GROKKING AS SLOW TRANSPORT: A QUANTITATIVE FRAMEWORK OF DELAYED GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Delayed generalization, or *grokking*, has been explained by many different mechanisms in the literature, with conclusions that sometimes appear to conflict. We show that these mechanisms can be understood as different kinetic regimes of a common process: the motion of a slow coordinate out of a metastable state. Starting from the minibatch update rule, we reduce plateau training dynamics to a one-dimensional first-passage problem and derive a governing equation for the grokking delay. This yields a quantitative regime classification, where each regime has a distinct delay law and seed-to-seed variability law. We further derive explicit delay laws for SGD and Adam-family optimizers by mapping their hyperparameters to the effective drift and diffusion of the slow coordinate. Our framework reconciles existing explanations as regime-specific limits and accounts for diverse grokking phenomena observed in prior work. Code will be publicly available at `https://github.com`.

1 INTRODUCTION

Grokking was first observed by Power et al. (2022) in modular arithmetic: a network rapidly memorizes the training set, remains near random test accuracy for many steps, and then abruptly generalizes. Since then, many mechanisms have been proposed, including *weight-norm decay* toward a generalizing shell Liu et al. (2022b), an *implicit-bias switch* from an early kernel predictor to a late min-norm or max-margin solution Lyu et al. (2024), *sparse–dense competition* between memorizing and generalizing subnetworks Merrill et al. (2023), a *lazy-to-rich transition* from fixed features to feature learning Kumar et al. (2024), and *progressive circuit formation and cleanup* Nanda et al. (2023); Chughtai et al. (2023). These explanations identify important mechanisms, but they are often setting-specific and sometimes point to different or seemingly conflicting causes. A unified quantitative framework is therefore needed to compare these findings, predict grokking delays, and explain why different experiments exhibit different signatures.

We develop such a framework by modeling grokking as slow transport along an effective one-dimensional coordinate. Once fast training modes have relaxed, progress toward generalization is governed by a slow coordinate θ. The grokking delay is the first-passage time for θ to reach the generalizing region, observed as the validation loss drops to its plateau. This converts grokking from a qualitative learning-curve phenomenon into a quantitative kinetic problem.

Our framework derives the governing delay equation and identifies three kinetic regimes: an escaping ridge, a flat stretch, and an activated barrier. Each regime has its own delay law, seed-to-seed variability, and hyperparameter dependence. This regime view reconciles prior observations: apparently inconsistent findings can arise because experiments lie in different kinetic regimes, or because optimizers induce different effective drift and diffusion. Experiments on representative settings from prior grokking studies show that these cases fit naturally into our classification and validate the predicted delay scaling, achieving a median error of just 3% in transfer prediction experiments.

Contributions. **(1)** We formulate grokking as first-passage transport along a slow coordinate and derive three kinetic regimes with closed-form delay and seed-variance laws. **(2)** We map SGD and Adam-family hyperparameters to effective kinetic quantities, yielding explicit delay laws and clarifying when hyperparameters change the regime versus only the timescale. **(3)** We analyze representative grokking settings from prior work and show that diverse observations are explained by their locations in our kinetic regime diagram, with quantitative experimental validation.

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.

2 GROKKING DELAY AND REGIMES

Referring to the typical grokking trajectory in Fig. 1, let s denote an arbitrary time point before the onset of the next grokking event, and let e denote the time at which the model generalizes. Grokking is governed by the evolution of the model parameters from the initial state θ_s to the generalized state θ_e. Denoting the grokking delay by N^*, we define the corresponding governing function as

$$\mathbb{E}[N^*] := f(\theta_s \to \theta_e). \tag{1}$$

The choice of s is flexible. In most prior studies, experiments focus on the first grokking event, corresponding to $s = 0$. This definition allows the delay to be quantified and makes it possible to study its extrema, together with the values of the independent variables at which they are attained.

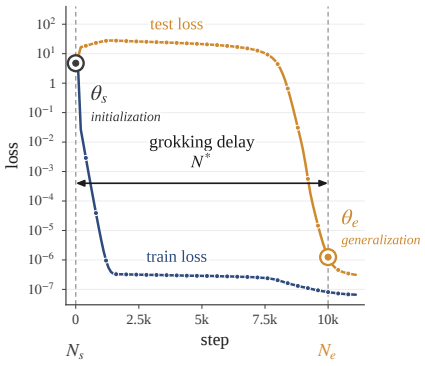


Figure 1: A typical grokking trajectory.

2.1 THE GOVERNING FUNCTION OF THE DELAY

Evolution statistics. Let us begin with the simplest setting: a one-dimensional parameter space optimized by SGD. We will later extend the resulting conclusions to higher-dimensional parameter spaces and other optimizers. At step t the parameter is updated by $\theta_{t+1} = \theta_t - \eta\,\hat{g}_t$ with learning rate $\eta > 0$ and minibatch gradient $\hat{g}_t = B^{-1}\sum_{i=1}^{B}\partial_\theta \ell_i(\theta_t)$. For the per-step displacement $\Delta_t := \theta_{t+1} - \theta_t$, the first two moments are

$$\mu(\theta) := \mathbb{E}[\Delta_t \mid \theta_t = \theta] = -\eta\,\partial_\theta L(\theta), \quad D(\theta) := \tfrac{1}{2}\mathrm{Var}[\Delta_t \mid \theta_t = \theta] = \eta^2 c(\theta)/(2B), \tag{2}$$

where L is the population loss and $c(\theta) := \mathrm{Var}_i[\partial_\theta \ell_i(\theta)]$ the per-example gradient variance. We refer to μ as the **drift** and D the **diffusion**. Third and higher cumulants are $O(\eta^3)$ and further suppressed by B^{-2} (Appendix D.1).

Governing Function of Expected Delay. Let $f(\theta)$ be the expected number of steps to first reach θ_e from θ. This gives $f(\theta) = 1 + \mathbb{E}[f(\theta + \Delta)]$. Taylor-expanding to second order and substituting $\mathbb{E}[\Delta] = \mu(\theta)$, $\mathbb{E}[\Delta^2] = 2D(\theta)$[1] yields $\mu(\theta)f'(\theta) + D(\theta)f''(\theta) = -1$, with $f(\theta_e) = 0$ and $f'(\theta_a) = 0$ at a reflecting lower boundary θ_a. Solving with integrating factor $e^{-\Phi(\theta)}$ yields

$$\mathbb{E}[N^*] \;:=\; f(\theta_s \to \theta_e) = \int_{\theta_s}^{\theta_e} \mathrm{d}x\, e^{\Phi(x)} \int_{\theta_a}^{x} \mathrm{d}y\, \frac{e^{-\Phi(y)}}{D(y)}, \quad \Phi(\theta) := -\int_0^{\theta} \frac{\mu(s)}{D(s)}\,\mathrm{d}s. \tag{3}$$

Eq. (3) quantifies the contribution of each point along the path from θ_s to θ_e to the overall crossing difficulty. The analyses in the following sections are centered on this governing function; its correspondence with standard results of non-equilibrium statistical physics is listed in Appendix H.

2.2 THREE REGIMES CLASSIFIED BY THE SHAPE OF Φ

Shape of Φ. From Eq. (3), the delay depends on the bottleneck only through Φ, which encodes the combined effects of (μ, D), together with a prefactor $1/D$. Take constant $D_0 := D(\theta^*)$, the diffusion at the bottleneck θ^*, Eq. (3) becomes $\mathbb{E}[N^*] = \frac{1}{D_0}\int_{\theta_s}^{\theta_e}\mathrm{d}x \int^{x}\mathrm{d}y\, e^{\Phi(x)-\Phi(y)-\ln(D/D_0)}$. The residual dependence on D appears only through the logarithmic term $\ln(D/D_0)$, whose variation across a stage is much smaller than that of Φ (Appendix D.3). We therefore neglect this term and, from here on, write D for D_0. Φ then becomes the sole determinant of Eq. (3), and it enters only through differences $\Phi(x) - \Phi(y)$. Consequently, the delay depends on the *shape* of Φ rather than its value. Since $\Phi(x) - \Phi(y)$ appears in the exponent, the double integral is dominated by the pair (y, x) that maximizes this difference. The location of this dominant pair, and therefore the functional form of the delay, is determined by the shape of Φ across the stage.

[1] We neglect the $\mu(\theta)^2$ contribution to $\mathbb{E}[\Delta^2]$ since $\mu(\theta)^2/2 \ll D(\theta)$ near a plateau.

Shape Variable. The delay primarily accumulates in regions where the drift is weakest, i.e., near the plateau. Let θ^* denote such a point, and define the stage width as $\ell := \theta_e - \theta^*$. A smooth drift near θ^* is approximated as $\mu(\theta) = v + \lambda(\theta - \theta^*) + \dots$, where $v := \mu(\theta^*)$ is the residual drift and $\lambda := \mu'(\theta^*)$ measures how rapidly the drift changes along the stage. Since $\Phi' = -\mu/D$, the resulting potential Φ takes the form of a tilted parabola: $\Phi = \Phi(\theta^*) - \Psi\, s - \frac{1}{2}\Theta\, s^2$, where $s := (\theta - \theta^*)/\ell$ in units of the stage width. Here, the **tilt variable** $\Psi := v\ell/D$ captures the change in Φ across the stage induced by its slope, while the **curvature variable** $\Theta := \lambda\ell^2/D$ captures the corresponding change induced by curvature.

Since the slope of Φ is determined by the drift, the shape of Φ reflects the drift profile along the stage. In particular, $\mu(\theta_e)/\mu(\theta^*) = 1 + \Theta/\Psi$. This motivates defining the natural shape variable as the ratio between curvature and tilt:

$$\Omega := \frac{\Theta}{\Psi} = \frac{\lambda\ell}{v} = \frac{\mu(\theta_e)}{\mu(\theta^*)} - 1. \tag{4}$$

Thus, Ω measures the fractional change in the drift across the stage; notably, the diffusion D cancels out. A parabolic profile yields exactly three qualitative shapes, determined by Ω[2]: 1) the slope remains approximately constant, when $|\Omega| \lesssim 1$ and an effectively linear profile; 2) the slope steepens, when $\Omega \gg 1$ and a downward-opening parabola; 3) the slope reverses before the exit, when $\Omega < -1$ and an upward-opening parabola whose minimum occurs at $s = 1/|\Omega|$).

Three regimes. These three shapes correspond precisely to the three distinct ways of evaluating Eq. (3). When $\Omega > -1$, Φ decreases monotonically, so the factor $e^{\Phi(x)-\Phi(y)}$ is significant only when y lies close to x. The inner integral is then local, $\int^x e^{-\Phi(y)}\mathrm{d}y/D \approx e^{-\Phi(x)}/\mu(x)$, and the expected delay reduces to the deterministic transit time induced by the drift,

$$\mathbb{E}[N^*] = \int_{\theta^*}^{\theta_e} \frac{\mathrm{d}\theta}{\mu(\theta)}, \tag{5}$$

with stochastic corrections entering only at relative order $1/\Psi$. In contrast, when $\Omega < -1$, the maximum of $\Phi(x) - \Phi(y)$ is attained with y at the minimum of Φ and x at its maximum. Thus, $\mathbb{E}[N^*] \propto e^{\Delta\Phi}$ where $\Delta\Phi$ is the height of the rise between these two points. In this regime, the delay grows exponentially with the barrier height, and crossing is possible only through noise.

This classification yields three regimes. **Regime A**, an *escaping ridge* ($\Omega \gg 1$), where the drift amplifies across the stage; **Regime B**, a *flat stretch* ($|\Omega| \lesssim 1$), where the drift is effectively constant and carries the run at speed v; **Regime C**, an *activated barrier* ($\Omega < -1$), where the drift reverses before the exit, so only noise can cross. Figure 2 illustrates these three geometries.

Reinterpretation of Previous Studies. Surprisingly, when we reproduced experiments from 12 grokking studies in the literature and computed the corresponding Ω values, we found that their grokking trajectories evolve in different regimes. As we will show later, this new perspective helps explain and reinterpret the findings reported in those studies (Sec. 4 & Appendix C).

2.3 Delay Laws by Regime

We now evaluate Eq. (3) in each regime using the parabolic approximation, keeping the leading noise correction. The resulting delay laws, summarized in Table 1, give the characteristic scaling in each case. All formulas are written in terms of the plateau quantities (v, λ, D, ℓ), measured before the escape event. Seed-to-seed variability, quantified by the coefficient of variation CV, is derived from the generalized governing function in Appendix D.2. The derivations are given below.

Regime A (escaping ridge, $\Omega \gg 1$). Let the linear drift vanish at the ridge top $\theta_r := \theta^* - u_0$ with $u_0 := v/\lambda$. Thus, by Eq. (4), we have $\Omega = \ell/u_0$: the entrance sits at offset u_0 from the top, and the ridge amplifies that offset. Writing $x := \theta - \theta_r$, the potential becomes $\Phi(x) = -x^2/(2\ell_T^2), \ell_T := \sqrt{D/\lambda}$, where ℓ_T is the radius of the *noise ball*, inside which noise dominates over the drift. The

[2]Comparing Θ with Ψ presumes $\Psi \gtrsim 1$, i.e. the drift moves the run further than noise does; otherwise Θ is compared with 1. In every grokking setting with sampling noise we measured, $\Psi \geq 16$, and $\Psi \geq 3.5 \times 10^2$ on all but sparse parity (Appendix A.9).

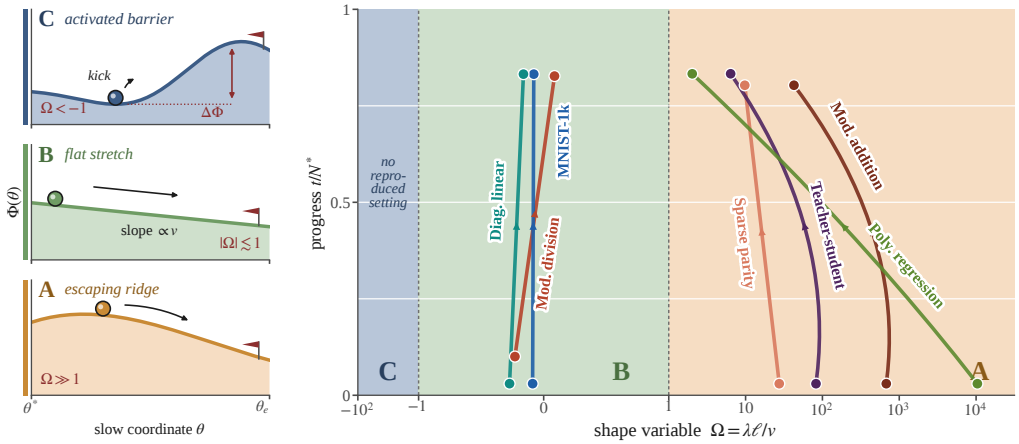


Figure 2: Three-regime overview. The regime map along Ω, where we place all prior literature experiment settings on this coordinate (per-setting values in Appendix B.2). Schematic diagrams for the shape of Φ in Regimes A, B, C.

reflecting boundary sits at the top, since the linearized drift is odd around this point, the problem reduces to the first passage of $|x|$ to $\ell + u_0$. Rescaling $w := x/\ell_T$, we obtain

$$\mathbb{E}[N_A^*] = \frac{1}{\lambda} \int_{u_0/\ell_T}^{(\ell+u_0)/\ell_T} F(w)\, \mathrm{d}w, \qquad F(w) := e^{-w^2/2} \int_0^w e^{t^2/2}\, \mathrm{d}t. \tag{6}$$

Deterministic transport dominates when the entrance lies outside the noise ball, $u_0 \gg \ell_T$, i.e. $\Psi \gg \Omega$ (since $u_0^2/\ell_T^2 = \Psi/\Omega$). Both limits of Eq. (6) then lie where $F(w) = w^{-1} + w^{-3} + O(w^{-5})$, and integrating gives, up to $O(\Theta^{-1})$,

$$\mathbb{E}[N_A^*] = \frac{1}{\lambda} \left[\ln(1+\Omega) + \frac{\Omega}{2\Psi} \right]. \tag{7}$$

The leading term is the transit time Eq. (5) of the linear drift, $\lambda^{-1} \ln(\ell/u_0)$ up to $O(1/\Omega)$: logarithmic in the initial offset. Diffusion enters the mean only at order $\Omega/\Psi = \ell_T^2/u_0^2$, but it jitters the effective offset, so $\mathrm{CV}_A \approx \sqrt{\sigma^2[\ln u_0] + \Omega/\Psi}\,/\ln(1+\Omega)$, which combines initialization spread and noise and is small when $\ell_T \ll u_0 \ll \ell$. If the drift vanishes to higher order at the top, $\mu \propto x^p$ with $p \geq 2$, the ridge is degenerate and Eq. (5) gives a power law u_0^{1-p} in place of the logarithm (Appendix D.5).

Regime B (flat stretch, $|\Omega| \lesssim 1$). With the curvature dropped, $\Phi = \Phi(\theta^*) - \Psi s$, and with the reflecting boundary at the entrance, Eq. (3) integrates exactly to the transit time at speed v as

$$\mathbb{E}[N_B^*] = \frac{\ell}{v} \left[1 - \frac{1 - e^{-\Psi}}{\Psi} \right] \simeq \frac{\ell}{v}, \tag{8}$$

where the wall shortens it by D/v^2. Keeping the curvature to first order multiplies Eq. (8) by $\ln(1+\Omega)/\Omega = 1 - \Omega/2 + \ldots$. Noise sets the spread: the first-passage time is asymptotically inverse Gaussian with $\mathrm{Var}[N^*] = 2D\ell/v^3$ (see Appendix D.6), so $\mathrm{CV}_B = \sqrt{2/\Psi}$.

Regime C (activated barrier, $\Omega < -1$). The drift reverses at the well $\theta_a = \theta^* + \ell/|\Omega|$, the minimum of Φ, and the stage exits beyond the barrier top θ_b, where $\mu(\theta_b) = 0$ and the drift turns forward again with slope $\lambda_b := \mu'(\theta_b) > 0$. The run relaxes into the well in $O(1/|\lambda|)$ steps and forgets θ_s. For $\Delta\Phi := \Phi(\theta_b) - \Phi(\theta_a) \gg 1$, the inner integral of (3) saturates at the Gaussian peak of $e^{-\Phi}$ around θ_a, where $\Phi'' = |\lambda|/D$, and the outer integral is dominated by the Gaussian peak of e^{Φ} around θ_b, where $\Phi'' = -\lambda_b/D$, giving

$$\mathbb{E}[N_C^*] \approx \frac{2\pi}{\sqrt{|\lambda|\,\lambda_b}}\, e^{\Delta\Phi}, \qquad \Delta\Phi = -\int_{\theta_a}^{\theta_b} \frac{\mu}{D}\, \mathrm{d}\theta = \Psi\, \frac{(|\Omega|-1)^2}{2|\Omega|}, \tag{9}$$

Table 1: Expected delay, seed variability, and optimizer laws by regime. Tags: (N) $s \ll 1$; (S) $s \gg 1$; (wd) decay-driven drift on $\theta = \ln \rho$; (∇) gradient-driven drift. Derivations in Appendix E.7.

Regime	Condition	$\mathbb{E}[N^*]$	CV	SGD	AdamW
A	$\Omega \gg 1$	$\frac{1}{\lambda}\ln(1+\Omega)$	$\frac{\sqrt{\sigma^2[\ln u_0] + \Omega/\Psi}}{\ln(1+\Omega)}$	$\frac{1}{\eta\lvert\kappa_*\rvert}\ln(1+\Omega)$	$\frac{1}{\eta\lvert\kappa_*^A\rvert}\ln(1+\Omega)$ (N); $\frac{\ell}{\eta}$ (S)
B	$\lvert\Omega\rvert \lesssim 1$	$\frac{\ell}{v}$	$\sqrt{2/\Psi}$	$\frac{1}{\eta\lambda_{\mathrm{wd}}}\ln\frac{\rho_0}{\rho_*}$ (wd); $\frac{\ell}{\eta\lvert\partial_\theta L\rvert}$ (∇)	$\frac{1}{\eta\lambda_{\mathrm{wd}}}\ln\frac{\rho_0}{\rho_*}$ (wd); $\frac{\ell\sqrt{1+s^2}}{\eta s}$ (∇)
C	$\Omega < -1$	$\frac{2\pi\, e^{\Delta\Phi}}{\sqrt{\lvert\lambda\rvert\lambda_b}}$	1	$\frac{2\pi}{\eta\sqrt{\kappa_a\lvert\kappa_b\rvert}}\exp\!\left[\frac{2B}{\eta c}\,\Delta\big(L + \frac{1}{2}\lambda_{\mathrm{wd}}\theta^2\big)\right]$	$\frac{2\pi}{\eta\sqrt{\kappa_a^A\lvert\kappa_b^A\rvert}}\exp\!\left[\frac{2}{\eta}\,\Delta\big(\sqrt{B/c}\,L + \frac{1}{2}\lambda_{\mathrm{wd}}\theta^2\big)\right]$ (N)

the last equality on the parabola; the exit θ_e lies past θ_b by more than $\sqrt{D/\lambda_b}$ (Appendix D.7). D survives only in the exponent. Relaxation is exponentially faster than escape, so the escape is memoryless: asymptotically exponential, with $\mathrm{CV}_C \to 1$ and constant hazard $1/\mathbb{E}[N_C^*]$.

2.4 From High-Dimensional Training to One-Dimensional Kinetics

After quotienting out exact network symmetries, high-dimensional training can be reduced to effective one-dimensional kinetics when there is a clear separation of timescales. Near a plateau, one direction evolves much more slowly than the others, while the fast modes rapidly relax to a conditional stationary distribution at fixed slow coordinate θ. Let $g \gg 1$ be the ratio between the slow relaxation time and the fastest remaining relaxation time. In the canonical grokking setting, the fast modes fit the training set: memorization corresponds to their relaxation onto the near-interpolation manifold, after which the slow coordinate moves along this manifold and controls the test readout.

Averaging the full-dimensional drift and diffusion over the fast equilibrium distribution yields effective one-dimensional coefficients $\mu_{\mathrm{eff}}(\theta)$, $D_{\mathrm{eff}}(\theta)$. The effective diffusion D_{eff} consists of the projected minibatch noise, plus an $O(1/g)$ correction arising from fast-mode fluctuations as they relax against the slow drift. Under the same separation condition $g \gg 1$ that justifies the reduction, this correction is negligible, and D_{eff} reduces to the projected-noise contribution used in Sec. 3. The full two-term expression, its $O(1/g)$ bound, and its validation on the constructed anchors are in Appendix D.9. Thus θ itself obeys an effective one-dimensional Markov diffusion, and the governing function, three-regime classification, and delay laws derived above all apply with (μ, D) interpreted as $(\mu_{\mathrm{eff}}, D_{\mathrm{eff}})$.

Multi-stage delays. When multiple directions are comparably slow, training progresses through a sequence of stages. Each stage is governed by one of the three regimes, and the total delay is the sum of the individual stage delays. For independent stages, $\mathrm{CV}_{\mathrm{total}} = \sqrt{\sum_j \mathbb{E}[N_j^*]^2 \mathrm{CV}_j^2} / \sum_j \mathbb{E}[N_j^*]$, so a short stage with high intrinsic variability may contribute little to the overall spread.

3 Optimizer Laws

Every law in Sec. 2.3 depends on the optimizer only through the effective pair (μ, D) it induces. Under the diffusion approximation, two optimizers that generate the same (μ, D) on the same landscape therefore have the same delay distribution. A hyperparameter can affect the delay only by changing v, λ, D, or the underlying landscape. Two general consequences apply to all optimizers considered below. First, the learning rate η rescales the entire drift, so the drift-contrast parameter $\Omega = \lambda\ell/v$ is independent of η. Thus the learning rate sets the timescale but does not determine the regime. Second, if $\mu/D = -U'/T$ with U independent of hyperparameters, then $\Phi = U/T$. In this case, the hyperparameters enter the exponential factor in Eq. (3), i.e. Ψ and Θ, through a single temperature T. Below, we specify (μ, D) for SGD and AdamW. These expressions are then substituted into the regime laws summarized in Table 1.

3.1 SGD

For SGD with decoupled weight decay, $\theta_{n+1} = \theta_n - \eta\hat{g}_n - \eta\lambda_{\mathrm{wd}}\theta_n$ induces

$$\mu = -\eta\,(\partial_\theta L + \lambda_{\mathrm{wd}}\theta), \qquad D = \frac{\eta^2 c}{2B}, \qquad T_{\mathrm{SGD}} = \frac{\eta}{B}, \quad U' = \frac{2(\partial_\theta L + \lambda_{\mathrm{wd}}\theta)}{c}. \tag{10}$$

The batch size B enters only through D, so for SGD Ω depends on neither η nor B: the regime is set by the landscape alone. In the drift-limited regimes A and B the leading-order delay scales as $1/\eta$, while the noise parameters satisfy $\Psi, \Theta \propto B/\eta$, so η and B trade off at fixed η/B only through the exponential factor. In regime C the prefactor of Eq. (9) is $2\pi/(\eta\sqrt{\kappa_a|\kappa_b|}) \propto 1/\eta$, which gives $\ln \mathbb{E}[N_C^*] = \Delta U\, B/\eta - \ln \eta + \text{const}$ with $\Delta U := \Delta\Phi\, \eta/B$. Heavy-ball momentum with coefficient β is SGD at learning rate $\eta/(1-\beta)$ in the gradient terms, and decoupled decay is not amplified (Appendix E.2).

When weight decay dominates the radial drift, the appropriate slow coordinate is $\theta' = \ln \rho$. On this coordinate weight decay induces a constant drift $v = \eta\lambda_{\text{wd}}$ with $\Omega = 0$ over $\ell = \ln(\rho_0/\rho_*)$, so the dynamics fall in regime B with $\mathbb{E}[N_B^*] = \frac{1}{\eta\lambda_{\text{wd}}} \ln \frac{\rho_0}{\rho_*}$. The logarithm appears because weight decay is uniform in $\ln \rho$, not in ρ.

3.2 AdamW

The Adam moment buffers equilibrate on timescales $\tau_{1,2} = (1-\beta_{1,2})^{-1} \ll N^*$ and can therefore be adiabatically eliminated. Because the first moment is an exponential moving average with unit gain at zero frequency, so β_1 does not affect the long-time effective drift μ and D. On a plateau $\hat{v}_i \approx (\partial_i L)^2 + c_i/B$, the dynamics are controlled by the per-coordinate signal-to-noise ratio

$$s_i := \frac{\sqrt{B}\,|\partial_i L|}{\sqrt{c_i}}, \qquad \sqrt{\hat{v}_i} = \sqrt{c_i/B}\,\sqrt{1+s_i^2}, \qquad P_i := \left(\sqrt{\hat{v}_i} + \varepsilon\right)^{-1}. \tag{11}$$

When $\varepsilon \ll \sqrt{\hat{v}_i}$, the induced coefficients are

$$\mu_i = -\,\eta\,\frac{s_i}{\sqrt{1+s_i^2}}\,\text{sgn}(\partial_i L) \;-\; \eta\lambda_{\text{wd}}\theta_i, \qquad D_i = \frac{\eta^2}{2\,(1+s_i^2)}. \tag{12}$$

The effective slow coordinate obtains its drift and diffusion by projecting these coefficients onto the slow direction. The relevant limiting behavior is determined by two measurable quantities, s_i and $\sqrt{\hat{v}_i}/\varepsilon$; derivations, corrections, and further discussion are given in Appendix E.3:

(N) Noise-dominated, $s_i \ll 1$ **and** $\sqrt{\hat{v}_i} \gg \varepsilon$**:** $\mu_i = -\eta\sqrt{B/c_i}\,\partial_i L - \eta\lambda_{\text{wd}}\theta_i$ and $D_i = \eta^2/2$. The preconditioner cancels the scale of the gradient noise, making D independent of B, c_i, and the loss. Consequently, $\Psi, \Theta \propto 1/\eta$ with no batch-size dependence when decoupled decay supplies the drift, and $\propto \sqrt{B}/\eta$ when the gradient does (Appendix E.4).

(S) Signal-dominated, $s_i \gg 1$ **and** $\sqrt{\hat{v}_i} \gg \varepsilon$**:** $\mu_i = -\eta\,\text{sgn}(\partial_i L) - \eta\lambda_{\text{wd}}\theta_i$ and $D_i \approx \eta^2 c_i/\big(2B(\partial_i L)^2\big)$. AdamW reduces to sign descent with constant step speed η, so the induced drift has $\Omega = 0$ regardless of the curvature of the loss: even a ridge is traversed like a flat stretch, with $N^* \approx \ell/\eta$.

(E) Floor-dominated, $\sqrt{\hat{v}_i} \ll \varepsilon$**:** $\mu_i = -(\eta/\varepsilon)\,\partial_i L - \eta\lambda_{\text{wd}}\theta_i$ and $D_i = \eta^2 c_i/(2B\varepsilon^2)$. The preconditioner saturates at $1/\varepsilon$, so Adam is SGD at learning rate η/ε in its gradient term, while the decoupled decay is not amplified.

Decay coupling. Decoupled decay bypasses P_i, so when it supplies the drift, $\mathbb{E}[N_B^*]$ is identical to SGD in all three limits above. By contrast, coupled L_2 decay passes through P. On a memorized plateau $\partial_i L$ collapses while $\lambda_{\text{wd}}\theta_i$ does not, so $\sqrt{\hat{v}_i} \to \lambda_{\text{wd}}|\theta_i|$. Each coordinate then enters the signal-dominated regime, with $\mu_i = -\eta\,\text{sgn}(\theta_i)$. The norm consequently shrinks approximately linearly, at a rate bounded by $\eta\|\theta\|_1/\rho$, independent of λ_{wd}. AdamW instead produces exponential shrinkage at rate $\eta\lambda_{\text{wd}}$ (Appendix E.3).

4 Experiments

4.1 Existing Experimental Settings on Grokking

Recent literature attributes grokking to diverse mechanisms, broadly categorized into: (1) **weight dynamics and regularization**, including weight-norm decay (Liu et al., 2022b), implicit-bias

shifts (Lyu et al., 2024), and efficiency competition (Varma et al., 2023); (2) **feature and circuit evolution**, spanning lazy-to-rich transitions (Kumar et al., 2024), multiphase representation learning (Liu et al., 2022a), sparse-vs-dense subnetwork competition (Merrill et al., 2023), hidden Fourier signals (Barak et al., 2022), and progressive circuit cleanup (Nanda et al., 2023; Chughtai et al., 2023); (3) **statistical physics analogies**, such as stochastic escape (Kozyrev, 2026) and first-order state transitions (Rubin et al., 2024); and (4) **optimization and numerical anomalies**, notably slingshot effects (Thilak et al., 2022) and Softmax collapse (Prieto et al., 2025).

We reproduce experiments on 9 experimental settings from the literature above (Appendix A). We measured their Ω (Appendix A.6) and place them in regimes in our framework. Runs that do not grok fall into three non-grokking phases outside the regime map (Appendix F). Table 6 (in the Appendix) reports a summary of these reproduced settings. Figure 3 shows the grokking trajectory results.

4.2 Insights Beyond Prior Studies: Resolving Conflicts

Is weight decay the clock? Omnigrok attributes the delay to weight decay, which slowly shrinks an oversized weight norm until the network generalizes (Liu et al., 2022b; Varma et al., 2023). Kumar et al. (2024) reply that "weight norm and weight decay cannot explain grokking": a small MLP trained on modular addition with plain gradient descent groks with no decay, and its weight norm *grows* while it does. We reproduce both observations, each on its own task, and the two tasks turn out to be waiting for different reasons. When both hosts get the same weight decay per step, grokking on MNIST comes about seven times sooner, while on modular addition it does not come sooner at all (Fig. 4a). During the wait, the MNIST weight norm shrinks steadily, while the modular-addition norm grows by up to 70%, with or without decay (Fig. 4b). MNIST is in regime B: the wait is a long, steady trip of the weight norm, and weight decay is what drives that trip, so more decay makes it shorter. Modular addition is in regime A: the wait is the time a small generalizing component needs to grow, and it grows on its own, compounding like interest. Decay does not feed that growth, so it cannot shorten the wait, and ten times more decay collapses the network instead. Each account is right about its own regime and wrong only in extending its claim to all grokking.

Does gradient noise matter? On sparse parity, Barak et al. (2022) report that "with larger batch sizes, the models were slower to converge". The weight-norm theories give delay formulas with no noise in them, and Lyu et al. (2024) prove grokking for noiseless full-batch gradient descent, with the delay set only by initialization and weight decay. At a fixed learning rate, changing the batch size changes only the gradient noise, so a batch sweep tests noise directly (Fig. 4c). On parity, a $16\times$ larger batch makes grokking take 42% longer. On Lyu's diagonal linear network, every batch from 4 up to the full dataset groks at the same step, to within 1%. Only at batch sizes 1 and 2 does the noise save a few percent. Parity is in regime A: the generalizing component starts from a tiny seed that noise helps plant, so quieter noise leaves a smaller seed that needs longer to grow. The diagonal network is in regime B: the delay is the same steady trip of the weight norm as on MNIST, and noise only jitters around that path and averages out. Noise is part of the clock in one regime and irrelevant in the other.

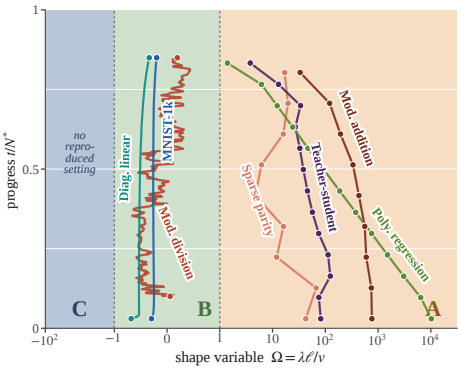


Figure 3: Exact grokking trajectories.

4.3 Transferring the Delay Laws Across Hyperparameters

To validate the transferability of our Grokking delay laws across hyperparameters, we vary one knob at a time on published settings: the target run supplies only its hyperparameters, while the remaining constants are estimated from a source setting on the same host, or a sibling's delay is rescaled a priori by the ratio prescribed by the law. Table 2 shows close alignment between the predicted and measured delays, with a median error of 3% (2.6% over all 23 transfers, Appendix B.4) across five task families, spanning GD, SGD, and AdamW. The predictions remain accurate under changes in weight decay, learning rate, batch size, and initialization scale, indicating that our framework

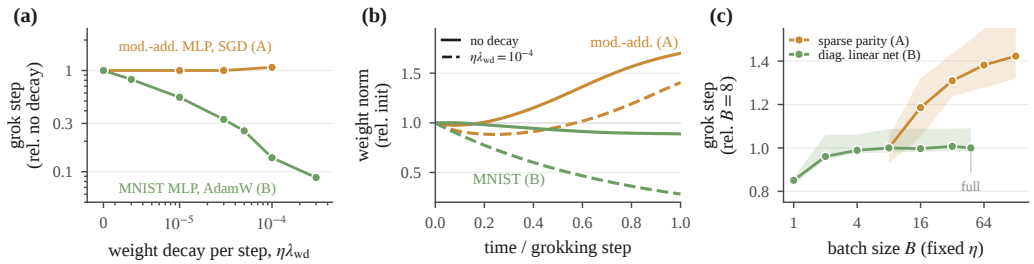


Figure 4: **Two published disagreements, resolved by regime.** (a) Grokking time vs. weight decay per step. (b) Weight norm dynamics with (dashed) and without (solid) weight decay. (c) Grokking time vs. batch size at a fixed learning rate. Bands show 95% intervals of the median.

Table 2: Delay laws transferred across hyperparameters. The regime is read from Ω on the host's slow coordinate ($\ln \|\theta\|$ for decay hosts). N^*: measured median delay. $\hat{N}$: prediction, with constants from the source setting, or (†) a priori, a sibling's delay rescaled by the law's ratio.

Setting	Transfer from	Optimizer	Regime	N^*	$\hat{N}$	$\hat{N}/N^*$
MNIST-1k, λ_{wd}=0.01	λ_{wd}=0, 0.003	AdamW	B	55,500	59,366	1.07
MNIST-1k, λ_{wd}=0.03	λ_{wd}=0, 0.003	AdamW	B	32,625	31,214	0.96
diag-net, $\lambda_{wd}=3\times10^{-3}$	$\lambda_{wd}=10^{-3}$	SGD	B	5,245	4,822†	0.92
diag-net, $\lambda_{wd}=10^{-2}$	$\lambda_{wd}=10^{-3}$	SGD	B	1,770	1,446†	0.82
diag-net, B=8, $\lambda_{wd}=10^{-2}$	B=8, $\lambda_{wd}=10^{-3}$	SGD	B	2,685	2,598†	0.97
diag-net, B=16, $\lambda_{wd}=10^{-3}$	B=8 (null in B)	SGD	B	26,320	25,975†	0.99
Power p=97, λ_{wd}=0.3	λ_{wd}=0.1	AdamW	B	8,000	7,900†	0.99
Power p=97, λ_{wd}=1	λ_{wd}=0.1	AdamW	B	2,400	2,370†	0.99
mod-add, $\eta=10^{-3}$	$\eta=5\times10^{-4}$	AdamW	A	9,050	8,175†	0.90
mod-add, $\eta=2\times10^{-3}$	$\eta=10^{-3}$	AdamW	A	3,200	4,525†	1.41
mod-add, $\eta=5\times10^{-4}$	$\eta=2\times10^{-3}$	AdamW	A	16,350	12,800†	0.78
mod-add, $\eta=10^{-3}$	$\eta=2\times10^{-3}$	AdamW	A	9,050	6,400†	0.71
mod-add, init scale 3	init scale 1	AdamW	A	8,400	8,616	1.03
poly GD, $x_s=10^{-3}$	$x_s=10^{-2}$	GD	A	940	926	0.99
poly GD, $x_s=10^{-4}$	$x_s=10^{-2}$	GD	A	1,320	1,292	0.98
poly GD, $x_s=10^{-5}$	$x_s=10^{-2}$	GD	A	1,700	1,658	0.98
poly GD, $x_s=10^{-6}$	$x_s=10^{-2}$	GD	A	2,080	2,024	0.97
poly GD, $x_s=10^{-7}$	$x_s=10^{-4}$	GD	A	2,460	2,454	1.00

captures their functional effects on the delay rather than fitting individual experimental settings. Reading a law forward on a single run, from plateau quantities measured at a pre-declared fraction of the delay, is evaluated in Appendix B.3.

Figure 5 further supports these findings at the trajectory level. We observe that panels (a) and (b) correspond to regime B: Ω remains within $|\Omega| \leq 1$, and the slow coordinate traverses a flat stretch with an approximately constant drift, producing comparatively gradual evolution over training steps. Panels (c) and (d) correspond to regime A: the trajectory remains near an escaping ridge for most of the plateau before the drift rapidly amplifies, yielding a long quiescent period followed by a sharp, seed-dependent change in loss and accuracy. Despite these distinct dynamics, the transferred predictions closely match the measured N^*.

4.4 Insights on Optimizers

The most-reproduced setting is Adam at its ε floor. We place every reproduced Adam setting in the limits of §3.2 from its stored buffers (Fig. 6a; Appendix A.8). Minibatch hosts sit in (N). Full-batch hosts, including Nanda et al. (2023)'s modular addition, sit in (E): there $\sqrt{\hat{v}_i} < \varepsilon$ on 99.9% of coordinates, so the plateau is gradient descent at rate η/ε plus decoupled decay. It is regime A, as GD would be ($\lambda > 0$ on 70/83 panels, $N^* \propto \eta^{-1.18}$). Removing decay at the floor lets the loss underflow and the drift $-(\eta/\varepsilon)\partial L$ vanish with it, and 0/8 seeds escape within 4×10^4 steps. Prieto

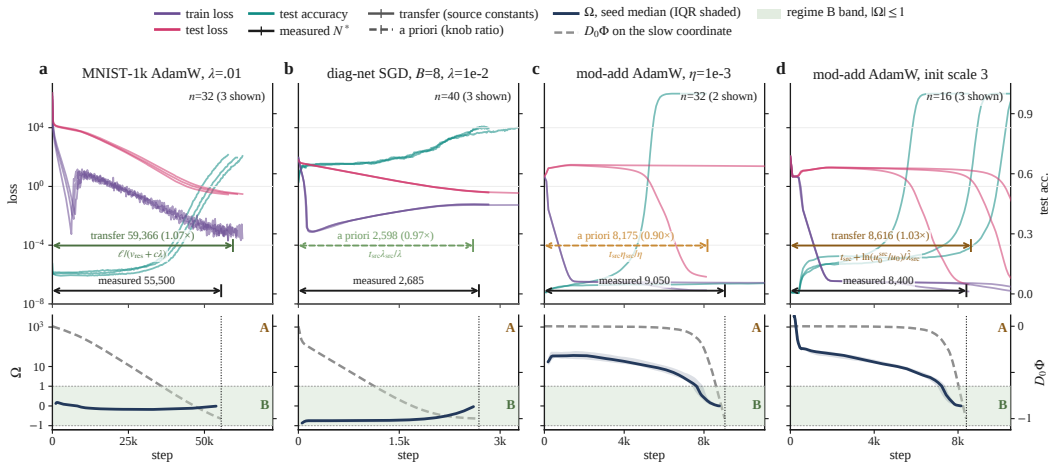


Figure 5: Transferred laws and associated Φ, Ω shapes. *Top:* Target run loss and accuracy. Arrows compare measured N^* with law predictions. *Bottom:* Runs on slow coordinates ($\ln \|\theta\|$ for (a), (b); embedding IPR for (c), (d)). Solid lines depict $\Omega(t)$; the shaded area denotes regime B ($|\Omega| \leq 1$).

et al. (2025)'s numerical freeze and Varma et al. (2023)'s "decay is necessary" are therefore one statement about this limit. Off it, grokking without decay is ordinary, as on MNIST-1k (§4.2).

Decay coupling changes the law, not the rate. At the same nominal $\eta = \lambda_{\text{wd}} = 10^{-3}$ on one host (Fig. 6b), decoupled decay is negligible: $e^{-\eta\lambda_{\text{wd}}t} = 0.996$ over the run, and $16/16$ seeds grok at a median of $2{,}100$ steps. Coupled L_2 decay instead becomes Adam's own normalizer, $\sqrt{\hat{v}_i} \approx \lambda_{\text{wd}}|\theta_i|$, which drives every coordinate into (S). The drift becomes sign descent, $\mu_i = -\eta\,\text{sgn}(\theta_i)$. It is constant along the run, $\Omega = 0$, and independent of λ_{wd}. The norm then falls linearly from $\approx$100 to 0.007, and $0/16$ seeds grok in 10^5 steps. A nominal λ_{wd} therefore does not transfer between AdamW and Adam+L_2.

5 Conclusion and Limitations

We analyze grokking from the first-passage escape process and show that its observed behaviors arise from three kinetic regimes governed by one measurable coordinate. Our framework derives the corresponding delay and seed-variability laws and connects these kinetics to common optimizers. Together, these results provide a unified, testable framework for analyzing grokking phenomena.

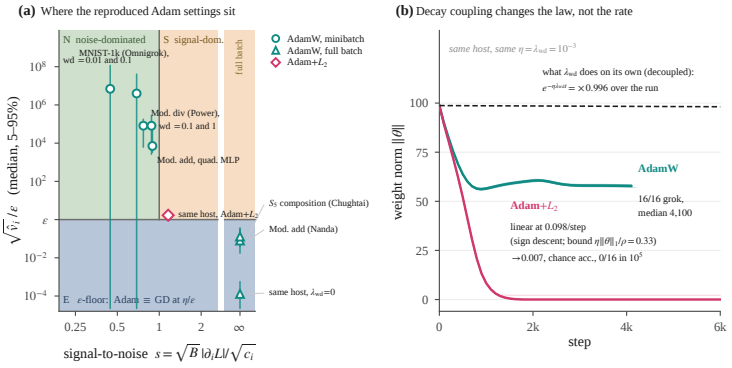


Figure 6: **(a)** Reproduced Adam settings placed in the limits of §3.2 by their stored buffers.**(b)** decoupled AdamW and coupled L2 comparison on the same setting.

Limitations. Existing grokking studies largely focus on small-scale settings, particularly toy algorithmic tasks. While this paper is scoped to analyze these existing grokking theories, it offers a foundation for extending to large models. A kinetic grokking theory of Large Models, with its whole additional theories, deserves its own paper, and it is an ongoing work of ours. Our theory also assumes near-Gaussian per-step increments; we monitor per-coordinate kurtosis throughout each run and apply the theory only when this assumption holds (see Appendix G).

AI USE STATEMENT

In this work, we used generative AI tools for paper editing, implementing methods from previous literature, assisting in the writing of proofs, and running experiments. We have not used generative AI tools for establishing core ideas or generating full paragraphs, and generating synthetic data sets or cleaning and reformating dataset, are not applicable to this work. Additionally, we used generative AI tools for editing illustrative figures with code and searching literature. We have reviewed all AI-assisted work. We want to emphasize especially that the training kinetics philosophy our paper wants to illustrate is completely original and comes from the author's experience with working on Thermodynamics and Fluid Dynamics. The author co-worked with the generative AI tool in drafting proofs, and all the claims have been verified by another co-author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments are conducted on a 4 $\times$ H20 GPU server. The code is available at *https://github.com* (anonymous github repository), with detailed guidance for reproducibility.

REFERENCES

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, pp. 6243–6267. PMLR, 2023.

Sergei Vladimirovich Kozyrev. How to explain grokking. *Computational Mathematics and Mathematical Physics*, 66(5):953–957, 2026.

Tanishq Kumar, Blake Bordelon, Samuel Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, volume 2024, pp. 23010–23035, 2024.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022a.

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022b.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Du, Jason Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, volume 2024, pp. 33897–33936, 2024.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Lucas Prieto, Melih Barsbey, Pedro Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *International Conference on Learning Representations*, volume 2025, pp. 81151–81168, 2025.

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, volume 2024, pp. 23881–23904, 2024.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

Vikrant Varma, Rohin Shah, Zachary Kenton, J'anos Kram'ar, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Table 3: Configuration of the reproduced settings. Architecture, data, loss and optimizer follow the cited study; one knob is swept per experiment (listed under "Sweep"). "Readout" is the quantity whose threshold crossing defines N^* (§A.2); the last column is the slow coordinate on which (v, D, ℓ) and Ω are measured (§A.3). $\varepsilon = 10^{-8}$ for every Adam run; SGD runs use no momentum.

#	Setting (source)	Task, data, model, loss	Optimizer	Sweep (seeds)	Readout	Slow coordinate
①	Modular addition (Nanda et al., 2023)	$a + b \bmod 113$, 30% of the 113^2 pairs for training; 1-layer transformer, d_{model}=128, 4 heads, MLP 512, ReLU, no LayerNorm; cross-entropy	full-batch AdamW, η=10^{-3}, β=(0.9, 0.98), λ_{wd}=1	$\eta \in \{0.5, 1, 2\}\times 10^{-3}$ (32/8/8); init scale $\{0.1, 0.3, 3\}$ (16); λ_{wd}=0 (8)	test acc. > 0.99	Fourier concentration of the embedding (IPR)
②	Modular arithmetic, Power config (Power et al., 2022)	$a - b \bmod 97$, 50% of pairs for training; 2-layer transformer, same widths; cross-entropy	AdamW, B=512, η=10^{-3}, β=(0.9, 0.98)	$\lambda_{\text{wd}} \in \{0.1, 0.3, 1, 3\}$ (32)	test acc. > 0.99	$\ln \|\theta\|$
③	Group composition (Chughtai et al., 2023)	S_5 composition, 40% of the 120^2 pairs; 1-layer transformer as ①; cross-entropy	full-batch AdamW, η=10^{-3}, λ_{wd}=1	— (32)	test acc. > 0.99	— (no seed exits)
④	MNIST-1k (Liu et al., 2022b)	1,000 training images, full test set; MLP $784{-}200{-}200{-}200{-}10$, ReLU, weights scaled $\times 8$ at init; MSE on one-hot targets	AdamW, B=200, η=10^{-3}, β=(0.9, 0.999); or SGD, B=200, η=10^{-2}, gradient clip 1	AdamW $\lambda_{\text{wd}} \in \{0, .003, .01, .03\}$ (32), $\{.05, .1, .3\}$ (16); SGD $\lambda_{\text{wd}} \in \{.003, .01, .03, .1\}$ (8)	test acc. > 0.6	$\ln \|\theta\|$
⑤	Sparse parity (Barak et al., 2022; Merrill et al., 2023)	$y = x_1x_2x_3$ on $x \in \{\pm 1\}^{40}$, pool of $2{\times}10^5$ examples; 1-hidden-layer ReLU MLP, width 1024; MSE	SGD, B=32, η=10^{-2}, no decay	$\eta \in \{.005, .01, .02\}$ (32); $B \in \{8, \ldots, 128\}$ (16)	test loss < 0.05	alignment with the support
⑥	Cubic teacher–student (Rubin et al., 2024)	$y = z^3 - 3z$, $z = w^* \cdot x$, $x \in \mathbb{R}^{32}$, 2,048 samples; student $\frac{1}{m}\sum_i (u_i \cdot x)^3$, m=64; MSE	full-batch GD, η=0.1	init scale $\{10^{-5}, \ldots, 10^{-2}\}$ (40)	teacher alignment > 0.9	teacher alignment
⑦	Diagonal linear net (Lyu et al., 2024)	$w = a \odot b$, n=64, 48 samples, logistic loss; sparse target (k=3) or dense target; init $\|a\|$, $\|b\|$ large ($40/\sqrt{n}$ per entry)	full-batch SGD, η=0.1; or minibatch $B \in \{8, 16\}$	$\lambda_{\text{wd}} \in \{.001, .003, .01\}$ (32 / 40); dense target (40)	test acc. > 0.82	$\ln \|\theta\|$
⑧	Polynomial regression (Kumar et al., 2024)	$y = (w^* \cdot x)^2 - 1$, $x \in \mathbb{R}^{32}$, 2,048 samples; student $\frac{1}{m}\sum_i (u_i \cdot x)^2$, m=64; MSE	full-batch GD, η=0.1	init scale $x_s \in \{10^{-9}, \ldots, 10^{-2}\}$ (32)	teacher alignment > 0.98	amplitude of the learned direction
⑨	Modular addition, quadratic MLP	$a + b \bmod 97$, 50% of pairs; $f = W_2(W_1x)^2$, width 512, init scale 4; MSE	AdamW or Adam+L_2, B=512, η=λ_{wd}=10^{-3}, β=(0.9, 0.98); SGD sibling: B=1024, η=0.3	decay coupling (16+16); SGD $\lambda_{\text{wd}} \in \{3.3, 10, 30\}\times 10^{-5}$ (8)	test acc. > 0.95	embedding IPR

A Reproduced Settings and Measurement Protocol

This appendix lists the nine settings we reproduce from the grokking literature, defines the delay N^* that every table and figure reports, and describes the instruments that read (v, λ, D, ℓ) and hence Ω off a running network. It ends with the two measured facts that §2.2 relies on: that the tilt Ψ is large on every setting, and that the diffusion varies too little across a stage to compete with Φ.

A.1 The nine settings

Table 3 gives the configuration of each setting. We keep the published architecture, data, loss and optimizer of the original study and change only what the experiment sweeps (one knob at a time). Each setting is referred to by its circled number throughout the appendices.

A.2 The delay N^*

N^* is the first step at which the readout of Table 3 crosses the threshold of Table 3 and stays there for three consecutive logged rows. Every run is stopped a fixed number of logging rows after the crossing, and the delay is the crossing step with that tail removed. All delays quoted in the paper are medians over the full seeded corpus of a cell, with the number of seeds n stated; seeds that do not cross within the budget are reported as censored and excluded from the median only when fewer than half are censored. The coefficient of variation CV is the seed standard deviation divided by the seed mean of the same delays. Because several settings share master seeds across runs, delays are de-duplicated by master seed before any statistic is formed.

A.3 The slow coordinate and the geometry rule

Each setting has a coordinate along which the plateau motion is slow and the readout eventually responds (last column of Table 3). On the decay-driven hosts (②, ④, ⑦) it is the log of the parameter norm, on which decoupled weight decay is a constant push. On the algorithmic hosts (①, ⑨) it is the concentration of the embedding on a few Fourier modes (its inverse participation ratio, IPR), which grows as the circuit forms. On the polynomial host (⑧) it is the amplitude of the learned direction,

which grows exponentially along the ridge; the teacher alignment saturates long before the exit and is not usable as a coordinate there. On the cubic host (⑥) and on parity (⑤) the alignment with the target is the coordinate.

The stage width ℓ is the distance, along this coordinate, from the current position to the value the coordinate takes at N^* (the median over seeds of that exit value, or each seed's own exit value where the dense panels allow it). The speed v is the finite-difference velocity of the same coordinate, and D is half the variance of its increments about that drift over a trailing window. One rule is not optional: *v and D must be measured in the same geometry as ℓ*. On hosts whose natural observable is a path length in parameter space, a drift block that reports a net displacement along one direction under-reads the speed by the ratio of path length to displacement, which on MNIST-1k reaches three orders of magnitude, and ℓ/v then over-predicts the delay by the same factor. Measuring the speed on the observable itself removes the error.

A.4 Removing the exact symmetries

Networks have exact continuous symmetries, such as rescaling the input weights of a ReLU neuron up and its output weights down, or rescaling the query and key matrices of an attention head in opposite directions, or shifting all logits of a softmax by a constant. Motion along these directions changes nothing and costs nothing, so they are flat by construction. Any curvature or drift measurement that does not remove them reads "flat" everywhere and misclassifies every plateau as regime B. Before every spectral, probe or drift measurement we assemble the generators of these symmetries at the current parameters (per-neuron rescalings for every ReLU or quadratic layer; per-head query–key and output–value rescalings and the logit-shift directions for the transformers; between 64 and 1,168 generators depending on the architecture), orthonormalize them, and project them out of every direction we measure along. If a generator cannot be enumerated for an architecture the measurement is refused rather than silently skipped.

A.5 The clone probe: reading λ

The local rate $\lambda = \mu'(\theta^*)$ tells how fast the drift changes along the slow coordinate; its sign separates a ridge ($\lambda > 0$) from a well ($\lambda < 0$). Rather than differentiate a model of the drift, we measure the response of the training dynamics itself. From a checkpoint we launch $2k+1$ clones ($k = 4$ to 6): the parent itself and a pair displaced by $\pm\delta$ along each of k symmetry-projected Hessian eigenvectors, half from the top and half from the bottom of the spectrum ($\delta = 10^{-3}$, scaled with the initialization on the small hosts). Every clone is run for a burst of 10^3 to 8×10^3 steps with the *same* minibatch sequence as the parent, and fit an exponential to the growth or decay of each clone's separation from the parent. Synchronized data order is essential: without it the separation is dominated by the different noise each clone sees, not by the dynamics. Under Adam each clone copies the full optimizer state and perturbs only the parameters.

The fitted rate is usable only inside a two-sided window. The burst must be long enough to resolve the rate, $|\lambda| \gtrsim 3/\text{burst}$, and short enough that the clones stay in the linear-response regime, burst $\lesssim N^* - t_{\text{probe}}$; beyond that the clones ride the parent's own escape and the fitted rate collapses toward half its value. A rate below the resolution floor is reported as *unresolved* with the floor as an upper bound on $|\lambda|$, never as evidence that $\lambda \approx 0$. Two further flags mark a panel as unusable: a response that is nonlinear in δ (which under Adam indicates the optimizer's own decorrelation rather than a rate), and a probe window that overlaps an instability burst (§A.7), where the response grows at the same rate along stiff and flat directions and measures the burst, not the landscape.

A.6 Ω: two estimators

With λ from the probe and (ℓ, v) from the drift block, $\Omega = \lambda\ell/|v|$. On hosts where the probe resolves λ along the slow coordinate (①, ⑤, ⑥, ⑧) this is the estimator used. On the decay hosts the probe's response is isotropic or burst-driven and does not act along $\ln\|\theta\|$; there we use the probe-free profile estimator that follows directly from Eq. (4), $\Omega(t) = v(N^*)/v(t) - 1$, the fractional change of the measured drift between the current panel and the exit. Where both are available we report both; disagreements are listed rather than averaged (Appendix B.2). Every Ω track is drawn

Table 4: Adam limits read from the stored buffers on the last pre-escape checkpoint. (E): $\sqrt{\hat{v}} < \varepsilon$ on most coordinates, Adam is gradient descent at η/ε. (N): noise-dominated, $s \ll 1$. (S): signal-dominated, sign descent.

Setting	Batch	$\sqrt{\hat{v}}/\varepsilon$ (median)	coords. below ε	s (median)	Limit
① Mod. addition, λ_{wd}=1	full	0.08	99.9%	—	(E)
① Mod. addition, λ_{wd}=0	full	1.3×10^{-4}	100.0%	—	(E)
③ Group composition	full	0.13	97.7%	—	(E)
② Mod. subtraction, λ_{wd}=0.1	minibatch	8.2×10^{4}	0.0%	0.77	(N)
② Mod. subtraction, λ_{wd}=1	minibatch	8.1×10^{4}	0.0%	0.88	(N)
④ MNIST-1k, λ_{wd}=0.01	minibatch	4.0×10^{6}	17.5%	0.69	(N)
④ MNIST-1k, λ_{wd}=0.1	minibatch	7.0×10^{6}	17.8%	0.44	(N)
⑨ Quad. MLP, AdamW	minibatch	7,108	0.0%	0.89	(N)
⑨ Quad. MLP, Adam+L_2	minibatch	1.7	0.0%	1.16	(S)

only to $0.8\,N^*$: beyond that the remaining distance $\ell \to 0$ and $\Omega \to 0$ by geometry, which is the end of the stage, not a change of regime.

A.7 Instability bursts

On the Adam hosts the training loss occasionally leaves its plateau value by many orders of magnitude and returns within a few hundred steps. We call such an excursion a burst when, after the training accuracy has first reached 0.99, the logged training loss rises by at least a factor of 100 between consecutive logging rows to a value of at least 10^{-4}; the burst lasts until the loss has fallen two decades below its peak. The counts are not sensitive to these thresholds at $\lambda_{\mathrm{wd}} \le 0.3$, where the plateau begins early; at $\lambda_{\mathrm{wd}} = 1$ memorization itself takes a third of the delay and the counts depend on where the plateau is taken to start, which we state wherever they are quoted. For each seed we record the number of bursts before N^*, whether N^* falls inside one, and the fraction of the pre-N^* plateau that bursts occupy, which is the chance level for the last count.

A.8 Reading Adam's limit from its buffers

Which of the three Adam limits of §3.2 a plateau is in is a measurement. From the last checkpoint before the escape we take the stored second-moment buffer, correct its bias ($\hat{v} = v/(1-\beta_2^t)$), and form two per-coordinate numbers: $\sqrt{\hat{v}_i}/\varepsilon$, whose median and below-one fraction decide between the floor (E) and the other two limits, and the signal ratio $s_i = \sqrt{B}|\partial_i L|/\sqrt{c_i}$, from the gradient and the per-example gradient variance on a probe batch, which decides between (N) and (S). Reading the *last* checkpoint of a run instead gives post-grokking values and inverts the verdict on the minibatch hosts, so the pre-escape checkpoint is not optional. Table 4 lists the result for every Adam host.

A.9 Two measured facts behind §2.2

Ψ is large. The three shapes of §2.2 are compared through $\Omega = \Theta/\Psi$ on the assumption that the drift carries the run further than the noise does, $\Psi = v\ell/D \gtrsim 1$. Table 5 gives Ψ on every setting where a diffusion can be measured. On every decay-transport host the run-median Ψ is at least 3.8×10^3, and the smallest value on any single panel where the drift points toward the exit is 359 (minibatch diagonal net) and 792 (MNIST-1k, AdamW). The one thermal ridge host, sparse parity, is the exception: its escape coordinate is noisy enough that the measured drift points away from the exit on a third of its panels, and where it points toward the exit Ψ falls to 16, still an order of magnitude above one. On the full-batch settings (①, ③, ⑥, ⑧, and the full-batch arm of ⑦) there is no sampling noise; the variance of the increments there measures deterministic variation of the step, not diffusion, so D is unavailable rather than small, and Ψ is formally infinite. These are the settings on which regime C is excluded by derivation rather than by measurement.

$\ln D$ varies little across a stage. Replacing D by its value at the bottleneck (§2.2) drops a term $\ln(D/D_0)$ from the exponent of Eq. (3). The same table lists the range of $\ln D$ over the plateau panels of each run next to the drop of Φ over the same panels. On every thermal run the range of $\ln D$ across the plateau is at most 17, against a drop of Φ of 10^4 or more; the ratio $\Delta \ln D/|\Delta\Phi|$ is below 1.1×10^{-3} on every host except sparse parity, where it reaches 0.064. The largest variation

Table 5: Tilt and diffusion variation on the reproduced settings, over the plateau panels ($t \leq 0.8N^*$) of the dense-trajectory seeds. $\Psi = v\ell/D$ (median and minimum over panels); $\Delta \ln D$ is the range of $\ln D$ across the panels of one run (median over seeds); $|\Delta\Phi|$ the drop of Φ across the same panels. Full-batch settings have no sampling noise and D is unavailable there (D n/a). †Sparse parity: run medians over panels where the drift points toward the exit; on 36% of its panels the measured drift points away and $\Psi < 0$.

Setting	Optimizer	cells / runs	Ψ (run medians)	$\min \Psi$	$\Delta \ln D$	$\lvert\Delta\Phi\rvert$
① Mod. addition	full-batch AdamW	4 / 8	D n/a	—	—	—
② Mod. subtraction (Power)	AdamW, B=512	4 / 8	3,837–2.4×10^6	2,407	0.2–14.2	9,734–1.9×10^6
③ Group composition	full-batch AdamW	1 / 2	D n/a	—	—	—
④ MNIST-1k	AdamW, B=200	7 / 21	4.0×10^4–1.3×10^9	792	4.6–9.4	2.2×10^5–4.8×10^9
④ MNIST-1k	SGD+clip, B=200	4 / 11	2.2×10^6–2.1×10^8	5.2×10^4	1.5–10.3	2.4×10^6–4.4×10^8
⑤ Sparse parity	SGD, B=32	8 / 16	1.6×10^1–7.0×10^4 †	16	0.6–9.5	22–1.1×10^4
⑥ Cubic teacher–student	full-batch GD	4 / 12	D n/a	—	—	—
⑦ Diag. linear net	full-batch SGD	4 / 12	D n/a	—	—	—
⑦ Diag. linear net	SGD, B=8, 16	6 / 18	2.5×10^5–2.6×10^8	359	3.3–16.7	8.5×10^5–6.1×10^8
⑧ Polynomial regression	full-batch GD	5 / 15	D n/a	—	—	—
⑨ Mod. addition, quad. MLP	AdamW, B=512	1 / 2	4.1×10^4–1.1×10^5	2.0×10^4	2.3–3.0	2.1×10^4–3.9×10^4
⑨ same, Adam+L_2	Adam+L_2, B=512	1 / 2	no exit	—	0.2–0.3	—

Table 6: Reproduced settings. "Limit" is the Adam limit read from the stored buffers (Appendix A.8); "Regime" is read from Ω on the slow coordinate of Table 3 (A: $\Omega \gg 1$; B: $|\Omega| \lesssim 1$); "Knob law" is the measured dependence of N^* on the swept knob with the law's exponent in parentheses; CV is the seed coefficient of variation of N^* (n seeds per cell). Full-batch settings are marked D n/a: they have no sampling noise and regime C is excluded by derivation.

#	Setting	Optimizer (limit)	Regime	Ω on the plateau	Knob law: measured (derived)	CV (n)
①	Mod. addition, 1-L transformer	full-batch AdamW (E)	A	763 → 33 on the embedding IPR; $\lambda > 0$ on 70/83 panels	$N^* \propto \eta^{-1.18}$ (-1); $\lambda_{\rm wd}$=0: 0/8 grok in 4×10^4	0.42 (32)
②	Mod. subtraction, 2-L transformer	AdamW B=512 (N)	B	-0.6 to 0.4 on $\ln \lVert\theta\rVert$ (probe unusable: bursts)	$N^* \propto \lambda_{\rm wd}^{-0.99}$ (-1); $\lambda_{\rm wd} N^* = 2{,}370/2{,}400/2{,}400$	0.24–0.27 (32)
③	S_5 composition	full-batch AdamW (E)	—	—	0/32 grok in 6×10^4 (budget-frozen)	—
④	MNIST-1k MLP	AdamW B=200 (N)	B	-0.4 to 0.5 on $\ln \lVert\theta\rVert$; probe A at $\lambda_{\rm wd}$=0.1 (disagreement)	$1/N^* = a + b\lambda_{\rm wd}$, R^2=0.993, $a \neq 0$; decay component $\propto \lambda_{\rm wd}^{-0.98}$ (-1)	0.010–0.071 (32/16)
	same host	SGD+clip B=200	B	-0.005 on $\ln \lVert\theta\rVert$; -0.24 to -0.75 on the path	$N^* \propto \lambda_{\rm wd}^{-0.75}$ over two channels (-1 for the decay channel)	0.055–0.096 (8)
⑤	Sparse parity, ReLU MLP	SGD B=32	A	5–67 where resolved (65% of panels at η=0.02); no decaying direction	+2,525 steps per doubling of B (R^2=0.93); η: 72,000 / 34,000 / 4,000	0.14–0.22 (16/32)
⑥	Cubic teacher–student	full-batch GD	A (degenerate)	82 → 13, > 5 on every panel	$N^* \propto u_0^{-0.995}$ (-1, p=2); 7/40 finish at 10^{-4}	0.18 (40)
⑦	Diagonal linear net	full-batch SGD	B	-0.4 to -0.15 on $\ln \lVert\theta\rVert$	$N^* \propto \lambda_{\rm wd}^{-0.91}$ (-1); a priori 0.92×, 0.82×	0.011–0.019 (32)
	same host	SGD B=8, 16	B	—	B=8 vs 16: 1.013× (null); $\lambda_{\rm wd}$×10: 0.97×	0.06–0.08 (40)
⑧	Polynomial regression	full-batch GD	A (linear)	$10^4 \to 6$, probe and profile agree	376 steps per decade of x_s ($\ln 10/\hat{\lambda}$ = 378)	0.004–0.009 (32)
⑨	Mod. addition, quadratic MLP	AdamW B=512 (N)	A	57 → 4 (below the probe floor)	16/16 grok, median 2,100	0.012 (16)
	same host	Adam+L_2 (S)	—	—	0/16 in 10^5; norm falls linearly at 0.095/step	—
	same host	SGD B=1024	B	—	$\lambda_{\rm wd}$×9 ⇒ N^*×1.12 (gradient channel)	— (8)

of D is on the SGD hosts with an MSE loss, where the per-example gradient variance shrinks with the training loss as the plateau deepens; even there it stays well inside the tilt.

B Additional Experimental Results

B.1 The reproduced settings at a glance

Table 6 summarizes every reproduced setting: the optimizer and, for Adam, the limit its buffers place it in; the regime read on its slow coordinate; the range of Ω on the plateau; the dependence of the delay on the knob each experiment sweeps, with the exponent the regime's law predicts in parentheses; and the seed-to-seed coefficient of variation. No reproduced setting is in regime C: every measured CV is below 0.5 (the largest, 0.42, is on the full-batch modular-addition host), no delay is exponential in a knob, and the full-batch settings have no noise with which to activate a barrier.

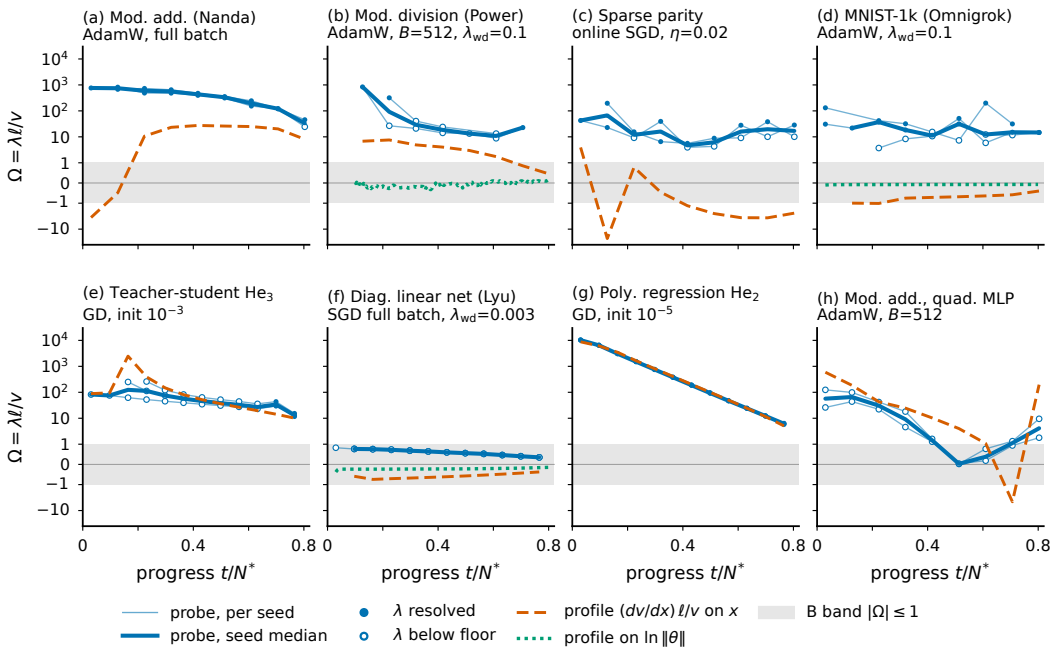


Figure 7: Ω along the plateau on every setting with dense panels. Thin lines: seeds; thick: seed median. Solid: probe estimator $\lambda\ell/|v|$; dashed: profile estimator $v(N^*)/v(t) - 1$. The shaded band is regime B, $|\Omega| \leq 1$; the axis is linear inside the band and logarithmic outside. Tracks end at $0.8N^*$, beyond which $\ell \to 0$ and $\Omega \to 0$ by geometry.

B.2 Ω ON EVERY SETTING

Figure 7 shows Ω along the plateau of every setting with dense gauge panels, seed by seed, and Table 7 lists the ranges. Two estimators are drawn where both exist (Appendix A.6). On the ridge hosts they agree. On modular addition (①) Ω falls from about 800 to about 30 as the run approaches the exit, with λ resolved on nearly every panel; on the polynomial host (⑧) it falls from 10^4 to 6 along a straight line on the log axis, the exponential growth of a linear ridge; on the cubic host (⑥) it stays above 5 for the whole plateau, which is what distinguishes a degenerate ridge from a flat stretch followed by a ridge. On the decay hosts the profile estimator on $\ln \|\theta\|$ reads $|\Omega| \leq 1$ throughout (②: -0.6 to 0.4; ④: -0.4 to 0.5; ⑦: -0.4 to -0.15), the flat stretch that the decay law requires; on the diagonal net the probe agrees (point estimate $0.77 \to 0.35$, unresolved but bounded by 3.8).

The disagreements are reported rather than hidden. On MNIST-1k at the largest decay ($\lambda_{\text{wd}} = 0.1$) the clone probe returns $\lambda > 0$ on half the panels, with Ω between 11 and 37, while the norm profile reads $\Omega = -0.1$ and the delay follows the decay law; the probe's response there grows at the same rate along stiff and flat directions, so it is not reading an instability along $\ln \|\theta\|$, and we label the host B. On the other MNIST-1k cells the probe is flagged optimizer-contaminated on three quarters of its panels and is not used. On modular subtraction (②) the probe-based Ω declines from about 800 early in the plateau to about 20 at $0.8N^*$ with only a quarter of its panels resolved, the unstable readings coincide with instability bursts (Appendix A.7), and the norm profile stays inside the band; the delay law is the decay law, so the host is labelled B with the early ridge on the circuit coordinate noted. On sparse parity the profile estimator is not interpretable, because the measured drift along the escape coordinate reverses sign on a third of the panels, and the probe is the only estimator. No reproduced setting crosses from one regime to another before its exit.

B.3 READING A LAW FORWARD ON A SINGLE RUN

The transfers of §4.3 take the law's constants from a different setting. A stricter and more useful test reads the law on the run itself, before it groks. We fix the observation point in advance at a quarter of the delay, take the regime from the clone probe and the drift block at that point, and evaluate the

Table 7: Ω on the plateau ($t \leq 0.8N^*$) by setting: probe estimator (range of the seed median; fraction of panels with λ resolved) and profile estimator, with the regime each implies.

Setting	Optimizer	Probe Ω (start$\to 0.8N^*$)	λ resolved	Profile Ω	Regime	Note
① Mod. addition, η=10^{-3}	AdamW, full	763 → 33	94%	x: A on 78% of panels	A	agree
① Mod. addition, η=5×10^{-4}	AdamW, full	700 → 35	100%	x: A on 78%	A	agree
① Mod. addition, η=2×10^{-3}	AdamW, full	1,400 → 43	94%	x: A on 89%	A	agree
② Mod. subtraction, λ_{wd}=0.1	AdamW, B=512	838 → 23 (bound $\leq$ 46)	27%	$\ln\|\theta\|$: 0.02 → 0.15; x: 6.8 → 0.44	B	probe unresolved; early ridge on x
② Mod. subtraction, λ_{wd}=0.3	AdamW, B=512	contaminated (16/18)	—	$\ln\|\theta\|$: 0.37 → −0.62	B	probe unusable
② Mod. subtraction, λ_{wd}=1	AdamW, B=512	contaminated (15/18)	—	$\ln\|\theta\|$: −0.59 → −0.21	B	probe unusable
④ MNIST-1k, λ_{wd}=0.003, 0.01, 0.03	AdamW, B=200	contaminated (74–88%)	—	$\ln\|\theta\|$: 0.46 → −0.14; 0.18 → −0.13; 0.09 → −0.04	B	probe unusable
④ MNIST-1k, λ_{wd}=0.1	AdamW, B=200	22 → 15 (range 11–37)	57%	$\ln\|\theta\|$: −0.10 → −0.07; x: −1.0 → −0.40	B	**disagree**: probe A
④ MNIST-1k, λ_{wd}=0.003–0.1	SGD+clip	0.2–4.2 (bounds 10–29)	0%	$\ln\|\theta\|$: −0.005; x: −0.24 to −0.75	B	probe unresolved
⑤ Sparse parity, η=0.02	SGD, B=32	43 → 17 (range 5–67)	65%	x: not interpretable (drift reverses on 36% of panels)	A	profile unusable
⑥ Cubic teacher–student, init 10^{-3}	GD, full	82 → 13 (> 5 on every panel)	19%	x: 89 → 10	A	agree; degenerate ridge
⑦ Diag. linear net, λ_{wd}=0.003	SGD, full	0.77 → 0.35 (bound $\leq$ 3.8)	0%	$\ln\|\theta\|$: −0.38 → −0.15; x: −0.59 → −0.37	B	agree
⑧ Polynomial regression, init 10^{-5}	GD, full	1.0×10^4 → 6.2	100%	x: 8.8×10^3 → 5.1	A	agree
⑨ Mod. addition, quad. MLP	AdamW, B=512	57 → 4 (bound $\leq$ 58)	0%	x: A on 89%	A	probe unresolved

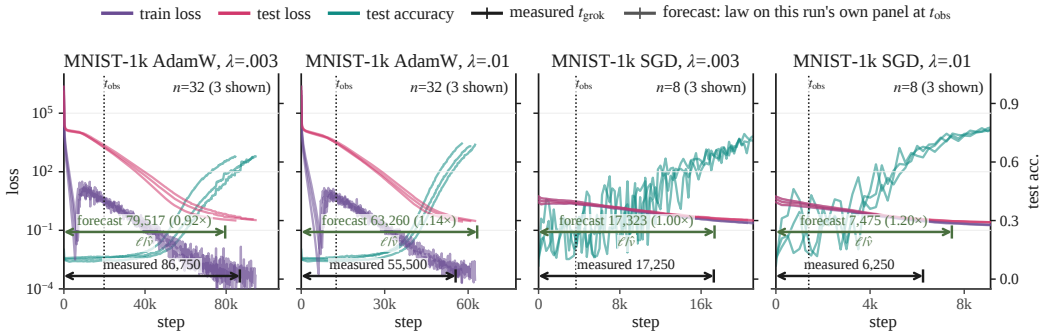


Figure 8: The flat-stretch law read forward on one run. Train loss, test loss and test accuracy for the seeds that carry dense panels; the black arrow is the measured N^* (median over all n seeds), the coloured arrow the law $\ell/\hat{v}$ evaluated on the run's own panel at t_{obs} (dotted), the nearest panel to a quarter of the delay.

regime's law with the quantities measured there: $\ell/\hat{v}$ in regime B, $\hat{\lambda}^{-1}\ln(\ell/u)$ in regime A with u the plateau-subtracted offset. Nothing is fitted, and the run's own delay is not used.

Figure 8 shows four MNIST-1k runs read this way (the nearest dense panel to a quarter of each seed's delay, 0.21–$0.23N^*$ of the full-corpus median): the flat-stretch law lands at $0.92\times$ and $1.14\times$ under AdamW and at $1.00\times$ and $1.20\times$ under minibatch SGD. Over the whole corpus the reading is less accurate than the transfers, and we report it as such. At a quarter of the delay, on the 43 dense cells (eight hosts, three optimizer families, delays from 560 to 108,125 steps), the median error is $2.5\times$ and 21 of 43 cells are within a factor of two. The two regimes fail differently. In regime B (31 cells) the reading is unbiased, with a median ratio $\hat{N}/N^*$ of 0.97 and a median error of $1.9\times$: the scatter is the drift measured on a trailing window of a noisy run. In regime A (12 cells) the reading is systematically early, with a median ratio of 0.29. This is not a failure of the functional form but of the observation point: on the ridge hosts the plateau is multi-stage, the linearized rate at a quarter of the delay belongs to an earlier and slower stage than the one that carries the run out, and the same law evaluated at 0.45–$0.6N^*$ reads $1.03\times$ on the seed median with 9 of 10 regime-A cells within a factor of two, with no change to its expression.

Figure 9 shows the full dependence on the observation point. Read at the first panel, three percent into the delay, the median error is $11\times$; it falls to $4.4\times$ at a tenth of the delay, crosses a factor of two near a third, and is $1.14\times$ past 80% once the speed is measured in the same geometry as ℓ. Every cell reaches a factor of two at some point of its plateau, so the forms are right; what a single-run forecast needs, and what the transfers of §4.3 supply by construction, is the constant of the rate-limiting stage.

B.4 All transfers

Table 2 in the main text lists the 23 transfers with a measured delay; their median error is 2.6%, and the two outside 25% are the learning-rate transfers on modular addition to and from $\eta = 2\times10^{-3}$ ($1.41\times$ and $0.71\times$), which rest on an 8-seed cell. Figure 10 shows eight of the transfers on their

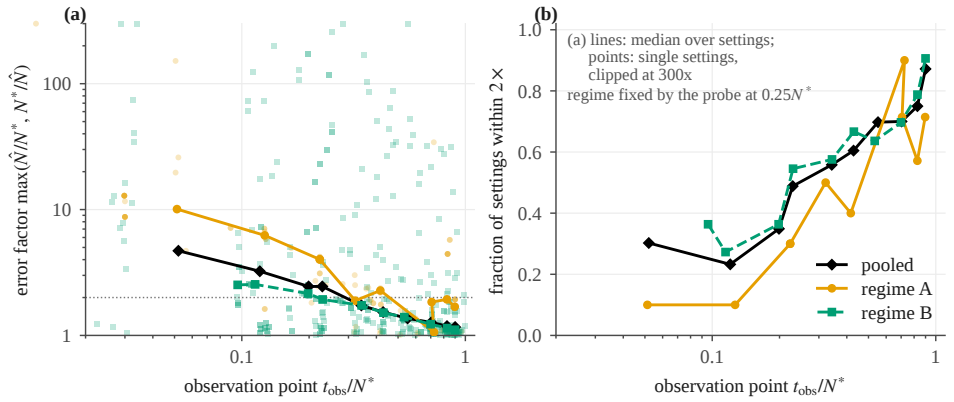


Figure 9: Forecast error against the observation point on the 43 dense cells. (a) Median error factor $\max(\hat{N}/N^*, N^*/\hat{N})$ over cells, pooled and by regime; light points are single cells. (b) Fraction of cells within a factor of two. The regime of each cell is the clone-probe verdict at a quarter of its delay.

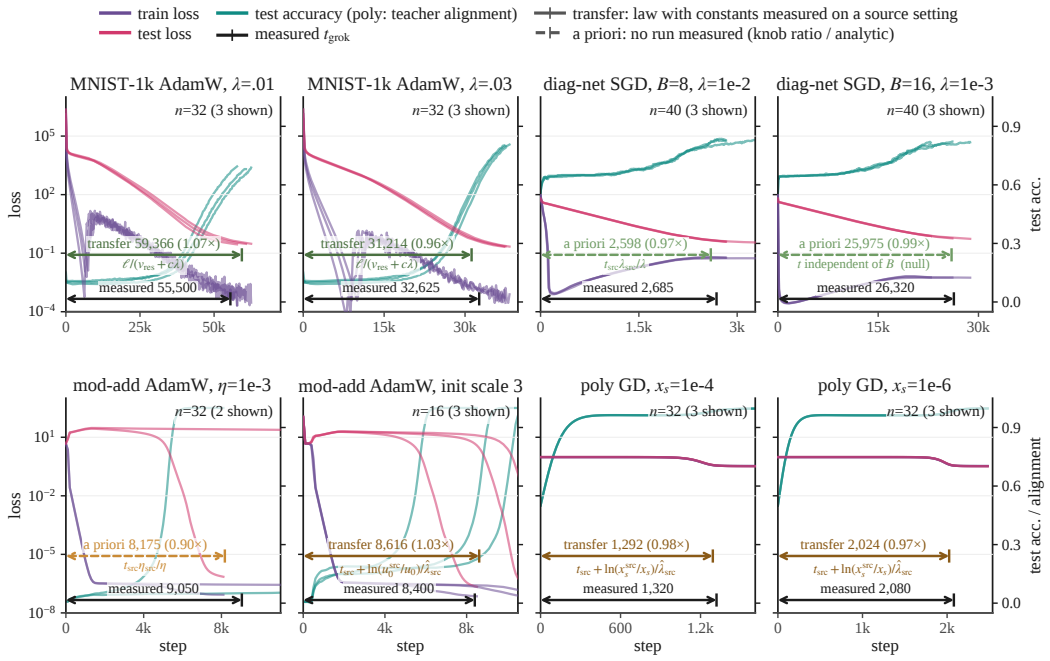


Figure 10: Eight transferred laws. Nothing is measured on the run shown: solid arrows take the law's constants from a source setting, dashed arrows rescale a sibling's measured delay by the ratio the law predicts. The fourth panel is a predicted null: the flat-stretch law contains no batch size.

learning curves, including the four of Fig. 5. Two further transfers on modular addition are censored and are not in the table: at initialization scales 0.3 and 0.1 the regime-A law with the source's rate predicts 9,526 and 9,960 steps, but 11 of 16 and 16 of 16 seeds do not grok within 6×10^4 steps. These are genuine misses, on the one host where Adam sits at its ε floor and the plateau is gradient descent at η/ε (§3.2); a smaller initialization there does not only move the entrance closer to the ridge top, it changes which coordinates are below the floor, and the law has no term for that. On the polynomial host the source rate $\hat{\lambda}$ is taken from the clone probe on the three dense seeds of one setting; its panels are individually below the resolution floor, and the rate is the fit across the whole plateau, which is why it is quoted with the 377 steps per decade it reproduces rather than as a resolved panel value.

B.5 THE THREE REGIMES IN ISOLATION

No reproduced setting is in regime C, and no single reproduced setting lets us dial a knob through all three regimes. To check the laws where they are sharpest we use two constructed systems on which every plateau quantity is known: a two-parameter matrix-factorization saddle, or double well, with post-optimizer injected Gaussian noise of controlled variance (S0a), and a diagonal linear net on sparse regression (S0b). These are first-passage problems, not grokking demonstrations: on S0a the training and test sets coincide, and on S0b the run starts on the interpolation manifold, so neither shows a memorize-then-generalize curve. What they show is the kinetics (Fig. 12).

Regime A. On the S0a saddle, with the rate $\lambda = 0.0198$ per step read by the clone probe, the delay is linear in $\ln(1/u_0)$ with slope 50.4 against $1/\lambda = 50$ (Fig. 12a). Starting instead at the ridge top and sweeping the injected temperature $T = \sigma^2/2\eta$, the delay is linear in $\ln(1/T)$ with slope 23.9 ± 0.5 against $1/(2\lambda) = 25$: the ratio of the two slopes is 2.10 ± 0.05 against the predicted 2, the signature of an entrance inside the noise ball (Appendix D.5). Seed spread does not separate the two arms (CV $= 0.05$–0.09 and 0.08–0.10 per cell); the factor of two does.

Regime B. On S0b the decay transport follows the integrated law $N^* = \ln(x_0/x_{\text{exit}})/(2\eta\lambda_{\text{wd}})$, the factor 2 coming from the product parameterization of the diagonal net; with no fitted constant it lands at 0.965–0.978 of the measured delay on four cells spanning a decade of $1/(\eta\lambda_{\text{wd}})$, and the measured decay rate of the coordinate is 0.93–1.05 times $2\eta\lambda_{\text{wd}}$ on every cell. Rescaling one measured delay by the law predicts a threefold and a 3.3-fold change of λ_{wd} and a doubling of η to within 1.5%. The batch-size null is shown on the literature host that has a batch sweep, the diagonal linear net (⑦): at $B = 8$ and $B = 16$ the 40-seed delay distributions coincide (medians $25{,}975$ and $26{,}320$, KS $p = 0.77$; Fig. 12d). The inverse-Gaussian spread of Eq. (8), CV $= \sqrt{2/\Psi}$, predicts 0.008–0.023 on the S0b cells against measured 0.011–0.025.

Regime C. On the S0a double well with injected noise the dynamics reduce exactly to one coordinate (gradient descent conserves $u^2 - v^2$, the runs start on $u = v$, and the noise is injected along that line), so $\mu = -\eta L'$, $D = \sigma^2/2$ and $\Phi = 2\eta L/\sigma^2$ with no approximation. The well relaxes in $1/|\lambda| = 4.4$ steps and the barrier top has $1/\lambda_b = 4.3$ steps, against delays of 10^3–10^4, and the readout fires 4–5 noise lengths past the top (Appendix D.7). Evaluating Eq. (3) on the measured L with no fitted constant predicts every cell within two standard errors of the seed mean (Fig. 11b): 2,300, 6,300, 10,400 and 18,700 steps at $\sigma = 0.050$, 0.045, 0.043 and 0.041 against measured $2{,}800 \pm 1{,}100$, $6{,}000 \pm 1{,}600$, $8{,}900 \pm 2{,}000$ and $13{,}600 \pm 2{,}600$ ($n = 21$–23), and 6,300 against $5{,}900 \pm 800$ on a second 62-seed cell at $\sigma = 0.045$. The barrier heights are $\Delta\Phi = 4.4$–6.5, and Eq. (9) with its 2π prefactor is within 6–9% of the exact integral there; the π prefactor that would apply to a readout placed exactly at the top is off by 2.1–$2.2\times$. The seeds are exponentially distributed on the held-out cell (CV $= 1.05$, KS $p = 0.76$ against an exponential of the same mean; Fig. 12f), 0.90 on the coldest rung and 1.04 on the 62-seed cell; the exact second moment gives CV $= 1.00$ on every rung. The hottest rung has CV $= 1.7$ and a rejected exponential ($p = 0.03$), which comes from two seeds with delays of $13{,}500$ and $20{,}500$ steps against a mean of 2,300; they escape normally when they do, the same two seeds are the latest on every rung (the rungs share master seeds), and without them the rung has CV $= 0.82$. Under the exact law a 21-seed cell reaches CV ≥ 1.7 with probability 10^{-3}; we report it as an unexplained fluctuation of this seed set. The same shared seeds explain a number we had earlier read as physics: the Arrhenius slope of the four-rung ladder is $0.75\,\Delta L$ ($0.64\,\Delta L$ on the calibration seeds) with an apparent $R^2 = 0.9999$, while the exact integral's slope over the same rungs is $0.98\,\Delta L$ and stays at 0.98–$1.00\,\Delta L$ for σ from 0.025 to 0.07. Because each seed runs the same noise sequence at every σ, the rungs are not independent and the R^2 overstates the precision: resampling seeds jointly across rungs gives a 90% interval of $[0.48, 1.11]\,\Delta L$ for the slope, and dropping the two late seeds moves it to $1.07\,\Delta L$. On this one-dimensional anchor, then, $\Delta\Phi = \Delta L/T$ exactly and the measured delays agree with it; the fitted-line prediction of the held-out cell (9,076 against 8,896 measured) is a fit to correlated data, and the no-parameter prediction is 10,438. Finally, doubling the injected temperature on this well speeds the escape by $10.7\times$ (90% interval $[6.4, 13.5]$, 64 seeds per arm), against $14\times$ from the exact integral, each of the two cells being individually within $1.3\times$ of its prediction; this is the largest acceleration of any intervention we measured.

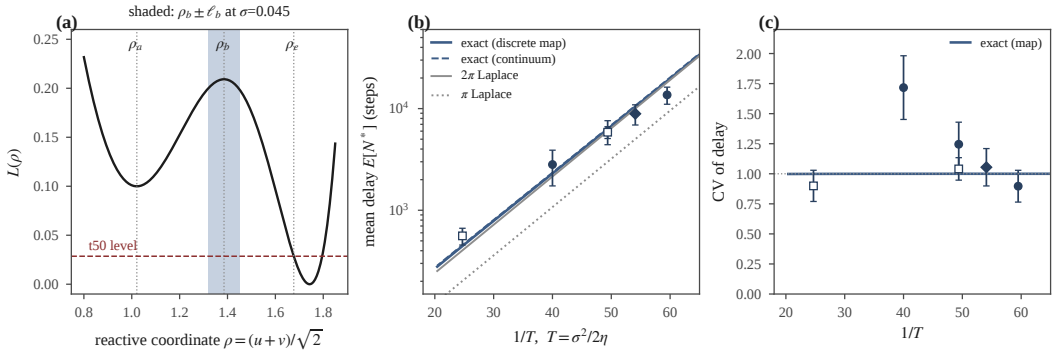


Figure 11: Regime C on the S0a anchor against the exact governing function. (a) The reduced loss $L(\rho)$ with the well, the barrier top, the absorbing point of the readout, and one top noise length shaded. (b) Mean delay against $1/T$: the exact integral (no fitted constant), the 2π and π Laplace forms, and the seed means with standard errors. (c) CV per cell against the exact value of 1.00.

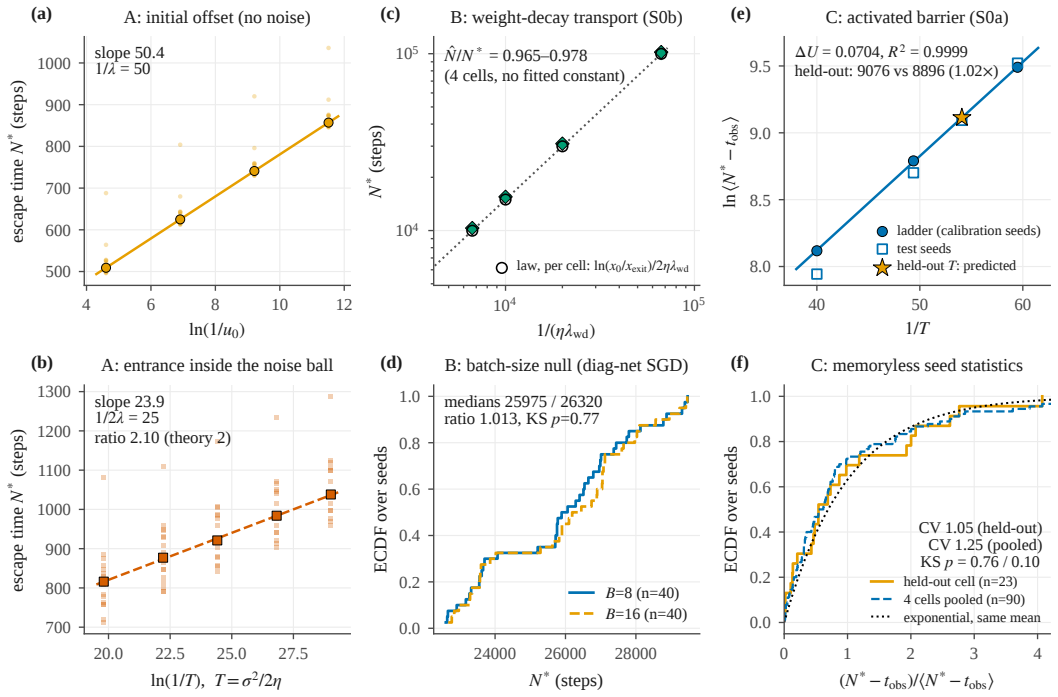


Figure 12: The three regimes in isolation. (a) Regime A on the S0a saddle: delay against $\ln(1/u_0)$, slope $1/\lambda$. (b) The same saddle entered from inside the noise ball: delay against $\ln(1/T)$, slope $1/(2\lambda)$; ratio of slopes 2.10. (c) Regime B on S0b: delay against $1/(\eta\lambda_{\text{wd}})$ with the integrated law and no fitted constant. (d) The batch-size null on the diagonal linear net, 40 seeds per batch size. (e) Regime C on the S0a double well: $\ln N^*$ against $1/T$, the Arrhenius fit on the calibration seeds, the test seeds, and the held-out temperature. (f) Seed distribution on the held-out cell against an exponential of the same mean.

C WHERE THE EXISTING THEORIES OF GROKKING SIT

This appendix walks through the twelve explanations of grokking listed in §4, one at a time, and says where each one sits in our picture. It is written to be read on its own. The only ideas it needs are the three regimes of §2.2, restated here in words.

The picture in words. After a network has memorized its training set, the part of training that still matters for generalization is slow motion along one direction in parameter space. We call the position along that direction the slow coordinate. The delay before grokking is the time this coordinate needs to travel from where the plateau begins to where the test readout responds. Two things drive the motion: a steady push (the drift, from the gradient and from weight decay) and a random jitter (the diffusion, from minibatch sampling). What decides the kind of delay a run has is how the push changes along the way, which we summarize by one number, Ω: the fractional change of the drift between the start of the stage and its exit.

- **Regime A, escaping ridge** ($\Omega \gg 1$). The run starts near a point where the push is almost zero and grows as the run moves away. The first part of the delay is spent creeping, the last part rushing. The delay is set by how close to the ridge top the run starts, so it depends on the initialization scale, logarithmically for an ordinary ridge and as a power law for a degenerate one. Seed-to-seed spread is small.
- **Regime B, flat stretch** ($|\Omega| \lesssim 1$). The push is roughly constant, so the delay is distance over speed. Under weight decay the speed is $\eta\lambda_{\mathrm{wd}}$ on the log of the parameter norm, which gives the $1/(\eta\lambda_{\mathrm{wd}})$ law. Batch size does not enter. Spread is small.
- **Regime C, activated barrier** ($\Omega < -1$). The push reverses before the exit, so the run sits in a dip and only the jitter can carry it out. The delay grows exponentially with the depth of the dip divided by the noise level, and the exit time is memoryless: seeds spread as widely as the mean itself (coefficient of variation near one).

Two more facts are used repeatedly. First, the learning rate rescales the push and the jitter together, so it sets the clock but never the regime. Second, Adam behaves as three different optimizers depending on the size of its second-moment buffer $\hat{v}$ relative to the gradient and to its ε constant (§3.2): noise-dominated (N), where it whitens the noise and the delay does not depend on batch size; signal-dominated (S), where it becomes sign descent and every coordinate moves at the same speed η; and floor-dominated (E), where $\hat{v}$ has fallen below ε and Adam is plain gradient descent at the very large rate η/ε. Full-batch Adam on a memorized plateau sits in (E). The settings are the nine of Appendix A, numbered ①–⑨ there.

C.1 Summary: the twelve accounts on one map

Table 8 places the twelve accounts. Reading down the table, the accounts fall into four groups by what they are actually about. Four describe the *push*: the flat-stretch transport of regime B, usually from weight decay. Four describe the *start*: the ridge of regime A, which produces both hidden progress and sudden transitions. Two describe *noise* as the cause of the escape, which is regime C. Three describe the *optimizer*, and all three turn out to be statements about Adam at its ε floor. One asks whether a delay exists at all. Several accounts belong to two groups, which is itself part of the resolution: an account can be right about the mechanism it isolates and silent about the one that sets the delay in its own experiments.

C.2 Transport accounts

C.2.1 Weight norm (Omnigrok)

Liu et al. (2022b) propose that grokking happens when weight decay has shrunk the parameter norm into a shell where the network generalizes, and that the delay is the time decay needs to get there. Read in our terms this is a flat-stretch law: on the coordinate $\ln\|\theta\|$ decay is a constant push of speed $\eta\lambda_{\mathrm{wd}}$, so the delay is $\ln(\rho_0/\rho_*)/(\eta\lambda_{\mathrm{wd}})$, Eq. (**??**). We confirm the law where the account was proposed and beyond it. On MNIST-1k (④) the decay component of the delay scales as $\lambda_{\mathrm{wd}}^{-0.98}$; on the Power configuration (②) the product $\lambda_{\mathrm{wd}}N^*$ is constant to 1% across a tenfold change in λ_{wd}; on the diagonal linear net (⑦) a threefold change in λ_{wd} rescales the delay to within 8% and 18% with nothing fitted. The account is right about its channel.

What it misses is that the channel is not alone. The account predicts no generalization at $\lambda_{\mathrm{wd}} = 0$, and two reproductions found generalization there anyway. On MNIST-1k every seed groks without decay, at $N^* = 108{,}125$. The reason is that the gradient itself supplies a second push along the same coordinate, and pushes add: $1/N^*$ is linear in λ_{wd} ($R^2 = 0.993$), with an intercept that is the gradient channel. The drift instrument sees the handover directly, as the fraction of the plateau

Table 8: The twelve accounts of grokking and where each sits. "Reads as" is the regime, drift channel or Adam limit that the account's own mechanism corresponds to. "Setting" is the reproduced setting of Appendix A on which we test it.

Account	Its mechanism, in one line	Reads as	Setting	Where checked
The push: transport along a flat stretch				
Weight norm (Liu et al., 2022b)	decay carries the norm to a generalizing shell; $t \propto 1/(\eta\lambda_{\mathrm{wd}})$	B, decay channel	④, ②, ⑦	§C.2.1
Implicit-bias switch (Lyu et al., 2024)	kernel phase, then a sharp switch to a min-norm solution	B (transport), then the exit	⑦	§C.2.2
Sparse vs. dense (Merrill et al., 2023)	a sparse subnetwork grows while a dense one decays	A (growth) with a B decay channel	⑤, ⑥	§C.2.3
Circuit formation and cleanup (Nanda et al., 2023; Chughtai et al., 2023)	circuit completes, then decay cleans up memorization	A on the circuit, B for cleanup; Adam (E)	①, ③	§C.2.4
The start: a ridge				
Lazy-to-rich (Kumar et al., 2024)	delay is the time to leave the lazy (kernel) point; log in laziness	A, linear ridge	⑧, ④	§C.3.1
Hidden progress (Barak et al., 2022)	the sparse solution is amplified invisibly; not a search	A, degenerate ridge	⑤, ⑥	§C.3.2
First-order transition (Rubin et al., 2024)	two states exchange stability; sharp jump	the exit of A; not C	⑥	§C.3.3
Representation phases (Liu et al., 2022a)	comprehension / grokking / memorization / confusion	the non-grokking edges	②, ③, ⑨	§C.3.4
Noise as the cause				
Eyring / Langevin escape (Kozyrev, 2026)	activated hop over a loss barrier at a temperature T	C	none native; S0	§C.4.1
Slingshots (Thilak et al., 2022)	Adam's cyclic instability supplies the kick	noise source, off the slow coordinate	②, ①	§C.4.2
The optimizer				
Circuit efficiency (Varma et al., 2023)	the generalizing circuit is more efficient; decay is necessary	statics; "necessary" is limit (E)	①, ④	§C.5.1
Numerical stability (Prieto et al., 2025)	logit growth freezes the gradient (softmax collapse)	limit (E) without decay	①	§C.5.2

motion pointing along the norm rises from 0.22 to 0.94 as decay is turned up from 0 to 0.03 (§4.2). A single power law fitted through the crossover gives a slope near -0.4, which is why a naive test of the $1/\lambda_{\mathrm{wd}}$ law can wrongly reject it. The additive law also says when decay is the right knob to turn: only above the crossover decay (≈ 0.014 on this host) does doubling λ_{wd} halve the delay.

C.2.2 Implicit-bias switch

Lyu et al. (2024) prove, for homogeneous networks with small weight decay, that training first behaves like a fixed kernel and then switches sharply to a phase that favors small-norm solutions, and that the switch is what produces grokking. This is a two-part description: a transport phase and an exit. We reproduce it on the account's own diagonal linear net (⑦, full-batch SGD). The transport is a flat stretch on the log-norm: Ω stays inside $|\Omega| < 1$ on both estimators (profile -0.4 to -0.15, probe point estimate 0.8 to 0.35), and the delay follows the decay law, scaling as $\lambda_{\mathrm{wd}}^{-0.91}$ across three decay strengths with the a-priori rescaling of a sibling cell landing at 0.92 and 0.82 of the measured delay. Seed spread is tiny ($\mathrm{CV} = 0.01$–0.02), as a deterministic transport requires. The account's theorem therefore describes a regime-B stage, and its "switch" is the exit of that stage: the point where the norm has fallen far enough for the test readout to respond. Because the account is deterministic (gradient flow), it is silent on what minibatch noise does. The framework fills that in: under SGD, batch size enters only the diffusion, and a transport delay does not depend on it. On the same host at $B = 8$ and $B = 16$ the delays agree to 1.3% (Table 2). The account also does not say whether the late phase helps: our dense-target arm on the same net, with the same decay, *loses* test accuracy from 0.99 to 0.68 as the norm falls, because no small-norm solution fits a dense target. Same push, opposite outcome; the regime decides the timing, the target decides the endpoint.

C.2.3 SPARSE VERSUS DENSE SUBNETWORKS

Merrill et al. (2023) describe grokking on sparse parity as a race: a sparse subnetwork that solves the task grows in norm while the dense memorizing subnetwork decays under weight decay, and the generalizing neurons "are inactive during memorization and then explode in norm". The growth is regime A, the decay is a regime-B channel, and the account does not say which of the two sets the delay. On the parity host (⑤) the clone probe answers this directly: wherever it resolves a rate, that rate is positive (a growing direction, with Ω between 5 and 67 on the fastest cell), and on no panel does it resolve a decaying one; on the slower cells the ridge is so gentle that the rate sits below the probe's resolution floor, which is itself what the delay implies ($\lambda \approx 10^{-4}$ per step). The explosion the account describes is the last part of a ridge, and on this host the ridge is degenerate (the target has degree three in the sense of Appendix C.3.2), which is why the explosion is preceded by a long stretch of almost nothing. Batch size on this host acts the way regime A predicts for a start inside the noise ball: larger batches are *slower*, by 2,500 steps per doubling of B ($R^2 = 0.93$), because less noise means a smaller effective initial offset from the ridge top.

C.2.4 CIRCUIT FORMATION AND CLEANUP

Nanda et al. (2023) and Chughtai et al. (2023) split training on full-batch AdamW into memorization, circuit formation and cleanup, with the generalizing circuit complete before the test accuracy moves and weight decay driving the late cleanup. Two things are true of these settings before any run. They are full batch, so there is no sampling noise and the delay must be a deterministic transport, never an activated escape. And their stored Adam buffers place them in the floor limit (E): on the Nanda configuration (①) $\sqrt{\hat{v}_i} < \varepsilon$ on 99.9% of coordinates, and on the group-composition task (③) on 97.7%. The plateau is therefore ordinary gradient descent at the rate η/ε plus decoupled decay. On the circuit coordinate (the concentration of the embedding on a few Fourier modes) the run is an escaping ridge: $\lambda > 0$ on 70 of 83 probe panels, Ω from about 800 down to about 30 before the exit, and the delay scales with the learning rate as $\eta^{-1.18}$ across three learning rates (8-seed cells at the ends), against the -1 every drift-limited law predicts. The staged picture is right; what the stages are made of is one ridge (formation) followed by the decay channel (cleanup), with no barrier anywhere. Seed spread is $\mathrm{CV} = 0.42$, far from the $\mathrm{CV} \approx 1$ a barrier hop would show. Chughtai et al. (2023)'s task is the same story with a longer ridge: none of 32 seeds exits within 6×10^4 steps, a budget-frozen cell (Appendix F).

C.3 RIDGE ACCOUNTS

C.3.1 LAZY-TO-RICH

Kumar et al. (2024) attribute grokking to the time a network needs to leave the lazy regime, where it behaves like a fixed kernel, and derive that this time grows logarithmically as the initialization is made lazier, with no role for noise. This is the regime-A law for a linear ridge, Eq. (7): the delay is $\lambda^{-1} \ln(\ell/u_0)$ with u_0 the initial offset from the ridge top. On the account's polynomial-regression setting (⑧) the law holds over seven decades of initialization scale at 376 steps per decade, against $\ln 10/\lambda = 378$ with λ read from the clone probe on a single source run, and the one-run transfer predicts every other scale to within 3% (Table 2). What the account cannot say is what a different ridge does. If the drift vanishes faster than linearly at the top, as it does for a cubic target, the same transit-time integral gives a power law in the initialization instead of a logarithm (§4.2). Nor does it say what noise does: on a minibatch host the same ridge has a noise ball around its top, and a start inside that ball makes the delay depend on the noise level with half the slope it has in the initialization (Appendix B.5). The lazy-to-rich picture is the noiseless, linear special case of regime A.

C.3.2 HIDDEN PROGRESS

Barak et al. (2022) show that on sparse parity SGD "gradually amplifies the sparse solution" while the loss and error stay flat, and argue that this is amplification of a signal already present in the gradient, not a search. We agree, and add what makes the progress hidden. For a target whose lowest nonzero Hermite degree is $k \geq 3$ (parity with $k = 3$ is the Boolean case) the population gradient near the ridge top vanishes like x^{k-1}, so the ridge is degenerate and the transit time is a

power law in the initial offset, $N^* \propto u_0^{1-p}$ with $p = k - 1$, rather than a logarithm. For $p = 2$ the trajectory is $x(t) = u_0/(1 - t/N^*)$: the offset only doubles over the first half of the delay and diverges at the end. That is exactly "gradual" then "sudden", one curve read at two times. On the cubic teacher–student host (⑥) we see both: alignment moves from 0.031 to 0.044 in the first half of the delay and from 0.24 to 0.97 in the last tenth, and the delay grows $9.9\times$ per decade of initialization, a power law, where a linear ridge (⑧) gives a fixed 376 steps per decade. The search-versus-amplification question the account raises is answered by the sign of λ, and on every panel we can resolve on parity and on the cubic host it is positive: amplification, never search.

C.3.3 First-order transition

Rubin et al. (2024) describe grokking as a first-order transition between two competing states, with hysteresis, and place it in an equilibrium free-energy picture. The sharpness they emphasize is real and we reproduce it on a cubic teacher–student network (⑥): the loss sits flat for most of the run while the alignment creeps, then train and test loss drop together within a few hundred steps. But the mechanism producing the sharpness here is the end of a ridge, not a barrier hop. The clone probe reads $\lambda > 0$ throughout (an unstable direction, $\Omega \gg 1$), the seed spread is $\mathrm{CV} = 0.18$ rather than ≈ 1, and the delay scales as a power of the initialization rather than exponentially in a barrier height. A first-order transition in the thermodynamic sense needs noise to cross a barrier, i.e. regime C, and a full-batch run has no noise to do it with. What survives of the account is the statement about *which* state wins once the run leaves the plateau, which is a question about endpoints, not about the delay; the delay itself is kinetic.

C.3.4 Representation phases

Liu et al. (2022a) chart four outcomes as the amount of training data changes: comprehension (immediate generalization), grokking, memorization (no generalization within budget) and confusion. Three of these are not regimes but ways a delay can fail to appear, which Appendix F lists as N3 (no plateau), N1 (budget exceeded) and N2 (no target). One setting shows two of them side by side. On the Power configuration at the largest decay we ran ($\lambda_{\mathrm{wd}} = 3$, ②), 8 of 32 seeds generalize almost as soon as they fit the training set (train and test times in ratio 1.3, against 51, 16 and 3.2 at the three smaller decays) and the other 24 never generalize within budget. The group-composition task (③) is entirely in N1, and the Adam+L_2 arm of ⑨ is N2: sign descent removes the network before any generalizing state exists to reach. Inside the region where a delay does exist, the data fraction is a landscape knob, and the three regimes and their laws apply unchanged.

C.4 Noise accounts

C.4.1 Activated escape (Eyring / Langevin)

Kozyrev (2026) model the delay as an activated hop over a loss barrier at a temperature set by the optimizer, with an Eyring-type rate. This is regime C, and our Eq. (9) is the same rate written on the slow coordinate, with two corrections. The barrier that matters is the integral of drift over diffusion, $\Delta\Phi$, not the loss barrier ΔL; the two coincide only when the noise is uniform along the path. And the temperature is not free: it is η/B for SGD and enters Adam differently in each of its three limits. We can realize regime C on a constructed system with injected noise (Appendix B.5), and there the account's signature is exactly as predicted: the delay is exponential in $1/T$ over the whole range tested, seeds spread with $\mathrm{CV} \approx 1$ and an exponential shape, and doubling the temperature speeds the escape up $10.7\times$. The point that matters for the literature is that *no reproduced setting is in this regime*. The account cites full-batch modular arithmetic, where there is no sampling noise at all; and on every minibatch setting we ran the seed spread is far below one (largest $\mathrm{CV} = 0.42$, on the full-batch modular-addition host) and the delay is a power law, not an exponential, in the knobs. Activated escape is a real regime that the standard grokking settings do not occupy.

C.4.2 Slingshots

Thilak et al. (2022) document a cyclic instability of Adam, in which the training loss spikes and the norm jumps, and report that grokking occurs at the onset of these slingshots and is absent without them. §4.2 tests this on the Power configuration (②; modular subtraction in our reproduction). The

bursts are real and frequent: at $\lambda_{\mathrm{wd}} \leq 0.3$ every seed passes through a median of 6 to 7 before it groks (under every burst definition we tried, Appendix A.7). But they neither time the transition (the grokking step falls inside a burst in only 0, 3 and 2 of about 31 seeds at the three decays, no more than the fraction of the plateau the bursts occupy) nor are they necessary: at $\lambda_{\mathrm{wd}} = 1$ most seeds grok without a single burst after memorization (28 of 31 under the definition of Appendix A.7). The delay follows the decay law to 1%. In the language of §2.4, a burst is a fast excursion off the slow coordinate: it moves the run a long way in directions that relax back, and leaves the slow coordinate, and hence the delay, to the decay transport. On the full-batch Nanda host (①) the origin of the bursts is also clear: the plateau is gradient descent at η/ε (limit E), and $\kappa = (\eta/\varepsilon)\lambda_{\max}(H)$ sits at 15 to 25, above the descent stability threshold of 2, so the run oscillates along its stiffest directions. That is an instability of the optimizer at the floor, not a source of the delay. Whether slingshots are the only route to grokking at $\lambda_{\mathrm{wd}} = 0$, the configuration the account was established on, is not tested here; at zero decay on the floor host the gradient underflows and nothing escapes (§C.5.2).

C.5 Optimizer accounts

C.5.1 Circuit efficiency and the necessity of weight decay

Varma et al. (2023) explain which solution wins, memorizing or generalizing, by which is more efficient at a given norm, and conclude that weight decay is necessary for the generalizing circuit to be preferred. The efficiency argument is about endpoints, and our framework does not contradict it. The necessity claim is about the delay, and it holds exactly on the setting it was made on and not beyond. On the Nanda host (①) the plateau is in Adam's floor limit (E). With decay on, the run is gradient descent at η/ε plus decay and groks in every seed. With decay off, the training loss keeps falling until it underflows, the gradient vanishes with it, the only remaining push $-(\eta/\varepsilon)\partial L$ goes to zero, and 0 of 8 seeds escape in 4×10^4 steps. Decay is necessary there because it is the only push left. On MNIST-1k (④), a minibatch host in limit (N), the gradient channel never dies and every seed groks at $\lambda_{\mathrm{wd}} = 0$. "Weight decay is necessary" is a true statement about limit (E), not about grokking.

C.5.2 Numerical stability

Prieto et al. (2025) identify softmax collapse: as the logits grow, the cross-entropy gradient falls below floating-point resolution and training freezes on the plateau, which a numerically stable loss (StableMax) prevents. This is the same limit (E) fact seen from the loss side. On the Nanda host without decay, the training loss falls below 10^{-6} and then underflows, $\sqrt{\hat{v}}$ falls to $10^{-4}\varepsilon$ on every coordinate, the clone probe resolves no rate on 26 of 28 panels (no push in any direction), and no seed escapes. The account is right that the freeze is numerical; our reading adds that it is Adam's ε floor and the absence of decay together that produce it, which is why the same host with decay groks normally through the same tiny gradients: decay bypasses the preconditioner and does not care how small ∂L is. On a host with an MSE loss and no softmax the mechanism has no channel to act through, and precision perturbations there are null. Prieto et al. (2025)'s fix, Varma et al. (2023)'s necessity claim and Thilak et al. (2022)'s bursts are three views of one configuration: Adam on a memorized full-batch plateau at its floor.

C.5.3 Coupled versus decoupled decay

A distinction none of the twelve accounts draws turns out to matter as much as any of them. Weight decay can be applied after Adam's preconditioner (AdamW, decoupled) or added to the loss (Adam+L_2, coupled). On the quadratic-MLP modular-addition host (⑨) at the same nominal $\eta = \lambda_{\mathrm{wd}} = 10^{-3}$, AdamW groks in 16 of 16 seeds at a median of 2,100 steps while Adam+L_2 groks in 0 of 16 within 10^5. The reason is the limit map of §3.2. On a memorized plateau the loss gradient collapses but the L_2 gradient $\lambda_{\mathrm{wd}}\theta_i$ does not, so it becomes Adam's own normalizer ($\sqrt{\hat{v}_i} \approx 1.16\,\lambda_{\mathrm{wd}}|\theta_i|$ at the median coordinate), every coordinate enters limit (S), and the update is sign descent at speed η regardless of λ_{wd}. The norm then falls linearly, at 0.095 per step, from about 100 to 0.007 in 15 of 16 seeds, and the network is removed. Decoupled decay over the same run shrinks the norm by a factor of only $e^{-\eta\lambda_{\mathrm{wd}}t} = 0.998$ over its 2,100 steps; the AdamW run's norm drop is the gradient's doing. A nominal λ_{wd} therefore does not transfer between the two, and comparisons across papers that mix them are not comparisons at matched decay.

Table 9: New statements the kinetic framework makes about grokking, with the core evidence for each. Settings are numbered as in Appendix A.

	Statement	Core evidence	Resolves / extends
1	Grokking delays follow closed-form laws with no fitted constant, and the laws transfer across learning rate, decay, batch size and initialization.	23 transfers on five hosts and three optimizers, median error 2.6% (Table 2); the diag-net batch-size null 0.99.	all quantitative accounts
2	Weight decay is one of two pushes along the same coordinate; pushes add, so rates add.	MNIST-1k: $1/N^* = a + b\lambda_{\mathrm{wd}}$, $R^2 = 0.993$; every seed groks at $\lambda_{\mathrm{wd}} = 0$; drift alignment with the norm $0.22 \to 0.94$.	Omnigrok vs. its reproductions
3	Slingshots are real but do not set the delay; decay transport does.	Power p=97: $\lambda_{\mathrm{wd}} N^*$ constant to 1%; N^* falls inside a burst in at most 3 of 32 seeds; at λ_{wd}=1, $28/31$ seeds grok without a burst after memorization.	Thilak vs. weight-norm accounts
4	"Gradual" and "sudden" are the two ends of one degenerate ridge; which one a study sees depends on when it looks.	He_3 host: alignment $0.031 \to 0.044$ over the first half, $0.24 \to 0.97$ over the last tenth; $\Omega \gg 1$ throughout.	Barak vs. Merrill
5	The delay is logarithmic in the initialization scale for a linear ridge and a power law for a degenerate one; the exponent is the degeneracy, not the task.	Poly host: 376 steps/decade vs. $\ln 10/\lambda = 378$; He_3 host: $9.9\times$ per decade, $7/40$ finish at 10^{-4}.	Kumar; Rubin's sharpness
6	The most-reproduced setting is Adam at its ε floor, i.e. gradient descent at η/ε; three accounts are one statement about it.	Nanda host: $\sqrt{\hat{v}} < \varepsilon$ on 99.9% of coordinates; $\kappa =$ 15–25 vs. threshold 2; wd$=0$: loss underflows, $0/8$ escape.	Prieto, Varma, Thilak
7	Coupled and decoupled decay are different optimizers at the same nominal λ_{wd}; coupled decay becomes sign descent and removes the network.	⑨: AdamW $16/16$ at 2,100 vs. Adam+L_2 $0/16$; $\sqrt{\hat{v}_i} = 1.16\lambda_{\mathrm{wd}}\lvert\theta_i\rvert$; norm linear to 0.007.	all AdamW-vs-Adam comparisons
8	Activated escape (regime C) exists but no standard grokking setting occupies it; full-batch settings cannot.	S0 with injected noise: the exact governing function predicts all six cells within 2 SE with no fitted constant, $\mathrm{CV} = 0.9$–1.05 on five of them; every reproduced setting has $\mathrm{CV} \le 0.42$.	Kozyrev, Rubin
9	The learning rate sets the clock, not the regime; in A and B the delay is $\propto 1/\eta$.	Nanda host $\eta^{-1.18}$; mod-add a-priori transfers across η at $0.90\times$ (Table 2).	—
10	Batch size does not enter a transport delay under SGD, nor any delay under AdamW in limit (N); on a ridge started inside the noise ball, larger batches are slower.	diag-net B=8 vs. 16: 1.3%; parity: $+2{,}500$ steps per doubling.	Barak (batch), Lyu (noise)
11	Seed-to-seed spread is a regime fingerprint: small in A and B, near one in C.	CV 0.004–0.42 across all nine reproduced settings; ≈ 1 only on the constructed C anchor.	Kozyrev
12	The non-grokking outcomes are three edges, not regimes, and one cell can show two of them.	Power config, wd$=3$: $8/32$ immediate, $24/32$ frozen; ③ frozen; Adam+L_2 no target.	Liu et al.

C.6 What the framework adds

Table 9 collects the statements above that go beyond the accounts they touch: conflicts that are resolved, predictions that no existing account makes, and the evidence for each.

D Derivations for the Kinetics

This appendix supplies every derivation deferred from §2: the drift–diffusion description and its error terms (§D.1), the governing function and its generalization to the full delay distribution (§D.2), the constant-D reduction (§D.3), the parabolic profile and the transit-time approximation (§D.4), the three regime laws with their distributions (§D.5–§D.7), a summary of hazard rates (§D.8), and the reduction from the full parameter space to one slow coordinate (§D.9). Notation follows §2: θ is the slow coordinate, Δ_t the per-step displacement, μ and D its drift and diffusion (Eq. 2), $\Phi = -\int \mu/D$, and $(v, \lambda, \ell, \Psi, \Theta, \Omega)$ the plateau quantities of §2.2.

D.1 From the Update Rule to Drift and Diffusion

Cumulants of one step. The displacement is $\Delta_t = -\eta\,\hat{g}_t$ with $\hat{g}_t = B^{-1}\sum_{i=1}^{B} \partial_\theta \ell_i(\theta_t)$ an average of B independent per-example gradients. Write κ_k for the k-th cumulant of Δ_t given $\theta_t = \theta$ and $\kappa_k^{(1)}$ for that of a single per-example gradient. Cumulants add over independent summands and are homogeneous of degree k under scaling, so

$$\kappa_k = (-\eta)^k \, \frac{\kappa_k^{(1)}}{B^{k-1}}, \qquad k \ge 1. \tag{13}$$

The first two are Eq. (2): $\kappa_1 = -\eta\,\partial_\theta L = \mu$ and $\kappa_2 = \eta^2 c/B = 2D$. Dividing by κ_2,

$$\frac{\kappa_k}{\kappa_2} = \Big(\frac{-\eta}{B}\Big)^{k-2}\frac{\kappa_k^{(1)}}{\kappa_2^{(1)}}, \qquad k \geq 3, \tag{14}$$

which is the statement in §2.1 that the third and higher cumulants are $O(\eta^3)$ and further suppressed by B^{-2}. The ratio $\kappa_k^{(1)}/\kappa_2^{(1)}$ is a property of the data and the loss, not of the optimizer setting; it is what the kurtosis monitor of Appendix G measures.

Error of the second-order truncation. Conditioning on the first step, $f(\theta) = 1 + \mathbb{E}[f(\theta + \Delta)]$, and expanding,

$$\mu f' + \tfrac{1}{2}M_2 f'' + \sum_{k\geq 3}\frac{M_k}{k!}f^{(k)} = -1, \qquad M_k := \mathbb{E}[\Delta^k \mid \theta], \tag{15}$$

with $M_2 = 2D + \mu^2$. Let $\ell_{\min}$ be the shortest length on which f, μ or D varies, and $\epsilon := (|\mu| + \sqrt{2D})/\ell_{\min}$ the per-step displacement in units of that length. Each derivative costs $\ell_{\min}^{-1}$, so term $k \geq 3$ is smaller than the retained diffusion term by $O\big((\kappa_k/\kappa_2)\,\ell_{\min}^{-(k-2)}\big) = O\big((\eta/B\ell_{\min})^{k-2}\big)$ by Eq. (14), and the μ^2 part of M_2 is smaller than the drift term $\mu f'$ by $O(|\mu|/\ell_{\min}) = O(\epsilon)$. Relative to the diffusion term it is $\mu^2/2D$; on a plateau in regime B, $\mu \approx v$ and $D = v\ell/\Psi$, so $\mu^2/2D = \Psi/(2N^*)$, which is 10^{-2} or less on every setting of Table 5 (the footnote of §2.1). Truncating Eq. (15) after the second derivative therefore leaves

$$D(\theta)f''(\theta) + \mu(\theta)f'(\theta) = -1 \tag{16}$$

with relative error $O(\epsilon) + O(\eta/B\ell_{\min})$. Truncation at second order is also the only consistent choice: by Pawula's theorem a truncation of the Kramers–Moyal expansion at any finite order above two admits negative densities, so the expansion is either kept in full or cut at two.

Stationary density (remark). The same moments define the forward (continuity) equation $\partial_t p + \partial_\theta J = 0$ with $J = \mu p - \partial_\theta(Dp)$, whose zero-current stationary solution on an interval with a reflecting end is $p \propto e^{-\Phi}/D$. This is how Φ acquires its meaning as the log of a stationary density; the derivation of Eq. (3) in §2.1 does not need it, and the escape problem, having an absorbing end, is never stationary. In one dimension there is no circulating current, which is why Φ has an equilibrium-like form although training is not at equilibrium.

D.2 The Governing Function and Its Generalization

Solution of Eq. (16). The boundary conditions are $f(\theta_e) = 0$ (absorption at the exit) and $f'(\theta_a) = 0$ (reflection at the lower end of the coordinate). Since $\mu/D = -\Phi'$, the operator has integrating factor $e^{-\Phi}$:

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[e^{-\Phi(x)}f'(x)\Big] = e^{-\Phi(x)}\Big[f'' - \Phi' f'\Big] = e^{-\Phi(x)}\Big[f'' + \tfrac{\mu}{D}f'\Big] = -\frac{e^{-\Phi(x)}}{D(x)}. \tag{17}$$

Integrating from θ_a to x with $f'(\theta_a) = 0$,

$$f'(x) = -\,e^{\Phi(x)}\int_{\theta_a}^{x}\frac{e^{-\Phi(y)}}{D(y)}\,\mathrm{d}y, \tag{18}$$

and integrating Eq. (18) from θ_s to θ_e with $f(\theta_e) = 0$ gives Eq. (3). The additive constant in Φ cancels between the two exponentials, so the lower limit in the definition $\Phi = -\int_0^\theta \mu/D$ is immaterial.

The generalized governing function. The same first-step conditioning applied to $\mathbb{E}_\theta[e^{-zN^*}] =: \hat{f}(\theta; z)$ gives, after the same truncation,

$$D\,\hat{f}'' + \mu\,\hat{f}' - z\,\hat{f} = 0, \qquad \hat{f}(\theta_e; z) = 1, \quad \hat{f}'(\theta_a; z) = 0. \tag{19}$$

Eq. (19) is the generalized governing function referred to in §2.3: it determines the full distribution of N^*, and Eq. (3) is its first-order coefficient in z. Expanding $\hat{f} = \sum_m (-z)^m M_m/m!$ with $M_m(\theta) := \mathbb{E}_\theta[(N^*)^m]$ gives the moment hierarchy

$$D\,M_m'' + \mu\,M_m' = -m\,M_{m-1}, \qquad M_m(\theta_e) = 0, \quad M_m'(\theta_a) = 0, \quad M_0 \equiv 1, \tag{20}$$

which the integrating factor solves recursively,

$$M_m(\theta_s) = m \int_{\theta_s}^{\theta_e} \mathrm{d}x\, e^{\Phi(x)} \int_{\theta_a}^{x} \mathrm{d}y\, \frac{e^{-\Phi(y)}}{D(y)}\, M_{m-1}(y). \tag{21}$$

$m = 1$ is Eq. (3); $m = 2$ gives $\mathrm{Var}[N^*] = M_2 - M_1^2$ and hence CV. Every CV in Table 1 is obtained from Eq. (21) or, where a closed form exists, from Eq. (19) directly. The survival function $S(n) = \Pr[N^* > n]$ and the hazard $h(n) = -\mathrm{d}\ln S/\mathrm{d}n$ follow from inverting the transform; §D.8 collects them.

D.3 The Constant-D Reduction

Write $D(y) = D_0\, e^{\ln(D(y)/D_0)}$ with $D_0 := D(\theta^*)$. Then Eq. (3) is exactly

$$\mathbb{E}[N^*] = \frac{1}{D_0} \int_{\theta_s}^{\theta_e} \mathrm{d}x\, e^{\Phi(x)} \int_{\theta_a}^{x} \mathrm{d}y\, e^{-\tilde{\Phi}(y)}, \qquad \tilde{\Phi}(y) := \Phi(y) + \ln \frac{D(y)}{D_0}, \tag{22}$$

which is the form quoted in §2.2. The variation of D across the stage therefore acts only as a correction to the potential seen by the inner integral, and it is a small one: if $\ln D$ varies by $\Delta \ln D$ over the stage while Φ drops by $|\Delta\Phi|$, replacing $\tilde{\Phi}$ by Φ changes the tilt of the inner exponent by a relative amount $\Delta \ln D/|\Delta\Phi|$, i.e. $\Psi \to \Psi \pm \Delta \ln D$ on the parabola, and the delay by the same relative order. Table 5 lists both quantities on every thermal setting: the ratio is below 1.1×10^{-3} on all hosts but sparse parity, where it reaches 0.064. Two remarks. First, in regimes A and B the leading term of the delay is the transit time Eq. (5), which contains no D at all (§D.4); the variation of D then enters only the $O(1/\Psi)$ corrections. Second, in regime C the inner integral is dominated by the well and the outer by the barrier top, so the prefactor of Eq. (9) uses D at the well while the exponent $\Delta\Phi$ is exact with the true $D(\theta)$; nothing in the exponent is approximated by the reduction. The largest variation of D we measure is on the SGD hosts with an MSE loss, where the per-example gradient variance shrinks with the training loss as the plateau deepens; even there $\Delta \ln D \le 17$ against $|\Delta\Phi| \ge 10^4$.

D.4 The Parabolic Profile and the Transit-Time Approximation

Tilt, curvature, shape. With $\mu(\theta) = v + \lambda(\theta - \theta^*)$ and $s = (\theta - \theta^*)/\ell$,

$$\Phi(\theta) - \Phi(\theta^*) = -\frac{1}{D} \int_0^{\theta-\theta^*} (v + \lambda u)\, \mathrm{d}u = -\frac{v\ell}{D}\, s - \frac{\lambda \ell^2}{D}\, \frac{s^2}{2} = -\Psi s - \tfrac{1}{2}\Theta s^2, \tag{23}$$

so $\Phi'(s) = -\Psi - \Theta s$ and $\mu(\theta_e)/\mu(\theta^*) = (v + \lambda\ell)/v = 1 + \Omega$ with $\Omega = \Theta/\Psi = \lambda\ell/v$, Eq. (4). For $\Theta < 0$ the profile has its minimum where $\Phi' = 0$, at $s = -\Psi/\Theta = 1/|\Omega|$, which lies inside the stage exactly when $|\Omega| > 1$. This is the boundary between the flat stretch and the activated barrier. On the other side, the drift $v + \lambda\ell s$ has doubled by $s = 1/\Omega$ when $\Omega > 0$, and the profile is a downward-opening parabola whose slope at the exit is $1 + \Omega$ times its slope at the entrance; $\Omega \gg 1$ is the escaping ridge.

Transit time with its first correction. When Φ decreases monotonically over the stage ($\Omega > -1$), evaluate the inner integral of Eq. (3) by expanding about its upper limit. With $u := y - x \le 0$, $\Phi(y) = \Phi(x) - \frac{\mu(x)}{D}u + \frac{1}{2}\Phi''(x)u^2 + \dots$ and $\Phi'' = -\mu'/D$, so

$$\int_{\theta_a}^{x} \frac{e^{-\Phi(y)}}{D}\, \mathrm{d}y = \frac{e^{-\Phi(x)}}{D} \int_{-\infty}^{0} e^{\frac{\mu(x)}{D}u + \frac{\mu'(x)}{2D}u^2}\, \mathrm{d}u + \dots = \frac{e^{-\Phi(x)}}{\mu(x)} \Big[1 + \frac{\mu'(x) D}{\mu(x)^2} + O\Big(\frac{D^2\mu''}{\mu^3}, \frac{D^2\mu'^2}{\mu^4}\Big)\Big], \tag{24}$$

where extending the lower limit to $-\infty$ costs $O(e^{-\Psi})$ when the reflecting end is at least a stage away. Substituting in Eq. (3),

$$\mathbb{E}[N^*] = \int_{\theta_s}^{\theta_e} \frac{\mathrm{d}\theta}{\mu(\theta)} \;+\; \int_{\theta_s}^{\theta_e} \frac{\mu'(\theta)\, D}{\mu(\theta)^3}\, \mathrm{d}\theta \;+\; \dots \tag{25}$$

The first term is Eq. (5). The second is the leading stochastic correction; at the entrance $\mu' D/\mu^2 = \lambda D/v^2 = \Omega/\Psi$, which is the "relative order $1/\Psi$" of §2.2. It is positive on a ridge (noise near the top delays the run on average, because the drift there is weakest and the run spends time on both sides of its deterministic path) and it vanishes on a flat stretch, where the only correction is the reflecting-wall term of Eq. (8). The next corrections are $O((\Omega/\Psi)^2)$ and $O(1/\Theta)$, as verified on the exact ridge integral below.

D.5 Regime A: Escaping Ridge

Reduction to the ridge integral. Let $\theta_r = \theta^* - u_0$, $u_0 = v/\lambda$, be the point where the linear drift vanishes, and $x = \theta - \theta_r$, so $\mu = \lambda x$ and $\Phi(x) = -\lambda x^2/2D = -x^2/2\ell_T^2$ with $\ell_T = \sqrt{D/\lambda}$. The entrance is at $x = u_0$, the exit at $x = \ell + u_0$, and $u_0^2/\ell_T^2 = v^2/\lambda D = \Psi/\Omega$. Because the linear drift is odd about $x = 0$, first passage of $|x|$ to $\ell + u_0$ on the whole line has the same law as first passage on the half-line with a reflecting end at $x = 0$, which is the boundary condition of Eq. (3). With $w = x/\ell_T$ and $t = y/\ell_T$, $\mathrm{d}x\,\mathrm{d}y = \ell_T^2\,\mathrm{d}w\,\mathrm{d}t$ and $\ell_T^2/D = 1/\lambda$, which gives Eq. (6) with $F(w) = e^{-w^2/2}\int_0^w e^{t^2/2}\mathrm{d}t = \sqrt{2}\,\mathrm{Daw}(w/\sqrt{2})$, a scaled Dawson function. It behaves as $F(w) = w - w^3/3 + \dots$ for $w \ll 1$ and

$$F(w) = \frac{1}{w} + \frac{1}{w^3} + \frac{3}{w^5} + O(w^{-7}), \qquad w \gg 1. \tag{26}$$

Entrance outside the noise ball, $u_0 \gg \ell_T$ (Eq. 7). Both limits $w_0 = u_0/\ell_T$ and $w_1 = (1+\Omega)w_0$ lie in the asymptotic range, and integrating Eq. (26) term by term,

$$\lambda\,\mathbb{E}[N_A^*] = \ln(1+\Omega) + \frac{\Omega}{2\Psi}\Big[1 - \frac{1}{(1+\Omega)^2}\Big] + \frac{3}{4}\Big(\frac{\Omega}{\Psi}\Big)^2\Big[1 - \frac{1}{(1+\Omega)^4}\Big] + \dots \tag{27}$$

The term dropped in Eq. (7) is $\Omega/[2\Psi(1+\Omega)^2] \approx 1/(2\Theta)$, and the next one is $O((\Omega/\Psi)^2)$, consistent with Eq. (25): the second term of Eq. (27) is exactly $\int \mu' D/\mu^3$ for $\mu = \lambda x$. The leading term is the transit time $\lambda^{-1}\ln((\ell+u_0)/u_0)$, logarithmic in the entrance offset: doubling u_0 saves $\ln 2/\lambda$ steps, regardless of ℓ.

Entrance inside the noise ball, $u_0 \ll \ell_T$. This limit is not used in the main text, because every reproduced ridge setting enters from outside the ball (Table 7), but it is realized on the constructed anchor of Appendix B.5 and it carries a distinct signature. With $w_0 \to 0$ and $w_1 = \sqrt{\Theta}$, $\int_0^W F(w)\,\mathrm{d}w = \ln W + \frac{1}{2}(\gamma + \ln 2) + O(W^{-2})$, where $\gamma = 0.5772$ is Euler's constant, so

$$\mathbb{E}[N_A^*] = \frac{1}{2\lambda}\Big[\ln\Theta + \ln 2 + \gamma\Big] = \frac{1}{2\lambda}\Big[\ln\Theta + 1.270\Big]. \tag{28}$$

Comparing with Eq. (7), the delay per e-fold of the controlling quantity halves: $\mathrm{d}\mathbb{E}[N]/\mathrm{d}\ln(1/u_0) = 1/\lambda$ outside the ball against $\mathrm{d}\mathbb{E}[N]/\mathrm{d}\ln(1/D) = 1/2\lambda$ inside it. This factor of two is the crossover measured on the S0 anchor (slopes 50.4 and 23.9 steps per e-fold, ratio 2.10, Appendix B.5). Its origin is that the noise sets the effective offset through its standard deviation $\ell_T \propto \sqrt{D}$, so a decade in D is only half a decade in offset.

Distribution and CV. The linear SDE $\mathrm{d}x = \lambda x\,\mathrm{d}t + \sqrt{2D}\,\mathrm{d}W$ solves to

$$x(t) = e^{\lambda t}\Big[u_0 + \int_0^t e^{-\lambda s}\sqrt{2D}\,\mathrm{d}W_s\Big]. \tag{29}$$

The stochastic integral converges for $\lambda t \gg 1$ to a Gaussian Z with variance $2D\int_0^\infty e^{-2\lambda s}\mathrm{d}s = D/\lambda = \ell_T^2$, so the late-time trajectory is $x(t) \approx (u_0 + Z)e^{\lambda t}$: the noise acts once, at the top, by jittering the effective offset. Solving $|u_0 + Z|e^{\lambda N} = \ell + u_0$,

$$N_A^* = \frac{1}{\lambda}\Big[\ln(\ell + u_0) - \ln|u_0 + Z|\Big], \qquad Z \sim \mathcal{N}(0, \ell_T^2). \tag{30}$$

Outside the ball, $\ln|u_0 + Z| = \ln u_0 + Z/u_0 + O(Z^2/u_0^2)$, and if the initialization also spreads $\ln u_0$ across seeds with variance $\sigma^2[\ln u_0]$ independently of Z,

$$\mathrm{Var}[N_A^*] = \frac{\sigma^2[\ln u_0] + \ell_T^2/u_0^2}{\lambda^2} = \frac{\sigma^2[\ln u_0] + \Omega/\Psi}{\lambda^2}, \qquad \mathrm{CV}_A = \frac{\sqrt{\sigma^2[\ln u_0] + \Omega/\Psi}}{\ln(1+\Omega)}, \tag{31}$$

the entry of Table 1. The two contributions are separable by experiment: the first is unchanged by η or D and the second scales as D. When $\ln u_0$ is Gaussian across seeds, N_A^* is Gaussian (the initialization enters through a logarithm, so a log-normal offset gives a normal delay), with rising hazard. Inside the ball, $N_A^* = \lambda^{-1}[\frac{1}{2}\ln\Theta - \ln|Z'|]$ with Z' standard normal; $\mathbb{E}[-\ln|Z'|] = \frac{1}{2}(\gamma + \ln 2)$ reproduces Eq. (28), $\mathrm{Var}[\ln|Z'|] = \pi^2/8$ gives

$$\mathrm{CV}_A^{\mathrm{inside}} = \frac{\pi/\sqrt{8}}{\frac{1}{2}\ln\Theta + 0.635}, \tag{32}$$

and the shape is log-half-normal. The survival function is $S(n) = \Pr[|Z'| < \sqrt{\Theta}e^{-\lambda n}] = \mathrm{erf}(\sqrt{\Theta/2}\,e^{-\lambda n})$, whose hazard rises from zero and saturates at λ itself, so the saturation value of the hazard reads the ridge rate from timing statistics alone. The spread in both cases shrinks with the number of e-folds $\ln(1+\Omega)$ the ridge amplifies: long ridges are reproducible across seeds whatever their noise.

Degenerate ridge. If the drift vanishes to higher order at the top, $\mu = \kappa_p x^p$ with $p \geq 2$ and $\kappa_p > 0$, the transit time from u_0 is

$$N^* = \int_{u_0}^{\ell+u_0} \frac{\mathrm{d}x}{\kappa_p x^p} = \frac{u_0^{1-p} - (\ell+u_0)^{1-p}}{(p-1)\kappa_p} \simeq \frac{u_0^{1-p}}{(p-1)\kappa_p} \qquad (\ell \gg u_0), \tag{33}$$

a power law in the entrance offset in place of the logarithm, as stated in §2.3. Written with the entrance rate $\lambda = \mu'(u_0) = pv/u_0$ this is $N^* = p/[(p-1)\lambda]$: on a degenerate ridge the delay is a fixed multiple of the local relaxation time at the entrance, with no logarithm because almost all of the delay is spent near the entrance. For $p = 2$ the deterministic trajectory is closed, $\dot{x} = \kappa_2 x^2$ giving

$$x(t) = \frac{u_0}{1 - t/N^*}, \qquad N^* = \frac{1}{\kappa_2 u_0}, \tag{34}$$

which is the "gradual, then sudden" curve of §4: the coordinate has moved only to $2u_0$ by $t = N^*/2$ and to $10u_0$ by $0.9N^*$, and then diverges. The seed spread follows from Eq. (33): $\sigma[\ln N^*] = (p-1)\,\sigma[\ln u_0]$, so a degenerate ridge amplifies initialization spread by $p-1$ rather than compressing it by $1/\ln(1+\Omega)$. The noise ball of a degenerate ridge has radius $(D/\kappa_p)^{1/(p+1)}$, where $\kappa_p x^p \sim D/x$.

D.6 Regime B: Flat Stretch

Exact mean with the reflecting entrance (Eq. 8). Set $\theta^* = 0$, $\mu = v$, $\Phi = -vx/D$, reflecting at the entrance and absorbing at ℓ. The inner integral is $\int_0^x e^{vy/D}\mathrm{d}y/D = (e^{vx/D} - 1)/v$, so

$$\mathbb{E}[N_B^*] = \int_0^\ell \frac{1 - e^{-vx/D}}{v}\,\mathrm{d}x = \frac{\ell}{v} - \frac{D}{v^2}\big(1 - e^{-\Psi}\big) = \frac{\ell}{v}\Big[1 - \frac{1 - e^{-\Psi}}{\Psi}\Big]. \tag{35}$$

The wall shortens the transit by D/v^2, one diffusive excursion length D/v at speed v: a run started at a reflecting wall can only be pushed forward by the noise. As $\Psi \to 0$ the bracket tends to $\Psi/2$ and the mean to the purely diffusive $\ell^2/2D$, a limit no measured grokking setting approaches ($\Psi \geq 16$, Table 5). With the curvature kept to first order, Eq. (5) gives $\int_0^\ell \mathrm{d}\theta/(v + \lambda\theta) = (\ell/v)\ln(1+\Omega)/\Omega$, the correction factor $1 - \Omega/2 + \Omega^2/3 - \ldots$ quoted in §2.3; at $|\Omega| = 1$ it is $\ln 2 = 0.69$ or ∞ on the two sides, which is why the flat-stretch law is used only for $|\Omega| \lesssim 1$ and the boundary with regime C is sharp.

Distribution: inverse Gaussian. For $\Psi \gg 1$ the wall is revisited with probability $e^{-\Psi}$ and can be removed. On the free line Eq. (19) has constant coefficients, $D\hat{f}'' + v\hat{f}' - z\hat{f} = 0$ with $\hat{f}(\ell) = 1$ and $\hat{f}$ bounded as $x \to -\infty$; the bounded root is $r_+ = (-v + \sqrt{v^2 + 4Dz})/2D > 0$ and $\hat{f}(x) = e^{r_+(x-\ell)}$, so from $x = 0$

$$\mathbb{E}\big[e^{-zN_B^*}\big] = \exp\Big[\frac{\ell}{2D}\Big(v - \sqrt{v^2 + 4Dz}\Big)\Big], \tag{36}$$

the transform of the inverse Gaussian (Wald) law with mean $m = \ell/v$ and shape parameter $\ell^2/2D$. Expanding $\sqrt{v^2 + 4Dz} = v(1 + 2Dz/v^2 - 2D^2z^2/v^4 + \dots)$,

$$\ln \mathbb{E}[e^{-zN_B^*}] = -\frac{\ell}{v}z + \frac{D\ell}{v^3}z^2 + O(z^3) \;\Rightarrow\; \mathbb{E}[N_B^*] = \frac{\ell}{v}, \quad \mathrm{Var}[N_B^*] = \frac{2D\ell}{v^3}, \quad \mathrm{CV}_B = \sqrt{\frac{2D}{v\ell}} = \sqrt{\frac{2}{\Psi}}. \tag{37}$$

The skewness is $3\mathrm{CV}_B$, and the density $f(n) = \sqrt{\ell^2/4\pi D n^3}\, \exp[-(n-m)^2 v^2/4Dn]$ has a hazard that vanishes at $n \to 0$, peaks near m, and decays to the constant $v^2/4D = \Psi/(4\mathbb{E}[N_B^*])$: a run that has overshot its scheduled arrival still escapes at a rate $\Psi/4$ times the memoryless reference, which is $\gg 1$ throughout the regime. Restoring the wall changes the mean by the $-1/\Psi$ term above and the variance by the same relative order, as one checks by inserting $M_1(y) = (\ell - y)/v - (D/v^2)(e^{-vy/D} - e^{-v\ell/D})$ into Eq. (21). Because the mean depends only on the drift and the spread only on D/v, a flat stretch is falsifiable from seeds: (v, ℓ, D) measured on the plateau predict CV with no fitted parameter (Table 6 compares the two on every host).

D.7 Regime C: Activated Barrier

Laplace evaluation of Eq. (3). With $\Omega < -1$ the drift is forward at θ^*, vanishes at the well $\theta_a = \theta^* + \ell/|\Omega|$ with $\mu'(\theta_a) = \lambda < 0$, and is backward beyond it until the barrier top θ_b, where it vanishes again with $\mu'(\theta_b) = \lambda_b > 0$. Hence Φ has a minimum at θ_a with $\Phi''(\theta_a) = |\lambda|/D$ and a maximum at θ_b with $\Phi''(\theta_b) = -\lambda_b/D$. (The reflecting end of Eq. (3) may sit anywhere below the well; it does not enter the result, which is why the main text does not distinguish it from θ_a.) For $\Delta\Phi = \Phi(\theta_b) - \Phi(\theta_a) \gg 1$ the inner integral, for any x past the well, is dominated by the Gaussian peak of $e^{-\Phi}$ at the well,

$$\int_{\theta_a}^{x} \frac{e^{-\Phi(y)}}{D}\,\mathrm{d}y \simeq \frac{e^{-\Phi(\theta_a)}}{D}\sqrt{\frac{2\pi D}{|\lambda|}} \qquad (x - \theta_a \gg \sqrt{D/|\lambda|}), \tag{38}$$

and the outer integral by the Gaussian peak of e^{Φ} at the top, $\int^{\theta_e} e^{\Phi(x)}\mathrm{d}x \simeq e^{\Phi(\theta_b)}\sqrt{2\pi D/\lambda_b}\cdot\chi$, where $\chi = 1$ if the absorbing point lies beyond the top by more than its noise length $\sqrt{D/\lambda_b}$, and $\chi = \frac{1}{2}$ if absorption is exactly at the top. The product is

$$\mathbb{E}[N_C^*] \simeq \chi\,\frac{2\pi}{\sqrt{|\lambda|\lambda_b}}\,e^{\Delta\Phi}, \tag{39}$$

which is Eq. (9) for $\chi = 1$. D cancels from the prefactor and survives only in $\Delta\Phi$. The factor χ has a simple meaning: a run that reaches the top exactly is equally likely to fall back or to go on, so the time to *commit* to the far basin is twice the time to first *touch* the top. The readouts of Table 3 fire only once the run has left the barrier region (test accuracy or alignment thresholds well past the point of no return), so $\chi = 1$ is the convention that matches how N^* is measured, and we use 2π throughout; if one instead defined the exit as the top itself, every regime-C prefactor in the paper would be halved and nothing else would change. The constructed anchor of Appendix B.5 settles this by measurement: its readout fires 4–5 top noise lengths past the top, the exact integral is 1.06–1.12 times the 2π form at $\Delta\Phi = 2.7$–6.5 and 2.1–2.2 times the π form, and the measured delays agree with the exact integral within two standard errors on every cell (Fig. 11).

Barrier height on the parabola. On the parabolic profile $\Phi(s) = \Phi^* - \Psi s + \frac{1}{2}\Psi|\Omega|s^2$, evaluating between the well $s = 1/|\Omega|$ and the exit $s = 1$,

$$\Delta\Phi = \Big(-\Psi + \tfrac{1}{2}\Psi|\Omega|\Big) - \Big(-\frac{\Psi}{|\Omega|} + \frac{\Psi}{2|\Omega|}\Big) = \frac{\Psi}{2|\Omega|}\big(|\Omega|^2 - 2|\Omega| + 1\big) = \Psi\,\frac{(|\Omega| - 1)^2}{2|\Omega|}, \tag{40}$$

the last member of Eq. (9). It vanishes quadratically at $|\Omega| = 1$, so the barrier grows continuously out of the flat stretch: there is no discontinuity in the delay at the A/B–C boundary, only a change in which limit of Eq. (3) applies. The parabola treats the curvature as constant across the stage; the true profile turns over at the top, so the parabolic $\Delta\Phi$ is an estimate of the barrier from plateau quantities, and the exact $\Delta\Phi = -\int_{\theta_a}^{\theta_b} \mu/D$ should be used when the drift is measured across the whole stage.

Memorylessness, CV, and finite-barrier corrections. Relaxation into the well takes $O(1/|\lambda|)$ steps, shorter than Eq. (39) by the factor $e^{-\Delta\Phi}$; after that the run has forgotten θ_s and escape is a Poisson event of rate $1/\mathbb{E}[N_C^*]$. Formally, Eq. (19) is then dominated by its smallest eigenvalue $z_0 = 1/\mathbb{E}[N_C^*]$, so $\mathbb{E}[e^{-zN}] \to z_0/(z_0+z)$, the transform of the exponential law: $\mathrm{Var} = \mathbb{E}^2$, $\mathrm{CV}_C = 1$, constant hazard, and the conditional distribution of the remaining wait is the same at every step, so no statistic of a single run's history can forecast its escape. The corrections are of two kinds. The relaxation time itself subtracts a short dead time, giving $\mathrm{CV}_C^2 = 1 - O(\tau_{\mathrm{rel}}/\mathbb{E}[N])$. At the barrier heights realized in practice Laplace's method also carries corrections $O(1/\Delta\Phi)$ from the non-Gaussian parts of the well and the top; on the anchor these are $+3$ to $+9\%$ from each, $+6$ to $+12\%$ in all at $\Delta\Phi = 2.7$–6.5. They are nearly independent of D, so they do not tilt the Arrhenius line: the slope $\mathrm{d}\ln\mathbb{E}[N]/\mathrm{d}(1/T)$ of the exact integral is 0.98–1.00 of the true barrier over the whole measured range. A finite seed ladder can nevertheless return a shallower slope with a deceptively high R^2 when the rungs share their noise sequences, which is what happened on the anchor ($0.75\,\Delta L$ with a 90% resampling interval of $[0.48, 1.11]\,\Delta L$; Appendix B.5); a barrier read from a ladder should therefore be quoted with a seed-resampled interval, not with the R^2 of the line.

D.8 Hazard Rates and Seed Statistics: Summary

Let $S(n) = \Pr[N^* > n]$ and $h(n) = -\mathrm{d}\ln S/\mathrm{d}n$, the escape probability per step given survival. Normalizing its late-time value by the mean, $\bar{h} := h(\infty)\,\mathbb{E}[N^*]$, removes units and makes the regimes comparable; $\bar{h} = 1$ is the memoryless floor.

Regime	Shape of N^*	CV	$\bar{h}$	Forecastable from one run's history
A, $u_0 \gg \ell_T$	Gaussian (normal in $\ln u_0$)	$\sqrt{\sigma^2[\ln u_0] + \Omega/\Psi}\,/\ln(1+\Omega)$	∞ (rises linearly)	yes: arrives on schedule
A, $u_0 \ll \ell_T$	log-half-normal	$(\pi/\sqrt{8})/(\frac{1}{2}\ln\Theta + 0.635)$	$\lambda\,\mathbb{E}[N^*] = \frac{1}{2}\ln\Theta + 0.635$	weakly: hazard saturates at λ
A, degenerate p	power of u_0	$(p-1)\,\sigma[\ln u_0]$	rises	yes
B	inverse Gaussian	$\sqrt{2/\Psi}$	$\Psi/4$	yes: peaked near ℓ/v
C	exponential	1	1	no: constant hazard

The ladder is monotone and $\bar{h}$ is estimable from a seed population without measuring (v, λ, D), which makes it an independent check on a regime read from the probe. Two cautions. A composite of several stages (§D.9.4) has CV bounded by that of its noisiest stage, so $\mathrm{CV} \approx 1$ in a seed population implies a regime-C stage somewhere but $\mathrm{CV} \ll 1$ does not exclude a short one. And a CV near 1 is also produced by a diffusive stretch in a compact geometry ($\Psi \to 0$, first-mode dominated), so shape alone does not separate B at small Ψ from C; the D-scaling of the mean does (power law against exponential), and every measured setting has $\Psi \geq 16$ in any case.

D.9 Reduction to the Slow Coordinate

D.9.1 Projection and the two-term effective diffusion

After quotienting the exact parameter symmetries (Appendix A.4), write the state as (θ, w) with θ the slow coordinate and $w \in \mathbb{R}^{P-1}$ the remaining directions. The joint update has drift (μ_θ, μ_w) and diffusion block $\left(\begin{smallmatrix} D_{\theta\theta} & D_{\theta w} \\ D_{w\theta} & D_{ww} \end{smallmatrix}\right)$, all functions of (θ, w). Let τ_s and τ_f be the relaxation times of the slow direction and of the fastest surviving mode, and $g := \tau_s/\tau_f \gg 1$. On times long compared with τ_f and short compared with τ_s the fast block reaches its conditional stationary density $\rho_\theta(w)$ at frozen θ. Averaging over ρ_θ and collecting the transport that the residual fast fluctuations feed into θ (the Zwanzig projection, to leading order in $1/g$),

$$\mu_{\mathrm{eff}}(\theta) = \big\langle \mu_\theta(\theta, w) \big\rangle_{\rho_\theta}, \tag{41}$$

$$D_{\mathrm{eff}}(\theta) = \underbrace{\big\langle D_{\theta\theta}(\theta, w) \big\rangle_{\rho_\theta}}_{\text{projected minibatch noise } D_0} + \underbrace{\int_0^\infty \big\langle \delta\mu_\theta(\theta, w_0)\, \delta\mu_\theta(\theta, w_s) \big\rangle_{\rho_\theta} \mathrm{d}s}_{\text{memory term } D_{\mathrm{mem}}}, \tag{42}$$

where $\delta\mu_\theta := \mu_\theta - \mu_{\mathrm{eff}}$ and w_s evolves under the fast dynamics at frozen θ. Eq. (42) is the two-term expression referred to in §2.4. The first term is the per-example gradient variance projected on the slow direction, the c of Eq. (2) and of the optimizer maps in §3. The second is the extra transport the slow coordinate inherits because the fast modes fluctuate around their conditional equilibrium rather than sitting at it.

D.9.2 THE $O(1/g)$ BOUND

Linearize about the plateau and keep one representative fast mode (the bound is controlled by the fastest). With drift Jacobian

$$J = \begin{pmatrix} \lambda_0 & \kappa \\ \kappa' & -\lambda_f \end{pmatrix}, \qquad \lambda_f = 1/\tau_f, \tag{43}$$

eliminating w by Schur complement gives the effective slow rate $\lambda = \lambda_0 + \kappa\kappa'/\lambda_f$ with $|\lambda| = 1/\tau_s$. Timescale separation, $|\lambda| \ll \lambda_f$, constrains the coupling: for the transmitted term not to exceed the slow rate, $|\kappa\kappa'| \lesssim |\lambda|\lambda_f$. This is the same condition that licenses the reduction, not a new assumption. At frozen θ the fast coordinate is Ornstein–Uhlenbeck with stationary variance D_{ww}/λ_f and correlation time $1/\lambda_f$; the fluctuating slow drift is $\delta\mu_\theta = \kappa\,\delta w$, so $\langle \delta\mu_\theta(0)\delta\mu_\theta(s)\rangle = \kappa^2(D_{ww}/\lambda_f)e^{-\lambda_f s}$ and $D_{\mathrm{mem}} = \kappa^2 D_{ww}/\lambda_f^2$. Dividing by $D_0 = D_{\theta\theta}$ and using the coupling bound,

$$\frac{D_{\mathrm{mem}}}{D_0} = \frac{\kappa^2}{\lambda_f^2}\frac{D_{ww}}{D_{\theta\theta}} \lesssim \frac{|\lambda|}{\lambda_f}\cdot\frac{\kappa}{\kappa'}\cdot\frac{D_{ww}}{D_{\theta\theta}} = \frac{R}{g}, \qquad R := \frac{\kappa}{\kappa'}\frac{D_{ww}}{D_{\theta\theta}}. \tag{44}$$

The memory term is $O(1/g)$ provided $R = O(1)$. R collects two asymmetries that can inflate it without violating timescale separation: a non-normal Jacobian ($\kappa \gg \kappa'$) and noise much louder in the fast directions than along the slow one ($D_{ww} \gg D_{\theta\theta}$). Where either holds the slow coordinate carries a diffusion that no optimizer map of §3 predicts, and the delay laws must be used with the measured D_{eff} rather than the projected D_0.

D.9.3 EMPIRICAL VALIDATION

Eq. (44) is testable without forming the high-dimensional generator, because it predicts a discrepancy between two independent measurements of the same quantity: D_0, the per-example gradient variance projected on the slow direction (from the noise block of the gauge panel, Appendix A.5), and D_{eff}, the quadratic variation of the slow coordinate about its fitted drift over a trailing window (from the drift block). Their relative difference should stay inside an envelope of slope $O(R)$ in $1/g$, with $g = \lambda_f/|\lambda|$ read from the clone probe as the ratio of the fastest resolved decaying rate to the slow direction's rate.

We ran this test on every stored gauge panel (2,484 panels; 736 usable after removing athermal full-batch cells, duplicated relaunches, and panels without a stored covariance), and we report what it can and cannot show (Fig. 13). On the constructed anchors both quantities are measured on the same coordinate and g is available from the dense Hessian. In regime C on S0a (122 panels, $1/g = 0$ because the well relaxes in 4 steps against escapes of 10^3–10^4) the median discrepancy is $r = (D_{\mathrm{eff}} - D_0)/D_0 = 0.14$, of which 0.136 is the slow mode's own relaxation within one step (an Ornstein–Uhlenbeck variable sampled once per step has one-step variance $2D_0\,[1 + |\lambda|/(2 - |\lambda|)]$, and $|\lambda| = 0.23$ per step there), leaving 0.002: $D_{\mathrm{eff}} = D_0$ at the origin, as Eq. (44) requires. On the diffusive S0b cells $r = -0.02$ (median $|r| = 0.04$). On the S0a ridge (184 panels, $g \approx 1$, i.e. no separation, so the bound is vacuous) the growth-corrected median is $r = 0.03$ with a tail to 3 on panels near the exit. On the decay-driven S0b cells at the smallest injected noise the estimator returns $r \approx 10$ on every panel; this is not a memory term (the reduction there is exact, the net is one-dimensional along the coordinate) but the deterministic variation of the drift inside the trailing window, which a single-lag quadratic variation with only the window mean removed counts as diffusion, and we flag it as a limitation of the stored estimator. On the minibatch literature hosts the test cannot be run from the stored panels: the probe resolves no decaying direction on the sparse-parity and Power hosts and is burst-contaminated on MNIST-1k, so g is unavailable; the drift block's D_{eff} is measured along the trailing-window valley direction, whose overlap with the probe's reactive direction is 10^{-6}–10^{-4}; and on the AdamW hosts the projected D_0 needs the full preconditioned covariance, of which only the diagonal is stored. What the panels do show there is that $|r|$ is $O(1)$ on the SGD hosts (parity: median 0.15, 90th percentile 0.5) and grows with a Hessian proxy for $1/g$ (Spearman 0.66, $n = 28$), consistent with but not a test of Eq. (44). The reduction is therefore validated where it can be, at $1/g = 0$, and its $O(1/g)$ envelope on the real hosts remains an assumption of the same kind as the timescale separation itself. Completing the test needs no training, only re-gauging from the stored checkpoints: probe directions orthogonal to the reactive one with longer bursts, the diffusion measured along the probe's reactive direction, and the coarse-grained rather than single-lag quadratic variation.

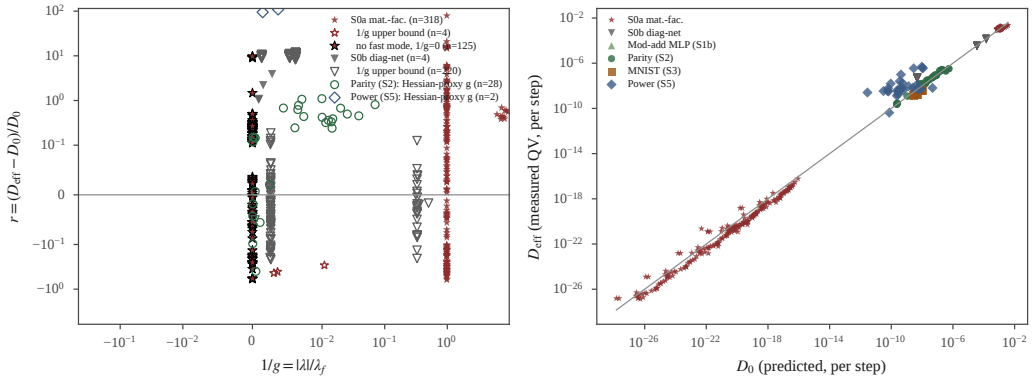


Figure 13: Slow-coordinate validation from the stored panels. Left: the relative discrepancy $r = (D_{\text{eff}} - D_0)/D_0$ against $1/g$ where g is resolved (S0 anchors); the cluster at $1/g = 0$ is regime C on S0a after the single-step relaxation correction. Right: D_{eff} against D_0 on all usable panels with the identity line; the decay-driven S0b cells at the smallest noise (upper left) are the estimator artefact discussed in the text.

D.9.4 Multi-stage composition and parallel channels

Stages in series. When J carries several comparably slow eigenvalues the escape resolves into successive stages, each with its own bottleneck and regime. Delays add, and for stages made independent by the fast-mode equilibration between them, variances add:

$$\mathbb{E}[N^*] = \sum_j \mathbb{E}[N_j^*], \qquad \mathrm{Var}[N^*] = \sum_j \mathbb{E}[N_j^*]^2 \mathrm{CV}_j^2, \qquad \mathrm{CV}_{\text{total}} = \frac{\sqrt{\sum_j \mathbb{E}[N_j^*]^2 \mathrm{CV}_j^2}}{\sum_j \mathbb{E}[N_j^*]}, \tag{45}$$

as stated in §2.4. $\mathrm{CV}_{\text{total}} \le \max_j \mathrm{CV}_j$, so a composite never looks noisier than its noisiest stage, and the sum is dominated by the stage with the largest mean, so the observed scaling is that of the rate-limiting stage. Published setups that disagree about the mechanism of one task are often reporting different rate-limiting stages of the same composite (Appendix C); and a law read forward at a fixed fraction of N^* (Appendix B.3) reads early exactly when the rate-limiting stage has not yet begun.

Drifts in parallel. When two mechanisms drive the *same* coordinate, their drifts add, $\mu = \mu_1 + \mu_2$, and on a flat stretch the rates add rather than the times:

$$\frac{1}{\mathbb{E}[N_B^*]} = \frac{v_1 + v_2}{\ell} = \frac{1}{N_1} + \frac{1}{N_2}. \tag{46}$$

On the norm coordinate $\theta = \ln \rho$ decoupled weight decay contributes exactly $-\eta\lambda_{\text{wd}}$ (§3.1), independently of the gradient's contribution v_∇, so

$$\frac{1}{N^*} = \frac{v_\nabla}{\ell} + \frac{\eta}{\ell}\,\lambda_{\text{wd}} =: a + b\,\lambda_{\text{wd}}, \tag{47}$$

the form fitted to MNIST-1k in §4.2 ($a = 9.7 \times 10^{-6}$, $b = 7.1 \times 10^{-4}$, $R^2 = 0.993$). The intercept is the decay-free delay, $1/a = 1.0 \times 10^5$ steps against the measured $N^* = 108{,}125$ at $\lambda_{\text{wd}} = 0$, and the slope fixes the stage width on the norm coordinate, $\ell = \eta/b = 1.4$ at $\eta = 10^{-3}$, i.e. the norm shrinks by $e^{1.4} \approx 4$ between the plateau and the exit. The crossover $\lambda_{\text{wd}}^{\times} = a/b = 0.014$ is where the two channels contribute equally; below it decay is a perturbation on a gradient-driven stretch, above it the stretch is decay-driven and Eq. (**??**) applies.

E Derivations for the Optimizer Laws

This appendix supplies the derivations deferred from §3: the elimination of the optimizer's internal buffers (§E.1), the SGD family including heavy-ball momentum (§E.2), the full derivation of

AdamW's (μ, D) with ε retained and its three limits (§E.3), the precise sense in which AdamW admits a single noise scale only in those limits (§E.4), the two decay channels (§E.5), instability bursts (§E.6), and the row-by-row derivation of the optimizer columns of Table 1 (§E.7). Throughout, i indexes coordinates, c_i is the per-example gradient variance of coordinate i, c its projection on the slow coordinate, and (μ, D) are the effective coefficients of Appendix D.9.

E.1 Elimination of Optimizer State

Momentum and adaptive optimizers carry buffers, so the state is (θ_t, m_t, v_t) rather than θ_t, whereas the kinetics of §2 require a Markov diffusion in θ alone. The buffers relax with time constants $\tau_1 = (1-\beta_1)^{-1}$ and $\tau_2 = (1-\beta_2)^{-1}$, between 10 and 10^3 steps for the β's of Table 3, while N^* is 10^3 to 10^5 steps. With $g_{\rm opt} := N^*/\max(\tau_1, \tau_2) \gg 1$ the buffers are the fast variables and θ the slow one, and the reduction of Appendix D.9.1 applies verbatim: averaging over the buffers' conditional stationary distribution at frozen θ gives effective coefficients for θ, with a memory correction $O(1/g_{\rm opt})$ by Eq. (44). The bound requires the buffer distribution at frozen θ to be stationary. For the second moment v_t this holds in two situations: when its input $\hat{g}_t^2$ has a non-vanishing mean (any minibatch run, limits N and S), or when the buffer has fallen below the floor, $\sqrt{\hat{v}} \ll \varepsilon$, so that the preconditioner is pinned at $1/\varepsilon$ and the buffer's continued decay no longer affects the update (limit E). The one case that escapes it is a buffer above the floor whose input keeps falling, as on a full-batch run before it reaches the floor; the step then inflates until either the floor or the stability edge is reached (§E.6).

E.2 The SGD Family

Plain SGD with decoupled decay (Eq. 10). The step is $\Delta_t = -\eta(\hat{g}_t + \lambda_{\rm wd}\theta_t)$ with $\mathbb{E}[\hat{g}_t] = \partial_\theta L$ and $\mathrm{Var}[\hat{g}_t] = c/B$. The decay term is deterministic given θ_t, so it shifts the mean and leaves the variance untouched: $\mu = -\eta(\partial_\theta L + \lambda_{\rm wd}\theta)$, $D = \eta^2 c/2B$. The ratio is

$$\frac{\mu}{D} = -\frac{2(\partial_\theta L + \lambda_{\rm wd}\theta)/c}{\eta/B} =: -\frac{U'}{T_{\rm SGD}}, \qquad T_{\rm SGD} = \frac{\eta}{B}, \quad U' = \frac{2(\partial_\theta L + \lambda_{\rm wd}\theta)}{c}, \tag{48}$$

so SGD always admits a single temperature: η and B rescale the noise and $\lambda_{\rm wd}$ deforms the landscape, two knobs acting on two different objects. With c constant along the stage, $U = 2(L + \frac{1}{2}\lambda_{\rm wd}\theta^2)/c$ and $\Phi = U/T$, which is what makes the regime-C exponent of Table 1 a function of B/η alone.

Heavy-ball momentum. With $m_t = \beta m_{t-1} + \hat{g}_t$ and $\Delta_t = -\eta m_t$, unrolling gives $m_t = \sum_{k\geq 0}\beta^k \hat{g}_{t-k}$, whose weights sum to $(1-\beta)^{-1}$ rather than to one. Hence $\mathbb{E}[m_t] \to \partial_\theta L/(1-\beta)$ and $\mu_{\rm mom} = \mu_{\rm SGD}/(1-\beta)$ in the gradient term. For the diffusion the relevant quantity is not the per-step variance, $\mathrm{Var}[m_t] = c/[B(1-\beta^2)]$, but the variance of the accumulated displacement, because D is the long-run spreading rate and successive m_t are correlated over $\tau_\beta = (1-\beta)^{-1}$ steps. Summing over $n \gg \tau_\beta$ steps and exchanging the order of summation,

$$\sum_{t=1}^{n} m_t = \sum_{t=1}^{n}\sum_{k\geq 0}\beta^k \hat{g}_{t-k} \approx \frac{1}{1-\beta}\sum_{t=1}^{n}\hat{g}_t, \qquad \text{so} \qquad D_{\rm mom} = \frac{\eta^2 c}{2B(1-\beta)^2} = \frac{D_{\rm SGD}}{(1-\beta)^2}. \tag{49}$$

Both μ and D take their SGD form with $\eta \mapsto \eta/(1-\beta)$ in the gradient terms: momentum SGD is SGD at learning rate $\eta/(1-\beta)$, with $T_{\rm mom} = \eta/[B(1-\beta)]$, and the whole classification transfers. (Using the per-step variance instead would give $(1-\beta^2)^{-1}$ and break this correspondence.) Decoupled decay is applied outside the buffer and is not amplified, so on the norm coordinate the decay-driven drift $v = \eta\lambda_{\rm wd}$ of Eq. (**??**) is unchanged by momentum, while a gradient-driven stage is faster by $(1-\beta)^{-1}$. Nesterov momentum has the same zero-frequency gain and the same result. This is the only place momentum enters the kinetics; it affects the discrete-time stability threshold separately (§E.6).

E.3 AdamW: Derivation of (μ, D) and the Three Limits

The update is, coordinatewise,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\hat{g}_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2)\hat{g}_t^2, \quad \theta_{t+1} = \theta_t - \eta\,\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \eta\lambda_{\rm wd}\theta_t, \tag{50}$$

with $\hat{m}_t = m_t/(1-\beta_1^t)$ and $\hat{v}_t = v_t/(1-\beta_2^t)$.

β_1 drops out. Unlike the heavy-ball buffer, Adam's first moment is a *normalized* average: its weights $(1-\beta_1)\beta_1^k$ sum to $1-\beta_1^t$, which the bias correction divides out. So $\mathbb{E}[\hat{m}_t] \to \partial_i L$ with no amplification, and in the accumulation argument above each $\hat{g}_t$ is collected with total weight $\sum_k (1-\beta_1)\beta_1^k = 1$, so the accumulated variance is also unchanged. β_1 smooths the trajectory on times below τ_1 and is invisible to the kinetics on times $N^* \gg \tau_1$. The contrast with heavy-ball momentum is entirely one of normalization.

Equilibrium of the second moment. At frozen θ, v_t is an exponential average of $\hat{g}_t^2$ with time constant τ_2, and its conditional mean is

$$\mathbb{E}[\hat{v}_i] = \mathbb{E}[\hat{g}_i^2] = (\partial_i L)^2 + \frac{c_i}{B} = \frac{c_i}{B}\big(1 + s_i^2\big), \qquad s_i = \frac{\sqrt{B}\,|\partial_i L|}{\sqrt{c_i}}, \tag{51}$$

which is Eq. (11). Its relative fluctuation is $O(\tau_2^{-1/2})$ and is absorbed in the $O(1/g_{\rm opt})$ error. Both terms are needed: keeping only the first is the deterministic idealization $P = 1/|\partial L|$, which is wrong on every minibatch plateau, where $\partial L \to 0$ but c/B does not.

Drift and diffusion with ε retained. Treating $P_i = (\sqrt{\hat{v}_i} + \varepsilon)^{-1}$ as constant over one step, the update is a preconditioned SGD step plus a decay term, $\Delta_i = -\eta P_i \hat{g}_i - \eta\lambda_{\rm wd}\theta_i$, so

$$\mu_i = -\eta P_i\, \partial_i L - \eta\lambda_{\rm wd}\theta_i, \qquad D_i = \tfrac{1}{2}\eta^2 P_i^2\, \frac{c_i}{B}. \tag{52}$$

Write the floor in the same units as the signal, $e_i := \varepsilon\sqrt{B/c_i}$, the ratio of ε to the noise scale $\sqrt{c_i/B}$. Then $P_i = \sqrt{B/c_i}\,/\big(\sqrt{1+s_i^2} + e_i\big)$ and

$$\boxed{\mu_i^{\nabla} = -\,\eta\,\frac{s_i}{\sqrt{1+s_i^2}+e_i}\,\operatorname{sgn}(\partial_i L), \qquad D_i = \frac{\eta^2}{2\big(\sqrt{1+s_i^2}+e_i\big)^2}} \tag{53}$$

where μ^{∇} is the gradient part of the drift. Everything in §3.2 is a limit of Eq. (53) in the two measurable ratios s_i and e_i ($\sqrt{\hat{v}_i}/\varepsilon = \sqrt{1+s_i^2}/e_i$):

Limit	Condition	μ_i^{∇}	D_i
(N) noise-dominated	$s_i \ll 1,\ e_i \ll 1$	$-\eta\sqrt{B/c_i}\,\partial_i L$	$\eta^2/2$
(S) signal-dominated	$s_i \gg 1,\ e_i \ll s_i$	$-\eta\operatorname{sgn}(\partial_i L)$	$\eta^2 c_i/[2B(\partial_i L)^2]$
(E) floor-dominated	$e_i \gg \sqrt{1+s_i^2}$	$-(\eta/\varepsilon)\,\partial_i L$	$\eta^2 c_i/(2B\varepsilon^2)$
general, $e_i \ll 1$	Eq. (12)	$-\eta\, s_i \operatorname{sgn}(\partial_i L)/\sqrt{1+s_i^2}$	$\eta^2/[2(1+s_i^2)]$

The (E) row follows from $\eta s_i/e_i = \eta|\partial_i L|/\varepsilon$ and $\eta^2/2e_i^2 = \eta^2 c_i/2B\varepsilon^2$: at the floor Adam is SGD at learning rate η/ε in its gradient term, in both drift and diffusion, while the decoupled decay is not amplified. Two consequences of the (E) row are used in §4. At full batch $c_i = 0$, so $D = 0$ and the plateau is gradient descent at η/ε plus decay; it is athermal and regime C is excluded by derivation. At minibatch, the floor amplifies the SGD diffusion by $1/\varepsilon^2$ together with the drift, so that μ/D is that of SGD at temperature $\eta/(\varepsilon B)$: an ε-floor plateau with sampling noise is a very hot SGD plateau.

Corrections at intermediate s. The minibatch hosts of Table 4 sit at $s = 0.4$–0.9 rather than at $s \ll 1$. From the general row, the drift is smaller than its (N) value by $1/\sqrt{1+s^2}$ and the diffusion by $1/(1+s^2)$, so a transit-time delay is longer than the (N) law by $\sqrt{1+s^2} = 1.09$–1.34 and the tilt $\Psi = v\ell/D$ is larger by the same factor. These are the corrections referred to in §3.2; they are applied wherever a law is evaluated on those hosts (Appendix B.3). At $s \geq 1$ the drift saturates at η and the diffusion falls as $1/s^2$: AdamW is loudest exactly where the signal is weakest, which is why the plateau is the noisy part of an AdamW run and the descent after escape is quiet.

Limit E with and without decay. At the floor the gradient drift is $-(\eta/\varepsilon)\partial_i L$. On a memorized plateau of a cross-entropy host the training loss keeps falling and with it every $\partial_i L$; without decay nothing else moves the parameters, so the drift vanishes as the loss underflows and the run freezes (Nanda host at $\lambda_{\rm wd} = 0$: $0/8$ seeds escape in 4×10^4 steps, and on the densely probed seeds the Hessian itself collapses with the loss, $\lambda_{\max}(H)$ falling from 4×10^{-4} at step 1,100 to 4×10^{-6} by step 4,500, against a plateau value of 2–4×10^{-4} with decay). With decay the term $-\eta\lambda_{\rm wd}\theta_i$ persists, shrinks the norm at the SGD rate $\eta\lambda_{\rm wd}$, keeps the loss from underflowing, and the run groks as gradient descent with decay would. This is the content of the statement in §4 that the numerical-freeze and decay-is-necessary accounts describe one limit.

E.4 When a Single Noise Scale Exists

The factorization $\mu/D = -U'/T$ of §3 needs U independent of the hyperparameters and T independent of position. From Eq. (53), for the gradient channel,

$$\frac{\mu_i^{\nabla}}{D_i} = -\frac{2}{\eta}\,\mathrm{sgn}(\partial_i L)\; s_i\big(\sqrt{1+s_i^2}+e_i\big). \tag{54}$$

The learning rate factors out. The batch size does not: it enters through $s_i \propto \sqrt{B}$ and $e_i \propto \sqrt{B}$, which also carry the position dependence. A valid split exists only where the right-hand side is a pure power of s_i or of e_i, i.e. in the three limits:

$$\text{(N)}\quad s\sqrt{1+s^2} \to s = \sqrt{B/c}\,|\partial L| \qquad \Rightarrow\; T = \frac{\eta}{\sqrt{B}},\quad U' = \frac{2\,\partial L}{\sqrt{c}}, \tag{55}$$

$$\text{(S)}\quad s\sqrt{1+s^2} \to s^2 = \frac{B(\partial L)^2}{c} \qquad \Rightarrow\; T = \frac{\eta}{B},\quad U' = \frac{2\,\partial L\,|\partial L|}{c}, \tag{56}$$

$$\text{(E)}\quad s\,e = \frac{\varepsilon B\,|\partial L|}{c} \qquad \Rightarrow\; T = \frac{\eta}{\varepsilon B},\quad U' = \frac{2\,\partial L}{c}. \tag{57}$$

At intermediate s the effective batch exponent, $\partial \ln[s\sqrt{1+s^2}]/\partial \ln B = \frac{1}{2} + s^2/[2(1+s^2)] \in (\frac{1}{2}, 1)$, depends on position, and no constant T reproduces Eq. (54). A batch-scaling experiment on AdamW therefore returns an exponent between $\frac{1}{2}$ and 1 whose value measures s at the bottleneck: 0.7 means $s \approx 0.8$, not a failure of the framework. Because s_i and e_i are read from the stored buffers (Appendix A.8), the applicable limit is determined by measurement.

Two further remarks. First, in limit N the noise parameters of a gradient-driven stage scale as $\Psi, \Theta \propto \sqrt{B}/\eta$, since $v \propto \eta\sqrt{B}$ while $D = \eta^2/2$ carries no B; only a decay-driven stage has $\Psi, \Theta \propto 1/\eta$ with no batch dependence, because there $v = \eta\lambda_{\rm wd}$. (The main-text sentence "$\Psi, \Theta \propto 1/\eta$ with no batch-size dependence" under (N) is the decay-driven case.) Second, when both channels act, the decay drift $-\eta\lambda_{\rm wd}\theta$ divided by $D = \eta^2/2$ gives $\Phi'_{\rm wd} = 2\lambda_{\rm wd}\theta/\eta$ with no B at all, so the two channels do not share a temperature: the regime-C exponent under AdamW-N in Table 1, $(2/\eta)\Delta(\sqrt{B/c}\,L + \frac{1}{2}\lambda_{\rm wd}\theta^2)$, has B inside the landscape rather than in a prefactor. Increasing B under AdamW steepens the loss part of the barrier relative to the decay part; under SGD it cools both equally.

E.5 The Two Decay Channels

Decoupled (AdamW). The decay term enters after the preconditioner and is exactly the SGD term $-\eta\lambda_{\rm wd}\theta_i$. On the norm coordinate $\theta' = \ln\rho$, $\rho_{t+1} = (1-\eta\lambda_{\rm wd})\rho_t$ from this term alone, so $\Delta\ln\rho = \ln(1-\eta\lambda_{\rm wd}) = -\eta\lambda_{\rm wd} + O(\eta^2\lambda_{\rm wd}^2)$: a constant drift $v = \eta\lambda_{\rm wd}$, $\Omega = 0$, exponential shrinkage at rate $\eta\lambda_{\rm wd}$, and the mean Eq. (**??**) identical across optimizers and across all three Adam limits. The seed statistics are not identical, because $\Psi = v\ell/D$ inherits the optimizer's diffusion:

$$\mathrm{CV}_B^{\rm SGD} = \sqrt{\frac{2}{\Psi}} = \sqrt{\frac{\eta\,c}{B\,\lambda_{\rm wd}\,\ell}}, \qquad \mathrm{CV}_B^{\rm AdamW\,(N)} = \sqrt{\frac{\eta}{\lambda_{\rm wd}\,\ell}}, \tag{58}$$

using $D = \eta^2 c/2B$ and $D = \eta^2/2$ respectively; they differ by $\sqrt{c/B}$. Identical means with different spreads is a separation no account phrased in terms of the mean can make.

Coupled (L_2 penalty through Adam). The penalty $\frac{1}{2}\lambda_{\mathrm{wd}}\|\theta\|^2$ is added to the loss, so its gradient $\lambda_{\mathrm{wd}}\theta_i$ passes through the buffers: $\hat{v}_i \to (\partial_i L + \lambda_{\mathrm{wd}}\theta_i)^2 + c_i/B$. On a memorized plateau $\partial_i L$ collapses while $\lambda_{\mathrm{wd}}\theta_i$ does not, so $\sqrt{\hat{v}_i} \to \lambda_{\mathrm{wd}}|\theta_i|$ (measured $1.16\,\lambda_{\mathrm{wd}}|\theta_i|$ on the host of Fig. 6b, Appendix A.8) and the decay signal ratio $s_i^{L_2} = \sqrt{B}\lambda_{\mathrm{wd}}|\theta_i|/\sqrt{c_i} \gg 1$: every coordinate is in limit S with respect to its own decay, and

$$\mu_i \to -\eta\,\mathrm{sgn}(\theta_i). \tag{59}$$

Three things follow. The drift is constant along the run, $\Omega = 0$, whatever the curvature of the loss. λ_{wd} has cancelled: the nominal decay strength sets nothing but the (fast) approach to this limit. And the shrinkage is linear rather than exponential: from $\rho\dot{\rho} = \sum_i \theta_i\dot{\theta}_i = -\eta\|\theta\|_1$,

$$\dot{\rho} = -\eta\,\frac{\|\theta\|_1}{\rho} \qquad (\le \eta\sqrt{P} \text{ by Cauchy–Schwarz}), \tag{60}$$

which is the rate quoted in §3.2; it is an equality in the pure sign-descent limit and an upper bound when some coordinates have $|\mu_i| < \eta$. On the host of Fig. 6b the norm falls at 0.095 per step from $\approx$99 to below 20 (secant), against 0.103 from a fit over the same window, and the coordinate $\ln\rho$ that made decoupled decay a flat stretch is not a slow coordinate of the coupled system at all: the run is carried to the origin in $\rho_0/(\eta\|\theta\|_1/\rho)$ steps regardless of whether a generalizing state lies on the way, which is the no-target phase N2 of Appendix F.

E.6 Instability Bursts

Bursts (Appendix A.7) are fast excursions off the slow coordinate; the kinetics treat them as part of the plateau's fast dynamics. Their origin differs between the Adam limits, and the difference is measurable.

Limit E: the adaptive edge of stability. Below the floor the plateau is gradient descent with normalized momentum β_1 at step η/ε. Its discrete-time stability is set by the preconditioned sharpness $\kappa_A := \eta\,\lambda_{\max}(P^{1/2}HP^{1/2})$, which for this update has threshold $(2 + 2\beta_1)/(1 - \beta_1) = 38$ at $\beta_1 = 0.9$ (**?**), against 2 for plain gradient descent. The gauge panels store κ_A (Appendix A.5). On the modular-addition host between bursts the plateau medians are $\kappa_A = 15.4$, 19.4 and 24.5 at $\eta = 0.5$, 1 and 2×10^{-3} ($n = 38$, 40, 36 panels; interquartile ranges 13–16, 18–22 and 20–30), and on the group-composition host 36.1 ($n = 28$), i.e. 0.4–0.95 of the threshold. As the memorized solution sharpens the plateau climbs toward the edge, overshoots, the gradient spikes, $\hat{v}$ refills and the step collapses, the sharpness drops, and the cycle restarts. κ_A grows sub-linearly with η (a factor 1.6 for a factor 4 in η) because $\lambda_{\max}(H)$ itself falls from 4.2 to 2.4×10^{-4}: the run settles on flatter memorizing solutions at larger steps, the self-regulation familiar from the edge-of-stability literature. At $\lambda_{\mathrm{wd}} = 0$ the plateau's κ_A falls through 1 by step 2,500 and reaches 0.3–0.7 by step 3,000–4,500 as $\lambda_{\max}(H)$ collapses to 4–7×10^{-6}: with the loss underflowed there is no curvature left to be unstable against, and no bursts, which is the frozen phase. Bursts at the floor are therefore excursions of a deterministic map past its stability limit, not a noise source in the sense of D; they are quasi-periodic, their spacing is set by the sharpening rate rather than by β_2, and they need not coincide with the escape.

Limit N: the $\hat{v}$-decay loop. Above the floor a plateau whose gradient drops sharply at step 0 has a buffer that decays as $v_t \approx \beta_2^t v_0$ while its input is negligible, so the step inflates geometrically, $\eta/\sqrt{v_t} = (\eta/\sqrt{v_0})e^{t(1-\beta_2)/2}$, until it reaches the stability edge; writing F for the inflation needed,

$$n_{\mathrm{burst}} \approx \frac{2\ln F}{1-\beta_2} = 2\tau_2 \ln F, \qquad D_{\mathrm{burst}} \approx \frac{\delta^2}{2n_{\mathrm{burst}}} = \frac{\delta^2(1-\beta_2)}{4\ln F}, \tag{61}$$

where δ is the displacement per burst. This mechanism predicts a spacing $\propto (1-\beta_2)^{-1}$ and no B dependence, unlike sampling noise. It is a scaling argument (neither F nor δ is fixed by it) and it applies only where $\hat{v}$ is above the floor; on floor-limited plateaus the step is pinned at η/ε and this loop is inactive, which is why the edge-of-stability description above, not this one, is the relevant account there. The minibatch Power config, the host whose bursts we count (Appendix A.7; a median of 5–9 bursts per seed before N^* at $\lambda_{\mathrm{wd}} \le 0.3$), reads (N) on its last pre-escape checkpoint (Table 4) but has checkpoints at the floor earlier on the plateau, so we do not assign its bursts to either mechanism.

E.7 Derivation of the Optimizer Columns of Table 1

Each entry follows three steps: take (μ, D) from Eqs. (10) and (53), expand μ at the bottleneck to read v and λ (or integrate μ/D for $\Delta\Phi$), and substitute into the regime law of §2.3. AdamW entries are in limit N unless tagged; c is taken constant along the stage.

Local rates. Differentiating the drift, $\lambda = \mu'(\theta^*)$, and with c slowly varying,

$$\lambda_{\mathrm{SGD}} = -\eta\big(\partial_\theta^2 L + \lambda_{\mathrm{wd}}\big) = -\eta\kappa_*, \qquad \lambda_{\mathrm{AdamW\,(N)}} = -\eta\Big(\sqrt{B/c}\,\partial_\theta^2 L + \lambda_{\mathrm{wd}}\Big) = -\eta\kappa_*^A, \tag{62}$$

where κ and κ^A are the curvatures defined in the caption of Table 1. A ridge has $\kappa_* < 0$ and $\lambda > 0$; a well $\kappa_a > 0$; a barrier top $\kappa_b < 0$ with $\lambda_b = -\eta\kappa_b > 0$. The AdamW curvature carries $\sqrt{B/c}$: the preconditioner rescales the loss curvature by the inverse noise scale but leaves the decay curvature alone, so on a ridge shaped by decay (λ_{wd} dominant in κ) the two optimizers have the same rate, and on a ridge shaped by the loss AdamW's rate is $\sqrt{B/c}$ times SGD's.

Regime A. From Eq. (7) with the rates above, $\mathbb{E}[N_A^*] = \ln(1+\Omega)/(\eta|\kappa_*|)$ for SGD and $\ln(1+\Omega)/(\eta|\kappa_*^A|)$ for AdamW-N. $\Omega = \ell/u_0$ is η-independent in both, since η rescales v and λ together; under SGD it is also B-independent, under AdamW-N it depends on B through κ^A only when loss and decay curvatures compete. In limit S the gradient drift saturates at η and is constant along the stage, so the ridge is traversed as a flat stretch: $\Omega = 0$, $\mathbb{E}[N^*] = \ell/\eta$, the (S) entry. The CV column is the same for all optimizers because it depends on (Ω, Ψ) and the initialization spread only.

Regime B. Decay-driven, on $\theta = \ln\rho$: $v = \eta\lambda_{\mathrm{wd}}$ for both optimizers (§E.5), $\ell = \ln(\rho_0/\rho_*)$, hence Eq. (**??**) in both columns, with the CVs of Eq. (58). Gradient-driven: $v = \eta|\partial_\theta L|$ for SGD, giving $\ell/(\eta|\partial_\theta L|)$, and $v = \eta s/\sqrt{1+s^2}$ for AdamW at any s with $e \ll 1$, giving $\ell\sqrt{1+s^2}/(\eta s)$, which interpolates between $\ell\sqrt{c/B}/(\eta|\partial_\theta L|)$ in limit N and ℓ/η in limit S.

Regime C. Integrating μ/D between the well and the top,

$$\Delta\Phi_{\mathrm{SGD}} = \int_{\theta_a}^{\theta_b} \frac{\eta(\partial_\theta L + \lambda_{\mathrm{wd}}\theta)}{\eta^2 c/2B}\,\mathrm{d}\theta = \frac{2B}{\eta c}\,\Delta\Big(L + \tfrac{1}{2}\lambda_{\mathrm{wd}}\theta^2\Big), \tag{63}$$

$$\Delta\Phi_{\mathrm{AdamW\,(N)}} = \int_{\theta_a}^{\theta_b} \frac{\eta(\sqrt{B/c}\,\partial_\theta L + \lambda_{\mathrm{wd}}\theta)}{\eta^2/2}\,\mathrm{d}\theta = \frac{2}{\eta}\,\Delta\Big(\sqrt{B/c}\,L + \tfrac{1}{2}\lambda_{\mathrm{wd}}\theta^2\Big), \tag{64}$$

and the prefactor of Eq. (9) with $|\lambda| = \eta\kappa_a$, $\lambda_b = \eta|\kappa_b|$ is $2\pi/(\eta\sqrt{\kappa_a|\kappa_b|})$ (SGD) or $2\pi/(\eta\sqrt{\kappa_a^A|\kappa_b^A|})$ (AdamW-N). Hence $\ln\mathbb{E}[N_C^*] = \Delta\Phi - \ln\eta + \mathrm{const}$ in both cases: the barrier is read from the slope of $\ln\mathbb{E}[N_C^*]$ against B/η (SGD) or against $\sqrt{B}/\eta$ at fixed decay (AdamW-N), and the $-\ln\eta$ is the prefactor correction, visible only in fits spanning more than a decade of η. Only this row places hyperparameters in an exponent, so it is the only regime in which a modest change of η or B moves the delay by orders of magnitude; every other entry is a power law. The $\sqrt{B}$ against B contrast in the exponent is the largest optimizer effect the framework predicts, and it follows entirely from the second moment equilibrating to Eq. (51) rather than to $(\partial L)^2$. At intermediate s the exponent is $\int(2/\eta)\,s\sqrt{1+s^2}\,\mathrm{sgn}(\partial L)\,\mathrm{d}\theta$ and interpolates between the two, as in §E.4; and at the floor (limit E) it is the SGD exponent with $\eta \to \eta/\varepsilon$, which at full batch is infinite ($D = 0$): regime C does not exist there.

F Three Non-Grokking Phases

The three regimes of §2.2 describe how a delay is produced. A run can also fail to show the memorize-then-generalize shape at all, and the framework says there are exactly three ways this happens. They are not regimes, because none of them is a value of Ω: two are statements about the endpoints of the stage and one about its length relative to the budget or the readout. We name them N1–N3, give the condition for each in terms of the plateau quantities, and list the reproduced cells that realize them (Table 10). Every count below is over deduplicated seeds on the paper clock of Appendix A.2.

N1: budget-frozen. A generalizing state exists and the run is moving toward it, but the expected delay exceeds the training budget N_{bud}. In the transit-time regimes the condition is simply $\int_{\theta^*}^{\theta_e} \mathrm{d}\theta/\mu(\theta) > N_{\mathrm{bud}}$, and its sharpest form is the frozen limit $\mu \to 0$: on the ε-floor host without decay the gradient drift $-(\eta/\varepsilon)\partial_\theta L$ vanishes with the underflowing loss (§E.3), and $0/8$ seeds escape in 4×10^4 steps with test accuracy never above 0.24 while training accuracy is 1.0. In regime C the condition is

$$\Delta\Phi > \ln\Big[\frac{N_{\mathrm{bud}}\sqrt{|\lambda|\lambda_b}}{2\pi}\Big], \tag{65}$$

from Eq. (9); because the delay is exponential in the barrier, a factor of ten in budget buys only $\ln 10 = 2.3$ units of $\Delta\Phi$, and seeds are then censored rather than slow. The group-composition host (③) is budget-frozen in our reproduction: $0/32$ seeds pass the readout in 6×10^4 steps, with training accuracy 1.0 on every seed and a median maximum test accuracy of 0.009 (best seed 0.14). Whether that cell is a long transit or a barrier cannot be decided from censored seeds; its plateau sits at $\kappa_A = 36$, close to the adaptive stability edge of 38 (§E.6), and it is the one reproduced setting where a longer budget is the only test. An apparent failure to grok is, in this phase, a statement about compute.

N2: no target. No generalizing state lies on the run's path, so the stage has no exit and the delay is undefined rather than long. This phase is orthogonal to Ω: the plateau quantities describe the motion, not whether the destination exists. It is the phase below the critical data fraction of Liu et al. (2022a), and we realize it in three other ways. (i) Coupled L_2 decay through Adam turns the drift into sign descent (§E.5); the norm is carried linearly to zero, the network never even memorizes (maximum training accuracy 0.34 over 16 seeds), and $0/16$ seeds generalize in 10^5 steps. (ii) On the diagonal linear net with a *dense* target, the large-initialization start already generalizes (test accuracy 0.99 on the first row) and the decay-driven transport toward the min-norm state moves it *away*: test accuracy falls to 0.68 by 6×10^4 steps on all 32 seeds. The kinetics are those of a flat stretch, identical to the sparse-target cell that groks; only the endpoint differs. (iii) At $\lambda_{\mathrm{wd}} = 3$ on the Power config, the decay caps the attainable norm: the median maximum training accuracy is 0.98 and the median maximum test accuracy 0.92, so $24/32$ seeds never reach the 0.99 readout although they generalize partially. A weaker readout would count them; the state the readout asks for does not exist under this decay.

N3: no plateau. The stage is traversed faster than the readout resolves it, or the generalizing state is reached before memorization completes, so there is no visible delay although the kinetics are unchanged. The remaining $8/32$ seeds at $\lambda_{\mathrm{wd}} = 3$ on the Power config are of this kind: they pass the test readout $1.31\times$ later than the training readout (median $1{,}350$ against $1{,}000$ steps), against ratios of 51, 16 and 3.2 at $\lambda_{\mathrm{wd}} = 0.1$, 0.3 and 1 where the same host groks. The transport law Eq. (**??**) continues smoothly through this point ($\lambda_{\mathrm{wd}} N^* \approx 2{,}400$ at $\lambda_{\mathrm{wd}} = 0.3$ and 1, which predicts 800 steps at $\lambda_{\mathrm{wd}} = 3$, below the memorization time); what disappears is the gap between the two readouts, not the mechanism. Sharpening the readout (a higher threshold, a stricter sustain rule) can move a cell from N3 back into a measurable delay, which is why the readout is part of the definition of N^* (Appendix A.2).

Relation to the four phases of Liu et al. (2022a). Their *comprehension* (immediate generalization) is N3; their *memorization* is N1; their *confusion* is N2 or the boundary of N1 and N2; and their *grokking* is the union of regimes A, B and C. The two classifications are complementary: theirs is by outcome within a budget, ours by why the outcome occurs, and the second is what a hyperparameter change acts on. Reducing the decay or increasing the budget moves N1 cells into grokking without touching the regime; adding data or changing the target moves N2 cells; sharpening the readout moves N3 cells. None of these changes Ω, which is why we keep the phases out of the regime map of Fig. 2.

G Gaussian-Increment Validity

The kinetics of §2 treat one optimizer step as a drift plus a noise increment whose third and higher cumulants are negligible (§D.1). That step is only as good as the central limit theorem applied to a

Table 10: Non-grokking phases realized in the reproduced settings. Counts are deduplicated seeds on the paper clock; "budget" is the training budget in steps.

Phase	Cell	Outcome	Diagnostic
N1	① mod-add, full-batch AdamW, $\lambda_{\mathrm{wd}} = 0$	$0/8$ in 4×10^4	train acc. 1.0; test acc. ≤ 0.24; $\lambda_{\max}(H) \to 4 \times 10^{-6}$ and $\kappa_A \to 0.4$ by step 4,500: drift and curvature gone with the underflowed loss (limit E without decay)
N1	③ S_5 composition, full-batch AdamW	$0/32$ in 6×10^4	train acc. 1.0; median max test acc. 0.009 (best 0.14); $\kappa_A = 36$ at the adaptive edge
N2	⑨ mod-add MLP, Adam+L_2	$0/16$ in 10^5	sign descent removes the network: max train acc. 0.34, norm $\to 0$ linearly at 0.095/step
N2	⑦ diag. net, dense target, SGD	$32/32$ misgrok	test acc. 0.99 at $t = 0$, 0.68 at 6×10^4; same transport law as the sparse target that groks
N2	② Power config, $\lambda_{\mathrm{wd}} = 3$	$24/32$ censored	decay caps the norm: median max train acc. 0.98, max test acc. $0.92 < 0.99$ readout
N3	② Power config, $\lambda_{\mathrm{wd}} = 3$	$8/32$ pass	test/train readout ratio 1.31 (vs. 51, 16, 3.2 at lower decay); transport law extrapolates below the memorization time

minibatch average of per-example gradients: the batch must be large enough, relative to how non-Gaussian a single example's gradient is, for the averaged increment to be near-Gaussian. We check this on every run rather than assume it, and this appendix resolves the reference in the Limitations paragraph of §5.

Protocol. At every gauge checkpoint the harness computes the per-example gradient on a fixed probe batch and projects it into the 48-dimensional measurement subspace of Appendix A.5. For each coordinate i of that subspace the excess kurtosis of the per-example gradient is $\gamma_{2,i} = m_{4,i}/m_{2,i}^2 - 3$, with m_2 and m_4 the central moments over the probe batch. Cumulants of independent sums add, so the excess kurtosis of the batch-mean gradient that enters the update, the *increment kurtosis*, is

$$\hat{\gamma}_{2,i} = \gamma_{2,i}/B, \tag{66}$$

with B the dynamics batch size; this is the $k = 4$ case of Eq. (14). Under Adam the step is $-\eta P_i \hat{g}_i$; if P_i were constant over one step the per-coordinate value would be unchanged, since kurtosis is scale-invariant, but the probe frame is a rotation of the coordinates and P is not constant on floor coordinates, so for the Adam hosts we compute the kurtosis of the exact one-step update from the stored buffers instead (below). We treat $|\hat{\gamma}_{2,i}| \leq 1$ as inside the near-Gaussian window, $1 < |\hat{\gamma}_{2,i}| \leq 3$ as a caution zone in which quantitative delay laws are downgraded to signs and orderings, and $|\hat{\gamma}_{2,i}| > 3$ as outside the domain of the diffusion description, mirroring how the other validity monitors of the harness (the signal ratio s, the floor ratio $\sqrt{\hat{v}}/\varepsilon$, and κ_A) gate which claims a run supports. On a full-batch run there is no increment to check: the numbers we report there use the probe size as a stand-in and describe the shape of the per-example distribution itself.

Results. Table 11 collects the increment kurtosis on the constructed anchors of Appendix B.5, on the two full-batch AdamW hosts, and on every minibatch host of Table 3: the Power config and MNIST-1k under AdamW, the modular-addition MLP under AdamW, and the minibatch SGD hosts (MNIST-1k, sparse parity, diagonal net). Where checkpoints exist on both sides of the escape we compare a plateau checkpoint with a later one, so that a burst or the escape itself disturbing the Gaussian approximation would show up as a difference between the two.

Every minibatch host of the paper is therefore covered, on the plateau (the last saved checkpoint before the stop-rule crossing, and a mid-plateau checkpoint at $0.5N^*$ from the dense-trajectory reruns) and at the escape. Three conventions of the sweep matter for reading the table. First, for the Adam hosts the number is the kurtosis of the *exact* one-step update, sampled by Monte Carlo from the stored per-example gradients and the stored buffers (m, v), rather than of the raw gradient: with P frozen over the step the per-coordinate value would be unchanged, but the probe frame is rotated and 17–36% of the MNIST coordinates sit at the ε floor, where P spans eight decades and a frozen-P projection produces spurious violations (a single example dominating every direction, at the sample ceiling); letting $\hat{v}$ react to the same gradient inside the step, as the optimizer does, removes them. Second, the conservative column samples examples independently; the runs actually sample without replacement within an epoch, which reduces the fourth cumulant of the increment by $1 - 6f(1-f)$ with $f = B/N$ (a factor 0.42 on the Power config and the modular MLP, 0.96 on

Table 11: Increment kurtosis across the reproduced settings and the constructed anchors. Thresholds: caution $= 1$, violation $= 3$. Medians and maxima are over probe-subspace coordinates, pooled across seeds and checkpoints unless a range is given (plateau $\rightarrow$ late).

Cell	Optimizer	B	Median $\lvert\hat{\gamma}_2\rvert$	Max $\lvert\hat{\gamma}_2\rvert$	Verdict
Constructed anchors (S0)					
Regime A, outside / inside the ball	GD + injected noise	32^*	0.06	0.06	valid; two-point per-example law, $-2/32$ exactly
Regime B	GD + injected noise	32^*	$0.06 \rightarrow 0.08$	0.45	valid; rises $\sim$40% over the transport
Regime C	GD + injected noise	—	undefined	—	noise is Gaussian by construction (injected)
Full-batch hosts (no sampling increment; probe size as stand-in)					
① mod-add, $\lambda_{\mathrm{wd}} = 1$	AdamW	512^*	0.03	0.11	valid, see caveat
① mod-add, $\lambda_{\mathrm{wd}} = 0$	AdamW	512^*	0.005	0.02	valid, same caveat
Minibatch hosts					
⑨ mod-add MLP, SGD sibling	SGD	1024	0.0002	0.001	valid, near-exactly Gaussian
② Power config, $\lambda_{\mathrm{wd}} = 0.1$	AdamW	512	$0.06 \rightarrow 0.06$ (0.16 mid)	5.1 (one seed, $0.07N^*$; 3.4 under the epoch sampler)	near-Gaussian; tail coords. on one seed
② Power config, $\lambda_{\mathrm{wd}} = 0.3$	AdamW	512	$0.05 \rightarrow 0.05$ (0.09 mid)	2.7 (mid-plateau)	near-Gaussian; tail coords.
② Power config, $\lambda_{\mathrm{wd}} = 1$	AdamW	512	$0.09 \rightarrow 0.06$	3.0 (plateau)	near-Gaussian; tail coords.
④ MNIST-1k, $\lambda_{\mathrm{wd}} \in \{.003, .01, .03\}$	AdamW	200	$0.05 \rightarrow 0.05$	0.6	valid
④ MNIST-1k, $\lambda_{\mathrm{wd}} = 0.1$	AdamW	200	$2.1 \rightarrow 0.8$ (0.5 mid)	4.9; 91% of coords. > 1 on the plateau	**caution** (i.i.d. sampling); 0.17 / 0.80 under the epoch sampler
④ MNIST-1k, $\lambda_{\mathrm{wd}} \in \{.003, .01\}$	SGD + clip	200	0.004–0.007	0.03	valid; clip active on every minibatch
⑨ mod-add MLP	AdamW	512	$0.05 \rightarrow 0.06$	0.27	valid
⑤ sparse parity, $B = 32$	SGD	32	$0.02 \rightarrow 0.17$	0.30	valid
⑤ sparse parity, $B = 8$ / 128	SGD	8 / 128	$0.13 \rightarrow 0.55$ / $0.005 \rightarrow 0.04$	0.97 / 0.06	valid; tail only after escape at $B = 8$
⑦ diag. net, $B = 8$ / 16 (escape ckpt only)	SGD	8 / 16	0.78 / 0.34	3.4 / 1.6	near-Gaussian; 0.09 / 0.29 under the epoch sampler (48-example population)

*Probe or full-batch size used in place of a dynamics batch.

MNIST-1k, and negative, i.e. platykurtic, on the diagonal net at $B = 16$); we report the epoch-shuffle value where it changes the verdict. Third, all values are single-step; averaging over the momentum buffer multiplies the excess kurtosis by $(1 - \beta_1)/(1 + \beta_1) \approx 0.1$, so the table is an upper bound on what enters the kinetics. The Monte-Carlo floor is a median of about 0.05 and a maximum of about 0.2 over 48 coordinates.

The result is that no cell reaches the violation threshold and one reaches caution. On every AdamW host at the decay strengths the main text uses, on the modular MLP, on the clipped-SGD MNIST runs and on sparse parity at all three batch sizes, the plateau medians are 0.005–0.16, indistinguishable from Gaussian at the Monte-Carlo floor, and the escape checkpoints are as clean or cleaner. The Power config shows a minority of tail coordinates on individual seeds early on the plateau (one seed at $0.07N^*$ has a median of 0.88 and a maximum of 5.1 under i.i.d. sampling, 0.48 and 3.4 under the epoch sampler), which disappear by the escape. The one caution cell is MNIST-1k under AdamW at $\lambda_{\mathrm{wd}} = 0.1$: on the plateau 91% of the probe coordinates have $\lvert\hat{\gamma}_2\rvert > 1$ and 29% exceed 3 under i.i.d. sampling, driven by the preconditioned eigen-directions of the saved subspace; under the epoch-shuffle sampler the same checkpoint reads 0.17 (maximum 0.80). We keep the conservative verdict and flag that cell's quantitative Adam-track numbers accordingly (it is the cell where the probe and profile estimates of Ω also disagree, Appendix B.2). Sparse parity, the host with the smallest tilt ($\Psi \geq 16$), is Gaussian on the plateau at every batch size and develops a mild tail only after the escape at $B = 8$ (median 0.55, maximum 1.4 under the epoch sampler), when the loss has fallen and few examples carry gradient. Fig. 14 shows the per-cell values. One caveat applies to the older rows: the 512-example probe used there cannot register an increment kurtosis above $(512 - 3)/B$, which is 0.99 at $B = 512$; the new sweep uses the full training population, and it is that estimator, not the probe, that found the Power-config tail seeds.

Two results from the anchors are worth stating explicitly. The regime-B anchor's increment kurtosis is not flat over a run: it is 0.06 on the plateau and 0.08 by the end of the transport on all four seeds checked, a consistent rise of about 40%. This is the phase dependence the checkpoint-by-checkpoint protocol is built to catch; it stays far from the caution zone here, but a cell where such a rise continued toward it near the escape is what this monitor exists to flag before a delay law is quoted past where it applies. And the two regime-A anchors return the same value, 0.0625, at nearly every checkpoint, by construction rather than coincidence: the host has two free parameters and is probed with 32 examples, so the projected per-example gradient takes only two distinct values, a two-point law with excess kurtosis exactly -2, and dividing by 32 gives $-2/32$. This is the batch-averaging step of the central limit theorem visibly at work on a maximally non-Gaussian per-example distribution, and it makes the point that the relevant quantity is the increment, not the per-example gradient.

Where the check does not apply. The regime-C anchor is the one cell where this instrument has nothing to measure: seven of its eight checkpoints return no defined kurtosis, because the intrinsic

per-example gradient variance collapses toward zero on an interpolating plateau. This is the same collapse that makes injected rather than intrinsic noise necessary for a clean regime-C cell in the first place (Appendix B.5); the injected noise is i.i.d. Gaussian after the optimizer step by construction, so for regime C the assumption holds by design rather than by measurement. The full-batch hosts have no sampling increment at all; the values above describe the per-example distribution and should not be read as validating the increment, which is athermal there in any case (Table 5).

When the theory applies. Across every cell with genuine sampling noise the increment kurtosis stays inside the near-Gaussian window through both the plateau and the escape, so the delay and seed-statistics laws apply to them without qualification. The two places the check does not simply confirm validity are informative on their own terms: regime C's noise is Gaussian by construction, and full-batch cells have no increment to describe. Every run logs $\hat{\gamma}_2$ with the fraction of coordinates past each threshold, so a future cell drifting into the caution range is caught by the harness rather than by inspection.

H Connection with Physics

Every derivation in the main text is self-contained: it starts from the update rule and uses only elementary probability, so that no physical law is assumed and the argument is readable without that background. The core ideas, however, came to us from thermodynamics and non-equilibrium statistical mechanics, which is where the picture of training as a whole system with its own kinetics, and of grokking as one escape phenomenon within it, originates. Table 12 makes the connection explicit for readers who know the physics and want the fastest route to intuition. Each row names the object in this paper and the standard result it arrives at, without assuming it; the physics name is a mnemonic, not a justification, and the correspondence is exact only where stated. Two objects have no classical name: the shape variable Ω, which is specific to the reduction to a stage, and the fact that the temperature of the process is set by the optimizer and exists as a single number only in limits.

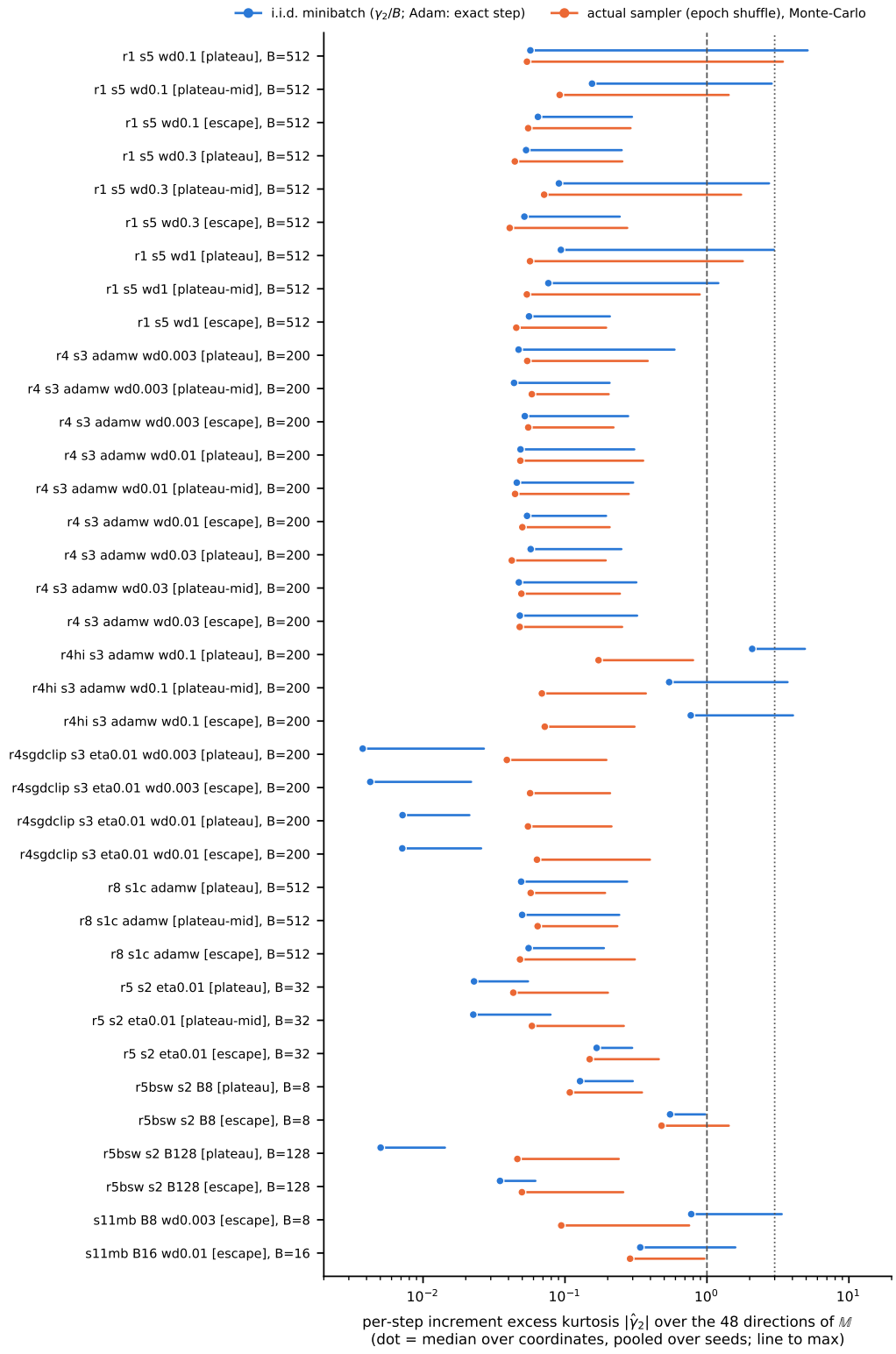


Figure 14: Increment kurtosis on the minibatch hosts, per cell and checkpoint (plateau, mid-plateau, escape): median and maximum over probe-subspace coordinates under the exact one-step update with i.i.d. sampling; the caution and violation thresholds are marked.

Object or result in this paper	Physics correspondence	Note
Drift and diffusion (μ, D) of one step, Eq. (2)	Kramers–Moyal expansion truncated at second order: the coefficients of a Fokker–Planck equation	Truncation is justified by the $(\eta/B)^{k-2}$ suppression of higher cumulants (§D.1), not assumed; Pawula's theorem makes second order the only consistent cut (**??**).
$\Phi = -\int \mu/D$	Log of the zero-current stationary density; in one dimension, the Freidlin–Wentzell quasipotential divided by the noise strength	A scalar diffusion on an interval carries no circulating current, so Φ has an equilibrium-like (Boltzmann) form although training is not at equilibrium (**?**).
Governing function Eq. (3) and the hierarchy Eq. (20)	Backward Kolmogorov (Pontryagin) equation for the mean first-passage time and its moments	Obtained here by first-step conditioning; the integrating-factor solution is the textbook one (**?**).
Tilt $\Psi = v\ell/D$ and curvature $\Theta = \lambda\ell^2/D$	Ψ is a Péclet number (drift transport over diffusive transport across the stage); $\Theta = (\ell/\ell_T)^2$ is the stage width in units of the noise-ball radius	$\Psi \gg 1$ is the drift-dominated (ballistic) transport limit; $\Theta \gg 1$ means the noise ball is small against the stage.
Shape variable $\Omega = \Theta/\Psi = \lambda\ell/v$	No classical name: the fractional change of the drift across the stage	Specific to the reduction to a stage with an entrance and an exit; D cancels, so it is a property of the deterministic flow.
Regime A, Eqs. (7) and (28)	Decay of an unstable state under linear instability plus noise; Suzuki's scaling theory	The logarithm in the offset, the factor two between the outside and inside-ball laws, and the log-half-normal shape are the classical results (**?**).
Regime B, Eq. (8) and $\mathrm{CV}_B = \sqrt{2/\Psi}$	First passage of a Wiener process with constant drift: the inverse Gaussian (Wald) law	Classical transport result; the late-time hazard $\Psi/4\mathbb{E}[N^*]$ is its exponential tail (**?**).
Regime C, Eq. (9)	Kramers escape rate in the overdamped limit; Arrhenius exponent with the Kramers curvature prefactor	The factor χ of Eq. (39) is the splitting probability at the top; D is not a physical temperature but comes from the optimizer map (**??**).
Constant hazard, $\mathrm{CV}_C = 1$	Escape as a Poisson process; a first-order rate constant in chemical kinetics	Consequence of well relaxation being fast against barrier crossing, proved from the spectral gap of Eq. (19).
$\mu/D = -U'/T$ (§3), where it exists	Einstein relation, fluctuation–dissipation: diffusion as mobility times a temperature	Holds unconditionally for SGD ($T = \eta/B$) and momentum SGD; for AdamW only in the limits N, S, E (§E.4), with $T = \eta/\sqrt{B}$ in limit N recovering the square-root scaling rule of **?**.
Reduction to a slow coordinate (§2.4, §D.9)	Adiabatic elimination of fast variables (Zwanzig projection); averaging principle for fast–slow SDEs	The $O(1/g)$ bound is the adiabaticity condition; the memory term is the projected-operator correction (**??**).
Momentum SGD as SGD at $\eta/(1-\beta)$ (§E.2)	Unit DC gain of a low-pass filter; Markovian limit of a short memory kernel	Fails when the filter memory is comparable to the local relaxation time $1/\lvert\lambda\rvert$, the Grote–Hynes regime; not reached on any measured setting.
Instability bursts at the ε floor (§E.6)	Loss of linear stability of a discrete map past its stability edge	The threshold $(2 + 2\beta_1)/(1 - \beta_1)$ is the adaptive edge of stability (**?**); bursts are deterministic excursions, not thermal noise.

Table 12: Correspondence between results derived in §2–§3 and standard objects in non-equilibrium statistical mechanics. Every left-column result is proved from the SGD or AdamW update rule; the right column is an intuition anchor for physics-literate readers.