Similarly, Figs. 10 and 12 show the singular mode trajectories for $L = 2$ and 8, respectively. In both cases, we observe that GD exhibits strong sequential learning and at the same time other optimizers also begin to show stronger bias with depth.

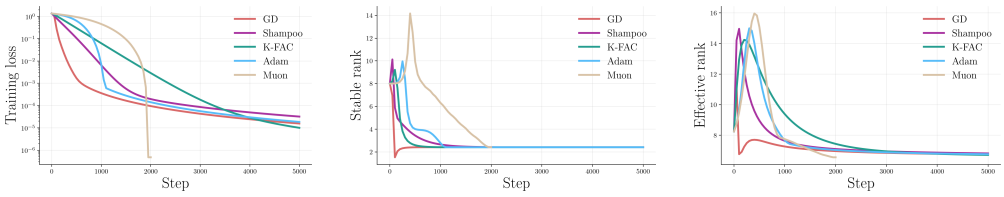
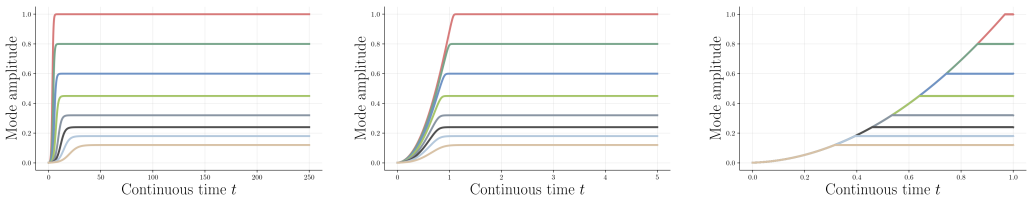
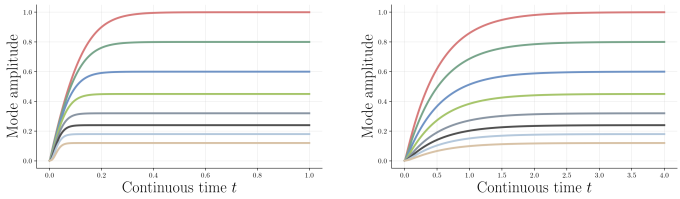


Figure 9: **Optimization and rank dynamics in full observation matrix factorization.** All optimizers train the same depth $L = 2$ deep linear factorization toward the same rank eight target and reach comparable final training loss.



(a) GD (b) Adam (c) Muon



(d) Shampoo (e) K-FAC (damped)

Figure 10: **Singular mode trajectories show optimizer dependent spectral clocks for $L = 2$.**

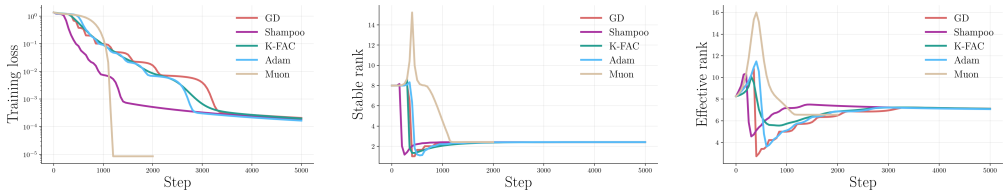
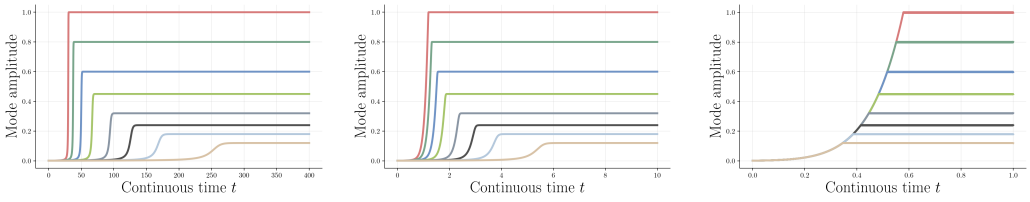
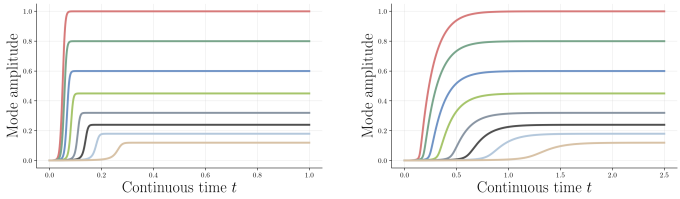


Figure 11: **Optimization and rank dynamics in full observation matrix factorization.** All optimizers train the same depth $L = 8$ deep linear factorization toward the same rank eight target and reach comparable final training loss.



(a) GD (b) Adam (c) Muon



(d) Shampoo (e) K-FAC (damped)

Figure 12: **Singular mode trajectories show optimizer dependent spectral clocks for $L = 8$.**

G.1.2 Adam Phase Transitions

In addition to the birth exponent plot in Fig. 3a, here we provide the validation of the theoretical decay exponent and saturation rate in Fig. 13. Both align well with the theory. For saturation rate, we choose to plot normalized residual $x = \dfrac{r\mu^q}{\epsilon}$ vs normalized saturation rate $\dfrac{\beta}{L\mu^p/\epsilon}$ for ease of representation. Theoretically, near the saturation $\sigma \approx \mu$, so $\beta = \dfrac{L\mu^p}{r\mu^q + \epsilon}$. Therefore, the normalized saturation rate reduces to

$$\frac{\epsilon}{r\mu^q + \epsilon} = \frac{1}{1+x}.$$

The normalized representation also clarifies why comparing the raw saturation rates can be misleading. The limiting rate scales as $1/\epsilon$, so changing ϵ by several orders of magnitude produces the same change in the vertical scale even when the underlying dynamics are identical after rescaling. In contrast, the normalized variables remove this trivial dependence.

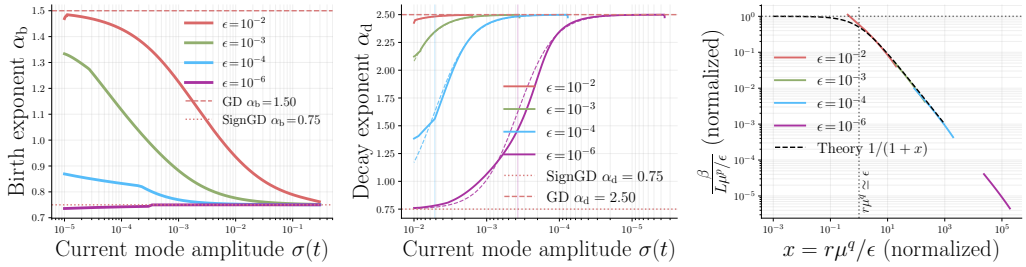


Figure 13: **Birth, decay and normalized saturation dynamics of Adam.** Dotted and solid lines mark the theoretical and empirical values for $L = 4$, respectively. These values align in both the decay and saturation rate plots across different values of ϵ.

G.1.3 Shampoo Phase Transitions

Fig. 14 examines how the cumulative Shampoo state changes the birth, decay, and saturation dynamics. The birth transition from the initial GD exponent p toward the history adapted exponent $p-1$ is discussed in the main text. Here we focus on the corresponding decay and saturation behavior. For an unsupported mode with $\mu = 0$, different values of λ_{Sh} produce different finite amplitude trajectories, but all curves approach the common asymptotic exponent $p+1 = 2.5$ for $L = 4$. Even in the case of smaller damping, the deviation from 2.5 is only in the second decimal place. Thus Shampoo retains a polynomial tail near zero rather than the finite time extinction of ideal Muon. For a positive target mode, the right panel shows each damping level approaches a distinct constant rate, with the empirical curves closely following the finite amplitude prediction. The limiting rate depends on the entire accumulated gradient history and this distinguishes Shampoo from memoryless spectral normalization and explains why modes with similar current residuals can saturate at different rates.

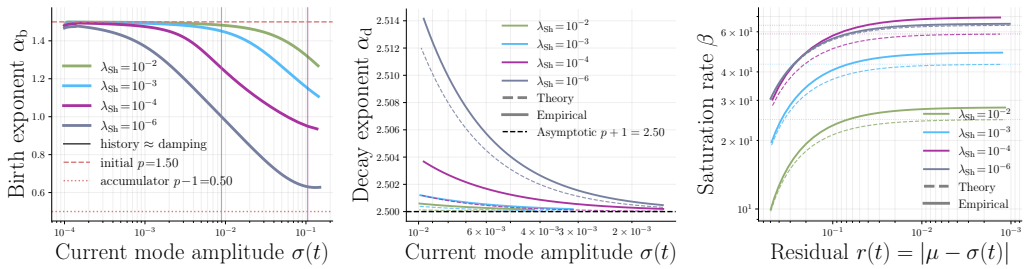


Figure 14: **Phase transitions induced by cumulative Shampoo state.** Solid curves show empirical dynamics and dashed curves show the corresponding aligned theory for depth $L = 4$. Together, the panels show that cumulative Shampoo changes its effective spectral geometry during training while retaining explicit memory of earlier gradient activity.

G.1.4 K-FAC Phase Transitions

Fig. 15 illustrates how additive damping in K-FAC controls the spectral dynamics across all three phases. For birth, positive damping restores the small mode barrier. Reducing λ_{KF} shifts this transition toward smaller mode amplitudes, recovering the undamped KFAC regime.

For an unsupported mode, the middle panel confirms that stronger damping causes the polynomial regime to appear at larger amplitudes, following the theory.

Near a positive target, K-FAC's saturation rate is nearly constant along the final approach to the target and approaches the undamped value $L = 4$ as $\lambda_{\mathrm{KF}} \to 0$. Thus, K-FAC damping controls not only numerical regularization of the curvature inverse, but also the spectral scale below which GD-like dynamics reappear.

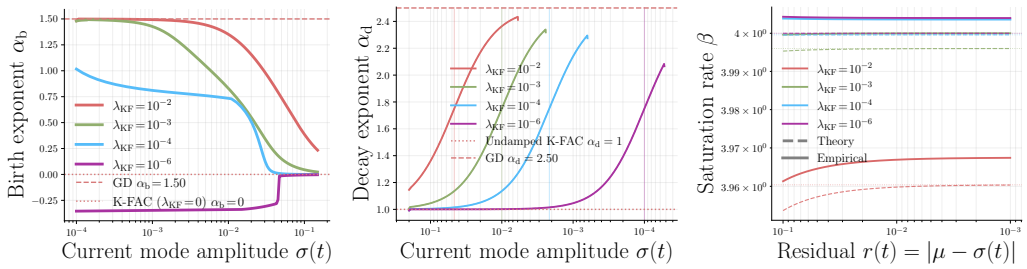


Figure 15: **Additive K-FAC damping controls the spectral phase transitions.** Solid curves show empirical dynamics and dashed curves show the corresponding aligned theory for depth $L = 4$. It shows that λ_{KF} defines a spectral mobility threshold separating curvature corrected dynamics from GD-like behavior.

G.2 Linear Regression

We next consider a less structured setting in which the effective predictor is trained by underdetermined linear regression as described in Sec. F. This experiment tests whether the qualitative predictions persist once our simplifying assumptions Assumptions 1 to 3 are removed.

Fig. 16 shows that the optimizers reach comparable training losses and converge to nearly identical final singular spectra. This agreement is important: the pronounced differences observed during training are not simply consequences of one method converging to a substantially different level of fit. Nevertheless, the stable and effective ranks evolve differently. Thus similar terminal predictors need not imply similar spectral learning dynamics.

The singular value trajectories in Fig. 17 make this distinction particularly clear. GD exhibits pronounced separation between the emergence times of successive singular values, consistent with the factorization induced small mode barrier derived in the aligned setting. Adam substantially compresses these delays. Muon produces a markedly more synchronized growth of the singular spectrum, consistent with normalization reducing mode dependent mobility differences. Cumulative Shampoo also accelerates weaker spectral components after its preconditioner has accumulated sufficient history, while K-FAC produces comparatively smooth growth across the spectrum through curvature correction. These effects remain visible even though the singular vectors are not fixed and the data covariance couples the modes.

The experiment therefore supports the interpretation of the theoretical phase laws as mechanisms governing transient spectral learning rather than merely properties of the special aligned model. Exact exponents depend on the aligned reduction, but the resulting qualitative distinctions persist in the more practical underdetermined regression problem.

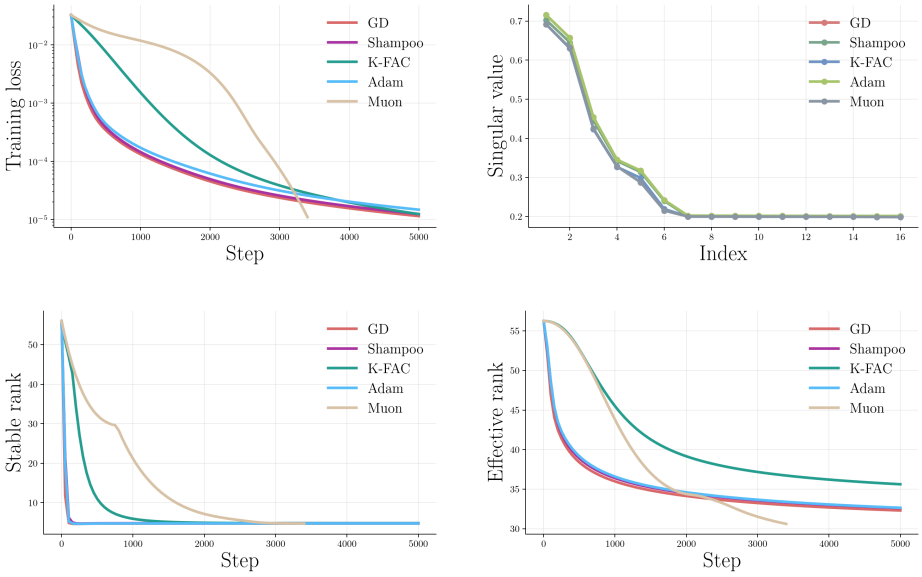
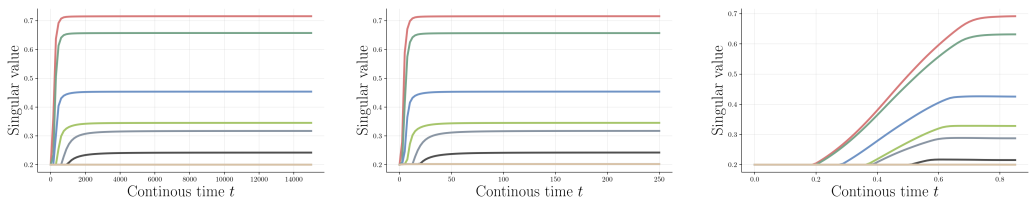
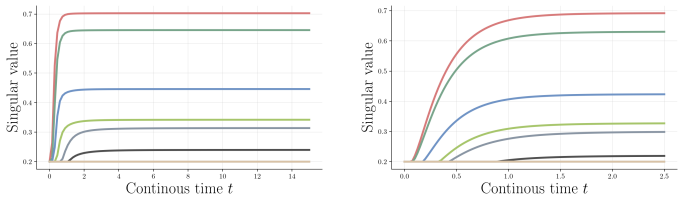


Figure 16: **Optimizer dependent dynamics persist in underdetermined linear regression.** We train the same deep linear predictor on an underdetermined, nonwhitened regression problem using GD, Adam, Muon, cumulative Shampoo, and KFAC. All methods reach comparable final loss and final spectra of W despite substantially different transient dynamics shown by the stable rank and effective rank throughout training. The optimizer geometry strongly changes how quickly spectral mass concentrates.



(a) GD (b) Adam (c) Muon



(d) Shampoo (e) K-FAC

Figure 17: **Singular value trajectories in nonaligned underdetermined linear regression.** Although all methods approach similar final singular spectra, their transient spectrum differ substantially. The qualitative optimizer dependent dynamics predicted by the aligned phase analysis remain visible even when the empirical data covariance couples singular directions and the exact scalar mode reduction no longer holds.