PERCOLATION DYNAMICS IN OPTIMIZATION: VARIANCE CASCADES AND NESTED SYMMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the dynamics of Stochastic Gradient Descent (SGD), which is known to steer deep neural networks toward invariant sets that correspond to simpler subnetworks. How this steering unfolds over time remains poorly understood. We answer this by modeling the stochastic gradient flow (SGF) as a percolation process, in which nested architectural symmetries force subnetworks to merge in discrete blocks rather than by single-edge attachment. These structural transitions register as variance spikes in a macroscopic order parameter echoing physical phase transitions. We further state sufficient conditions under which the trapping argument carries over to Adam and AdamW under heavy-tailed gradient noise and measure them on a trained Transformer.

1 INTRODUCTION

Unlike classical statistical learning, deep learning systems rely on a complex interaction between the dataset, architecture, and choice of optimizer. Furthermore, they are known to achieve strong generalization without explicit regularization, a phenomenon attributed to implicit biases introduced during training. Recent work has established that Stochastic Gradient Descent (SGD) is a central source of this implicit bias collapsing networks onto invariant sets that behave like much simpler subnetworks (Wei et al., 2008; Chen et al., 2023). However, the dynamics by which these invariant sets are reached are poorly understood.

This gap limits our ability to mechanistically explain anomalous training behaviors. Delayed generalization in transformer networks (Power et al., 2022; Liu et al., 2022) is one example. A network may memorize the training data perfectly, but it only discovers a generalizing solution after thousands of epochs, challenging the classical view of smooth optimization. Similar patterns can also be seen in exact deep linear networks (Saxe et al., 2013) and task shifts (Goodfellow et al., 2013; Kirkpatrick et al., 2017). This suggests that the topological structure of these dynamics contains crucial information about how optimization progresses.

To chart this evolution, we describe the optimization continuously with a stochastic differential equation (SDE) that records how parameters drift and diffuse over time (Li et al., 2017). Architectural symmetries produce invariant sets which draw parameter trajectories (Chen et al., 2023). Independent parameters entering these sets come together and fuse into equivalence classes, and the converging weights collide and merge, folding the parameter space into a simpler topology in the settings we examine (Naitzat et al., 2020). We represent this dynamics as a percolation process where isolated components link and fuse, potentially causing sudden macroscopic transitions (Achlioptas et al., 2009; Chen et al., 2014). We develop this framework for SGD, where the mechanism is cleanest to state and prove, but most large-scale training uses adaptive optimizers, so we state sufficient conditions under which the trapping and cascade arguments carry over to Adam and AdamW with heavy-tailed gradient noise.

Building on this percolation framework, our contributions are as follows:

1. **Percolation Model of Topological Condensation:** We construct a framework that maps stochastic gradient flow near invariant sets onto a graph that tracks merges and splits (Reeb graph), which is then renormalized into a percolation process. Nested architectural symmetries force merges across subtrees to occur as block merges, alongside single-edge

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.

attachment within them, and we give the condition under which this discontinuity survives as the network grows large.

2. **Variance Peaks and Topological Cascades:** We use relative variance across training trajectories to isolate discrete microtransitions. Each block merge of multiplicity n produces a peak of height $(n-1)^2/(4n)$ at a fixed quantile of its hitting law. Modeling the block merges as a coalescence process, we show that the peaks form a geometric cascade exhibiting the Discrete Scale Invariance (DSI) of physical critical phenomena (Sornette, 1998). The ratio between successive levels is n in the merge fraction and $n^{1-\gamma_K}$ in training time.
3. **Empirical Observations:** We test this framework across environments and observe the predicted cascade structure. In toy models we track parameters clustering under shifting data distributions, and a comparison of a flat layer with a binary tree built from the same units shows that nested symmetry produces the macroscopic block merges and that a weight-decay-paced clock spaces the levels uniformly in time. Cascades with graded statistical support appear in MLPs on UCI tabular classification and vision benchmarks and in a Transformer on modular arithmetic, on which we also measure the Adam trapping condition.

Key takeaways. Implicit bias is frequently characterized as a continuous shift toward simpler networks. We show that it instead happens via discrete merges of subnetworks, where each merge is marked by a peak in relative variance. Deep networks inherit nested symmetry from composition, leading to discontinuous joint merges of groups of subnetworks that are absent in a single flat layer. Such merges follow a geometric cascade and are evenly spaced under an implicit clock. We observe that in AdamW the weight decay additionally functions as a driver of subnetwork merges.

2 Stochastic Collapse in SGD

As our principal aim is to study the transient dynamics of a network, we characterize the invariant sets of the training dynamics and the conditions under which they are attractive. We rely on permutation symmetries for this purpose.

Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote a model's parameter vector and $\mathcal{L}(\boldsymbol{\theta})$ a nonconvex loss, for example a neural network's empirical risk. At iteration t, optimizing $\mathcal{L}$ with learning rate η over a mini-batch $\mathcal{B}_t$ yields the discrete update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta}_t).$$

By separating the exact full-batch gradient $\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ from the zero-mean batch noise, we approximate this discrete sequence continuously using a Stochastic Gradient Flow (SGF) modeled by an Itô SDE (Li et al., 2017):

$$d\boldsymbol{\theta}_t = -\nabla\mathcal{L}(\boldsymbol{\theta}_t)dt + \sqrt{\eta\Sigma(\boldsymbol{\theta}_t)}dW_t, \tag{1}$$

where W_t is a standard d-dimensional Wiener process and $\Sigma(\boldsymbol{\theta})$ is the position-dependent noise covariance matrix of the mini-batch gradients.

Architectural symmetries, such as permutation invariances among neurons, generically create flat, degenerate regions in the parameter space. We formally define these regions as invariant sets, following (Chen et al., 2023).

Definition 2.1 (Invariant Set)**.** A Borel set $A \subset \mathbb{R}^d$ is invariant for a stochastic process $\{\boldsymbol{\theta}_t\}$ if every trajectory initialized in A stays in A almost surely. For discrete SGD, t ranges over the iterations.

We note that for affine invariant sets, invariance under SGD carries over exactly to SGF.

Proposition 2.2 (Affine Invariance in SGF)**.** *Consider an SGD process where the individual sample gradients $\nabla\ell(\cdot; x_i, y_i)$ are Lipschitz continuous and bounded. If an affine subset $A \subseteq \mathbb{R}^d$ is invariant under every sample gradient step, so that each $\nabla\ell(\cdot; x_i, y_i)$ is tangent to A on A, then A is an invariant set of the continuous SGF process.*

Proof Sketch (Chen et al., 2023). As the symmetric root of $\eta\Sigma$ is only Hölder-$\frac{1}{2}$ on A, we drive the SDE by the Lipschitz factor $G(\boldsymbol{\theta})$ of the centered sample gradients, $GG^\top = \eta\Sigma$, which admits a unique strong solution. Since the drift and G are tangent to A on A, the projected SDE matches the original one and trajectories stay in A (Appendix A). □

We note that our analysis uses SGF while our experiments run on discrete SGD, which shares the same invariant sets (Proposition 2.2). The discretization error vanishes on the invariant set (Li et al., 2017; Chen et al., 2023). Further, since the cascade predictions are ratios of times, they are unaffected by the conversion from steps to continuous time.

Once parameters drift near these invariant sets, the model predicts a qualitative change in the optimization dynamics. If the mini-batch variance $\Sigma(\boldsymbol{\theta})$ decays as $\boldsymbol{\theta}$ approaches A, an inward pull emerges that can trap trajectories near A. We formalize this as stochastic attractivity:

Definition 2.3 (Stochastic Attractivity via Drift Dominance)**.** Write $\boldsymbol{x} = \boldsymbol{\theta}^{(i)} - \pi_A(\boldsymbol{\theta}^{(i)})$ for the transverse vector, $Y_t^{(i)} = \|\boldsymbol{x}_t\|_2^2$ for the transverse distance process, and $\Sigma_\perp = P^\perp \Sigma_{ii} P^\perp$ for the transverse noise covariance. An invariant set $A \subset \mathbb{R}^{d_i}$ of the block $\boldsymbol{\theta}^{(i)}$ (Definition 3.1) of the stochastic process $\{\boldsymbol{\theta}_t \in \mathbb{R}^d : t \geq 0\}$ is locally stochastically attractive of order $\alpha \in (0, 1]$ if there exists a basin $\mathcal{N}(A, \epsilon)$ such that for all $\boldsymbol{\theta} \in \mathcal{N}(A, \epsilon) \setminus A$,

$$-2\,\boldsymbol{x}^\top \nabla_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}) + \eta \operatorname{tr} \Sigma_\perp(\boldsymbol{\theta}) \;\leq\; 2(1-\alpha)\,\eta\, \frac{\boldsymbol{x}^\top \Sigma_\perp(\boldsymbol{\theta}) \boldsymbol{x}}{\|\boldsymbol{x}\|_2^2}. \tag{2}$$

A is strictly attractive with rate $\kappa > 0$ if the gap between the two sides is at least $(\kappa/\alpha)\|\boldsymbol{x}\|_2^2$, so that $\mathscr{G}(Y^{(i)})^\alpha \leq -\kappa\,(Y^{(i)})^\alpha$. The infinitesimal generator applied to $(Y_t^{(i)})^\alpha$ equals $\alpha(Y_t^{(i)})^{\alpha-1}$ times the difference of the two sides of equation 2 (Appendix B, Equation equation 15), so equation 2 states that the inward drift, helped by the Itô correction of the concave map $Y \mapsto Y^\alpha$, overpowers the outward transverse diffusion.

Sending $\alpha \downarrow 0$ recovers the collapse threshold $\mu \leq \zeta^2/2$ of Appendix E (Ziyin et al., 2022; Chen et al., 2023). With isotropic noise in k transverse directions, noise alone traps only for $k < 2(1-\alpha)$, and beyond that the inward drift or weight decay must carry the condition. SGF behaves as a non-equilibrium system driven by anisotropic, position-dependent noise (Chaudhari & Soatto, 2018; Mandt et al., 2017; Mori et al., 2022), which motivates the analysis below.

3 Topological Dynamics of SGF via Attractor Equivalence

To characterize how the network approaches these sets, we analyze the convergence of subnetworks under symmetries, e.g. the neurons of a layer, which are interchangeable under permutation and whose redundancy is exploited in pruning and dropout (Şimşek et al., 2021; Frankle & Carbin, 2019; Srivastava et al., 2014).

Definition 3.1 (Subnetwork Partition)**.** Partition the global parameter vector into N distinct subnetworks, $\boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}]$, such that each $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^{d_i}$ evolves over a marginal function space $C([0, \infty), \mathbb{R}^{d_i})$.

We model the collapse of a network as a sequence of topological phase transitions between its subnetworks. To this end, we track the marginal path measure of each subnetwork over a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. We treat subnetworks separately, as symmetry gives each pair of identical subnetworks its own invariant set, while neural network training is documented to possess a simplicity bias that decouples parameter blocks (Şimşek et al., 2021; Huh et al., 2021; Shah et al., 2020).

The constructions below use only the transverse distance of each subnetwork to its invariant set, so no separate bound on the cross-covariance between blocks is needed: per-sample symmetry makes the pairwise transverse noise vanish on the coincidence set. To formalize structural binding, we project these trajectories into a shared quotient space by tracking their squared transverse distance $Y_t^{(i)} = \|\boldsymbol{\theta}_t^{(i)} - \pi_A(\boldsymbol{\theta}_t^{(i)})\|_2^2$ to a target invariant set A. Under stochastic attractivity, the inward deterministic gradient dominates the transverse diffusion, opposing outward drift. We show this is enough to trap the subnetwork within the local basin.

Theorem 3.2 (Local Supermartingale Trapping)**.** *Given a stochastically attractive invariant set A of order α, there exists a local basin $\mathcal{N}(A, \epsilon)$ such that the stopped transverse process $(Y_{t\wedge\tau_\epsilon}^{(i)})^\alpha$ operates as a non-negative supermartingale, where $\tau_\epsilon = \inf\{t \geq t_0 : Y_t^{(i)} \geq \epsilon\}$ marks the boundary of escape. By Doob's maximal inequality, the probability of escaping the ϵ-neighborhood is then at most $\epsilon^{-\alpha}\,\mathbb{E}[(Y_0^{(i)})^\alpha]$.*

Proof Sketch. By Itô's lemma, the drift of $(Y_t^{(i)})^\alpha$ equals $\alpha(Y_t^{(i)})^{\alpha-1}$ times the difference of the two sides of equation 2, so stochastic attractivity makes it non-positive before escape. The remaining stochastic integral has integrand of order $(Y_t^{(i)})^\alpha$, since the factor $(Y_t^{(i)})^{\alpha-1}$ is offset by transverse noise vanishing on A at Lipschitz rate. As the stopped process is bounded by ϵ (Lemma B.2), the integral is a genuine martingale (Karatzas & Shreve, 1991) and vanishes under conditional expectation. □

Theorem 3.2 traps decoupled subnetworks within shared invariant subspaces and drives their path measures toward synchronization. We formalize this binding as *Attractor Linkage*.

Definition 3.3 (Attractor Linkage, $\sim_A$)**.** For two subnetworks of identical architecture let $A_{ij} = \{\boldsymbol{\theta} : \boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(j)}\}$ and $Y_t^{(ij)} = \frac{1}{2}\|\boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^{(j)}\|_2^2$ its squared transverse distance, with escape time $\tau_\epsilon^{(ij)} = \inf\{s \geq t : Y_s^{(ij)} \geq \epsilon\}$. Fix an inner radius $\epsilon_{\text{in}} < \epsilon$. Then $\boldsymbol{\theta}^{(i)} \sim_A \boldsymbol{\theta}^{(j)}$ at time t if and only if some $s \leq t$ has $Y_s^{(ij)} < \epsilon_{\text{in}}$ and $Y_u^{(ij)} < \epsilon$ for all $u \in [s, t]$, so a link forms at the inner radius and breaks at the next escape from the tube. The link probability

$$\mathcal{P}_{\text{link}}^{(i,j)}(t) := \mathbb{P}\left(\tau_\epsilon^{(ij)} = \infty \mid \mathcal{F}_t\right) \tag{3}$$

measures whether the link survives. When A_{ij} is stochastically attractive of order α, Lemma B.6 gives $\mathcal{P}_{\text{link}}^{(i,j)}(t) \geq 1 - \epsilon^{-\alpha}(Y_t^{(ij)})^\alpha$. The class $[i]_t$ is the connected component of i in the graph G_t^ϵ whose edges are the links. Components partition $\mathcal{S}$, so the classes are transitive by construction.

Links are reversible, and we project them onto a continuous Reeb graph (Edelsbrunner & Harer, 2008; Carlsson, 2009). Let $\mathcal{S} = \{1, \ldots, N\}$ index the subnetworks. A *topological transition* is any discrete change in the membership of $[i]_t$. We project these classes onto the quotient space $\mathcal{R} := (\mathcal{S} \times [0, \infty))/\sim_{\mathcal{R}}$, where $(i, t) \sim_{\mathcal{R}} (j, t)$ identifies two subnetworks whenever $j \in [i]_t$.

Theorem 3.4 (Non-Equilibrium Condensation and Fragmentation)**.** *The Reeb graph $\mathcal{R}$ realizes these topological transitions via two mechanisms:*

- ***Condensation*** *($\cup$): When $Y_t^{(ij)}$ falls below ϵ_{in}, the edge forms and Lemma B.6 bounds the chance that it ever breaks by $(\epsilon_{\text{in}}/\epsilon)^\alpha$. If the classes were disjoint, the quotient map binds them ($[i]_t = [i]_{t-\delta} \cup [j]_{t-\delta}$) into a single merge node.*
- ***Fragmentation*** *($\emptyset$): If a variance deviation breaches the basin, $\tau_\epsilon^{(ij)} \leq t$ triggers, the edge is severed, and the class splits ($[i]_t \cap [j]_t = \emptyset$) into decoupled branches.*

Proof Sketch. Doob's bound at the inner radius gives the link probability, the quotient projection $\pi_{\mathcal{R}}$ maps merging branches into a vertex of in-degree > 1, and a large deviation (Freidlin & Wentzell, 2012) that breaches $\mathcal{N}(A_{ij}, \epsilon)$ deletes the edge and produces a vertex of out-degree > 1 (Edelsbrunner & Harer, 2008). □

4 Percolative Collapse and Discrete Scale Invariance

As merges accumulate during the transient, we analyze the structure of this process as training progresses. We show that it admits a regular structure, namely a percolation process, in which merges can be tracked by peaks in relative variance.

4.1 Architectural Symmetry and Graph Renormalization

The equivalence classes defining the Reeb graph's vertices correspond to topological attractors within the loss landscape, since permutation invariances among functionally equivalent neurons partition the parameter space into lower-dimensional geometric subspaces (Chen et al., 2023).

Definition 4.1 (Approximate Q-Symmetry)**.** Let $Q \in \mathbb{R}^{d\times d}$ be orthogonal, $QQ^\top = I_d$. A loss functional $\mathcal{L}(\boldsymbol{\theta})$ exhibits approximate Q-symmetry around $A = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid Q\boldsymbol{\theta} = \boldsymbol{\theta}\}$ if, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|\mathcal{L}(Q\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| < \epsilon \cdot d(\boldsymbol{\theta}, A) \quad \forall \boldsymbol{\theta} \text{ s.t. } d(\boldsymbol{\theta}, A) < \delta, \tag{4}$$

where $d(\boldsymbol{\theta}, A) = \inf_{\boldsymbol{a}\in A} \|\boldsymbol{\theta} - \boldsymbol{a}\|_2$.

Proposition 2.2 needs the per-sample version, which permutation symmetries satisfy exactly.

Theorem 4.2 (Affine Trapping and Transverse Diffusion)**.** *If $\mathcal{L}$ satisfies approximate Q-symmetry around A, deterministic gradient drift is tangent to A on A. The affine geometry of A decouples the transverse stochastic distance $Y_t = d(\boldsymbol{\theta}_t, A)^2$ from the tangential motion without any curvature drift.*

Proof Sketch. On A, along a normal vector $\boldsymbol{n} \perp A$, approximate symmetry forces $\nabla_{\boldsymbol{n}}\mathcal{L}(\boldsymbol{\theta}) = 0$, so the drift keeps A invariant. Because A is affine, all principal curvatures vanish ($H_A = 0$), so the Itô correction of Y_t carries no geometric term and the transverse generator takes the flat form of Definition 2.3. Details are in Appendix C. □

Stochastic gradient noise breaks and reforms edges on the transient Reeb graph, and the ensemble tube occupancy supplies the monotone control parameter.

Assumption 4.3 (Condensation Phase)**.** On the training interval under study, the expected number of connected components $\mathbb{E}[M(\tau)]$ of G_τ^ϵ is nonincreasing in τ.

Proposition 4.4 (Reeb Graph Renormalization to Percolation Graph)**.** *Let $M(\tau)$ be the number of connected components of G_τ^ϵ and define the merge fraction $p(\tau) := 1 - \mathbb{E}[M(\tau)]/N \in [0, 1 - 1/N]$. Under Assumption 4.3, p is nondecreasing. Indexing the realization graphs $G_\tau := G_\tau^\epsilon$ by p through $\tau_p := \inf\{\tau : p(\tau) \geq p\}$ turns the transient Reeb graph $\mathcal{R}_t$ into a percolation process $G_p := G_{\tau_p}$ with control parameter $p \in [0, 1 - 1/N]$.*

Proof Sketch. $\mathbb{E}[M]$ is nonincreasing by assumption, while individual realizations still fragment, and those events feed the relative variance below. Links form under strict attractivity (Proposition C.10) and persist with probability at least $1 - (\epsilon_{\text{in}}/\epsilon)^\alpha$ (Lemma B.6), and the generalized inverse τ_p handles plateaus of p (Appendix C). □

We track the connectivity of G_p via the structural order parameter, the fractional size of the largest connected component:

$$\mathcal{O}(p) = \frac{|C_{\max}(p)|}{N}. \tag{5}$$

Unlike Erdős-Rényi percolation, where the largest component grows by absorbing components of vanishing relative size (Stauffer & Aharony, 1994), symmetry-induced percolation can merge components in blocks. A condensation event has *multiplicity* n if it joins n components of equal size i into one of size ni. Pairwise coincidence sets A_{ij} have the smallest codimension among the S_n-fixed subspaces, so generic condensation is pairwise and $n = 2$ (Proposition C.18). We keep n as a parameter.

Proposition 4.5 (Non-Erdős-Rényi Discontinuity)**.** *A condensation event of multiplicity n on components of size i produces a jump $\Delta\mathcal{O}(p) = (n-1)i/N$, a genuine discontinuity as $N \to \infty$ precisely when the merging components are already macroscopic ($i = \Theta(N)$), and a vanishing one when $i = O(1)$ (Corollary C.14). In $G_{ER}(N, p)$ every jump of $\mathcal{O}_{ER}$ is $o(1)$ with high probability, so the symmetry mechanism departs from Erdős-Rényi percolation only in the macroscopic regime.*

Proof Sketch. Joining n components of size i raises $|C_{\max}|$ to ni. In Erdős-Rényi graphs the largest component only ever absorbs components of size $O(N^{2/3})$ inside the critical window and $O(\log N)$ outside it (Appendix C). Bounded-choice Achlioptas processes are continuous in this limit too (Riordan & Warnke, 2011). □

Because stochastic noise shifts the density at which these discrete merges occur, single-trajectory observations obscure the critical thresholds. We isolate structural shocks via the relative variance of the order parameter across an ensemble of training trajectories:

$$R_v(p) = \frac{\mathbb{E}\left[(\mathcal{O}(p) - \mathbb{E}[\mathcal{O}(p)])^2\right]}{\mathbb{E}[\mathcal{O}(p)]^2}. \tag{6}$$

Theorem 4.6 (Bounded Peak of Relative Variance)**.** *Assume that near a microtransition the ensemble law of $\mathcal{O}$ is a two-point mixture of $\mathcal{O}_1$ and $\mathcal{O}_1 + \Delta\mathcal{O}$ with weight w on the merged state, as a single block merge produces. Then*

$$R_v(w) = \frac{w(1-w)\,\Delta\mathcal{O}^2}{(\mathcal{O}_1 + w\,\Delta\mathcal{O})^2} \tag{7}$$

vanishes at $w = 0$ and at $w = 1$ and peaks at $w^* = \mathcal{O}_1/(2\mathcal{O}_1 + \Delta\mathcal{O})$. For a block merge $\Delta\mathcal{O} = (n-1)\mathcal{O}_1$, so the peak occurs at $w^* = 1/(n+1)$ with height $(n-1)^2/(4n)$, independent of i and N. Each microtransition therefore leaves a sharp, bounded peak in $R_v(p)$.

Proof Sketch. The mixture has mean $\mathcal{O}_1 + w\,\Delta\mathcal{O}$ and variance $w(1-w)\,\Delta\mathcal{O}^2$, which gives the ratio above. Setting the derivative of its logarithm in w to zero yields w^* (Chen et al., 2014). For a block merge, $\Delta\mathcal{O} = (n-1)\mathcal{O}_1$, so $\mathcal{O}_1$ cancels and the peak height $(n-1)^2/(4n)$ depends on neither i nor N. The first merge out of isolated vertices has $\mathcal{O}_1 = 1/N$ and gives the same height (Appendix C). □

The spectral observable of Section 6 records the peak times, not this height (Proposition F.1).

Block merges break the continuous group of dilations into a discrete subgroup. An observable exhibits Discrete Scale Invariance (DSI) if it is self-similar under a discrete set of magnification factors λ^k rather than under every dilation (Sornette, 1998). Let T_i denote the time at which the mean component size $N/M(t)$ of a realization first reaches i, and call its law the hitting law of size i. In the balanced cascade below every component at a level has the same size, so T_i is also the time at which the largest component reaches i. Merged subnetworks coincide on A_{ij}, so a component has a single representative and its coincidence set with another component is again an affine subspace of the form of Theorem 4.2. We assume that the transverse process between two components depends on them only through their sizes, so that the merges form a Marcus–Lushnikov coalescence process (Marcus, 1968; Lushnikov, 1978; Aldous, 1999) with a kernel.

Assumption 4.7 (Homogeneous Coalescence Kernel)**.** During the condensation phase, two components of sizes a and b link at rate $K(a,b)/N$, where K is symmetric, depends on the components only through their sizes, is homogeneous of degree $\gamma_K < 1$, that is, $K(ca, cb) = c^{\gamma_K} K(a,b)$, and admits a dynamical scaling solution of the Smoluchowski equation. In the coalescence idealization, links formed during the phase are not broken.

Dynamical scaling is proved for constant and additive kernels and conjectured in general (Aldous, 1999).

Lemma 4.8 (Self-Similar Hitting Laws)**.** *Under Assumption 4.7, in the mean-field limit the typical component size grows as $s(t) \propto t^{1/(1-\gamma_K)}$, so $T_{ni}/T_i \to \lambda_t := n^{1-\gamma_K}$. If in addition the merges are balanced, so that components of equal size merge in groups of n (Definition C.12), then $T_{n^{k+1}} \overset{d}{=} \lambda_t T_{n^k}$ up to a relative correction of order $\lambda_t^{-k} + (n^k/N)^{1/2}$, and the merge fraction satisfies $1 - p_{n^k} = n^{-k}$. For the constant kernel $T_{n^k} = 2(n^k - 1)$, so consecutive ratios are $(n^{k+1}-1)/(n^k-1)$ and tend to n (Appendix C).*

The measured ratio gives $\gamma_K = 1 - \log_n \lambda_t$, into which a time-varying link rate also enters (Appendix E.5).

Theorem 4.9 (Coalescence with Symmetry Yields DSI)**.** *Under Assumption 4.7 with a time-homogeneous kernel and balanced merges, the mean transition time, every quantile, and the location q_i of the R_v peak, the w^*-quantile of T_i by Theorem 4.6, all satisfy $q_{ni}/q_i = \lambda_t = n^{1-\gamma_K}$ up to the corrections of Lemma 4.8, asymptotically in the level k. The peak times are invariant in law under $t \mapsto \lambda_t t$, which is DSI in training time (Sornette, 1998), and the peak merge fractions satisfy $1 - q_{ni}^{(p)} = (1 - q_i^{(p)})/n$.*

Proof Sketch. Under dynamical scaling, the typical component size grows as $t^{1/(1-\gamma_K)}$, so the time to reach size i scales as $i^{1-\gamma_K}$ and $T_{ni}/T_i \to n^{1-\gamma_K}$ (Lemma 4.8). Rescaling a law by λ_t rescales its mean and every quantile by λ_t. The R_v peak is the w^*-quantile of T_i, so its location inherits the same factor. Balanced merges divide the number of components by n at each level, which gives the relation for the merge fraction (Appendix C). □

The theorem describes the balanced, time-homogeneous coalescence idealization, and Section 6 tests trained networks for the geometric spacing and peak structure it predicts. A weight-decay-paced clock keeps this level structure but spaces the levels uniformly in t (Appendix E.5).

5 Conditional Extension to Adam and AdamW

As most large-scale training uses adaptive optimizers, we extend the trapping and cascade arguments of Sections 3 and 4 to Adam (Kingma & Ba, 2015) and AdamW. We show that they carry over under a trapping condition, which we measure on a trained Transformer in Section 6, with full derivations in Appendix D. Throughout, ϵ_{opt} denotes Adam's numerical constant and ϵ the tube radius.

Adam and AdamW maintain a joint state $(\boldsymbol{\theta}_t, m_t, v_t)$, and the elementwise squaring in v_t narrows the admissible symmetry group to coordinate permutations P_π, a class that already contains the S_n neuron-permutation symmetry, so the joint invariant set $\tilde{A} = \{(\boldsymbol{\theta}, m, v) : P_\pi \boldsymbol{\theta} = \boldsymbol{\theta}, P_\pi m = m, P_\pi v = v\}$ extends Theorem C.2 without modifying the geometry of A.

Assumption 5.1 (Heavy-Tailed Gradient Noise)**.** There exist $p \in (1, 2]$ and $\sigma > 0$ such that $\mathbb{E}[|g_{t,i}|^p \mid \mathcal{F}_{t-1}] \le \sigma^p$ almost surely for every coordinate i and every t, matching the heavy-tailed regime documented for attention-based architectures (Zhang et al., 2020).

As heavy tails permit infinite variance for $p < 2$, we analyze Adam with gradients clipped at a threshold τ, $\hat{g}_t := \text{clip}_\tau(g_t)$, as attention-based models are trained in practice. Clipping bounds every moment at a bias of at most $\sigma^p \tau^{1-p}$. We show that $\hat{v}_t$ and $\hat{m}_t$ then track the second moment and the gradient with high probability, where $\tau_2 = 1/(1-\beta_2)$ is the memory of $\hat{v}_t$ (Appendix D). Across a symmetric pair of units, the preconditioner splits as $c_t + E_t$, with c_t diagonal, role-wise and shared by both units. Theorem 5.3 then states the condition that fixes the sign of the transverse drift.

Assumption 5.2 (Local Quasi-Static Regime and Macroscopic Merges)**.** We assume that $\hat{v}_i(\boldsymbol{\theta}_t) \ge v_{\min} > 0$ for every coordinate under analysis and $t \le \tau_\epsilon^{\text{blk}}$, that $1 - \beta_2$ is small enough for $\hat{v}$ on the block to stay frozen over the trapping horizon, so that $c_t = c$ and $E_t = E$, and that the merging components satisfy $i(N) = \Theta(N)$.

Assumption 5.2 describes a local window of a run, and Remark D.30 measures the variation of $\hat{v}$ inside such windows. Under the frozen preconditioner, the transverse coordinate $\boldsymbol{x}_t$ of the block follows the Euler–Maruyama step of the reduced SDE equation 36, with drift $-\eta c \nabla_\perp \mathcal{L} - \eta\lambda \boldsymbol{x}_t \mathbb{1}_{\text{AdamW}}$, diffusion $\eta c (\Sigma_\perp^\tau)^{1/2}$ for the transverse covariance $\Sigma_\perp^\tau$ of the clipped gradient, a frozen forcing $\boldsymbol{f} = -\eta P^\perp E \nabla \mathcal{L}$, and a residual $\|\boldsymbol{\varrho}_t\| \le \rho_{\max}$ collecting the momentum lag, clipping bias and tracking errors (Proposition D.24). The white-noise reading of the filtered momentum noise in that proposition is a modeling step of the same kind as the one that turns SGD into the SGF of Section 2. The forcing is $O(\sqrt{\epsilon})$ and moves the trap off A by $O(\|\boldsymbol{f}\|)$.

Theorem 5.3 (Conditional Trapping Criterion for Adam and AdamW)**.** *Fix $\alpha \in (0, 1]$ and $0 < r_0 < \sqrt{\epsilon}$, let $\rho_\parallel \le \rho_{\max}$ be a constant with $|\boldsymbol{x}_t^\top \boldsymbol{\varrho}_t| \le \rho_\parallel \sqrt{Y_t}$ on the tracking event of Lemma D.16, and suppose that on $\{r_0^2 \le Y < \epsilon\}$*

$$-2\eta\, \boldsymbol{x}^\top c\, \nabla_\perp \mathcal{L} - 2\eta\lambda Y \mathbb{1}_{\text{AdamW}} + 2\, \boldsymbol{x}^\top \boldsymbol{f} + 2\sqrt{Y}\, \rho_\parallel + \eta^2 \operatorname{tr}(c\, \Sigma_\perp^\tau c) \;\le\; 2(1-\alpha)\, \eta^2\, \frac{\boldsymbol{x}^\top c\, \Sigma_\perp^\tau\, c\, \boldsymbol{x}}{Y}. \tag{8}$$

Then, under equation 36, Y^α, stopped when Y leaves the annulus $[r_0^2, \epsilon)$, is a non-negative supermartingale, and over T steps the trajectory escapes the tube before reaching the core $\{Y < r_0^2\}$ with probability at most $\epsilon^{-\alpha}\mathbb{E}[Y_{t_0}^\alpha \mid \mathcal{F}_{t_0}] + 2\delta$, the 2δ from Lemma D.16.

Proof Sketch. Itô's lemma on Y^α under equation 36 produces the left side of equation 8 and the concavity correction on the right. The diffusion enters at second order in c on both sides, so positivity of c alone does not determine the sign, and equation 8 is the hypothesis that determines it. The forcing and the residual enter through their projections on $\boldsymbol{x}$, and Cauchy–Schwarz gives the cruder sufficient condition with $2\sqrt{Y}(\|\boldsymbol{f}\| + \rho_{\max})$ in their place. Doob's inequality gives the escape bound (Appendix D). □

Theorem 5.4 (Conditional Cascade Transfer to Adam and AdamW)**.** *Under Assumptions 5.1, 5.2, 4.3 and 4.7, and the confinement condition of Corollary D.27, on $t \le \tau_\epsilon^{\text{blk}}$ the Adam- or AdamW-trained*

trajectory admits a percolation process whose R_v peaks form the cascade of Theorem 4.9 with the same factor λ_t.

Proof Sketch. The confinement bound of Corollary D.27 supplies the link survival that Lemma B.6 provides for SGD, and the frozen preconditioner c never enters the ratio (Appendix D, measured in Remark D.30). □

Corollary D.27 certifies less than one inter-slingshot window, so the guarantee applies window by window.

6 Empirical Validation

To test whether this percolation holds beyond our assumptions, we track merges in trained networks through the order parameter $\mathcal{O}$ and its relative variance $\mathcal{R}_v$ across training seeds, which realizes the two-point mixture of Theorem 4.6. We start from toy models, where merges can be observed directly, and move to MLPs and a Transformer, where we rely on the spectral effective rank (Appendices E and F).

6.1 Stochastic Condensation and Reactive Fragmentation

In a controlled toy setting adapting Chen et al. (2023), pairwise merges ($n = 2$) under a constant kernel ($\gamma_K = 0$) give $\lambda_t = 2$ asymptotically (Corollary C.24). A constrained $N = 3$ kinematic model checks that the detection pipeline recovers imposed merge times, and unconstrained $N = 6$ label-noise SGD gives a two-peak ratio $\lambda_t \approx 2.00$ from localized $\mathcal{R}_v$ peaks, with the finite-N prediction for six units at 2.5 (Appendix E), and a task shift at $t = 3000$ confirms the predicted reversibility, producing fragmentation at the rate $\eta(\mu - \zeta^2/2)$ sets (Appendix E, Figure 3). We also compare a nested binary-tree network with a flat layer built from the same units and trained by SGD with weight decay, whose exponential clock lies outside Lemma 4.8 but keeps its level structure. The tree also produces the macroscopic block merges of Proposition 4.5 that make the R_v peaks sharp, with the first-level peak at the height $1/8$ and weight $w^* = 1/3$ of Theorem 4.6, reaching one cluster in 47% of seeds at $N = 32$, while the flat layer condenses to two clusters by single attachments (Figure 1, Appendix E.5).

6.2 Universality, Grokking, and Benchmark Generalization

Deeper networks carry the nested structure of the tree through the composition of their layers; however, moving from the toy models to general, unconstrained networks requires replacing distance-based clustering with Spectral Effective Rank as the order parameter.[1]

In a Transformer trained on modular arithmetic, delayed generalization (grokking; Power et al., 2022) coincides with a 3-peak cascade (epochs 4500, 9500, 20000; Figure 2) immediately preceding the performance spike ($\lambda_t = 2.11$, phase-randomized spectral null false positive rate (FPR) $\leq 0.1\%$), though we do not claim this cascade is the sole driver of the transition. On the same AdamW run we measured the quantities in condition equation 8 on permutation pairs of feed-forward units (Remark D.30). With dropout active, the trapping condition holds at 50 to 62% of checkpoints (62 to 66% inside the tube), the transverse drift is inward at 63 to 74%, and contraction phases reach 3.6 decades. Within the horizon that Corollary D.27 certifies, these measurements agree with the local assumptions of Section 5. Slingshot instabilities (Thilak et al., 2022) punctuate the run about once per $1.4\,\tau_2$ and restart the condensation phase (Remark D.30). In contrast, a cascade structure appears with cleaner transient dynamics and fractional λ_t across UCI tabular classification and vision benchmarks trained with SGD, with graded surrogate support (Appendix F, Figures 7 and 11).

[1] Spectral Effective Rank measures soft dimensionality collapse via the exponentiated Shannon entropy of a weight matrix's singular value spectrum, generalizing hard, exact-zero collapse to the case where weights rarely vanish exactly.

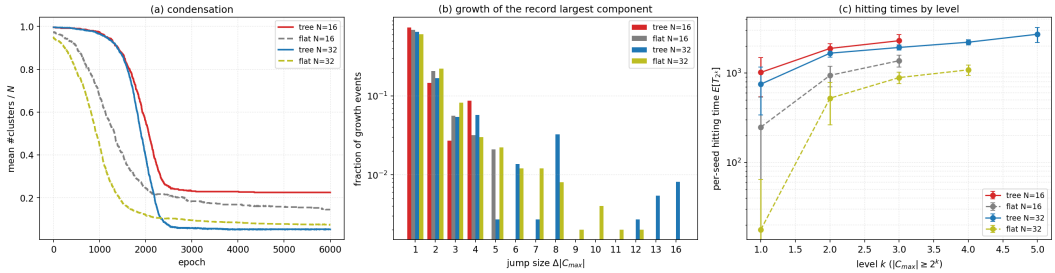


Figure 1: **Block merges under nested symmetry.** Binary tree (solid) and flat layer (dashed) at $N = 16$ and 32 over 32 seeds. **(a)** Only the tree reaches a single cluster. **(b)** Jumps of $+16$, merging two half-trees, occur only in the tree. **(c)** Hitting times per level are evenly spaced under weight decay.

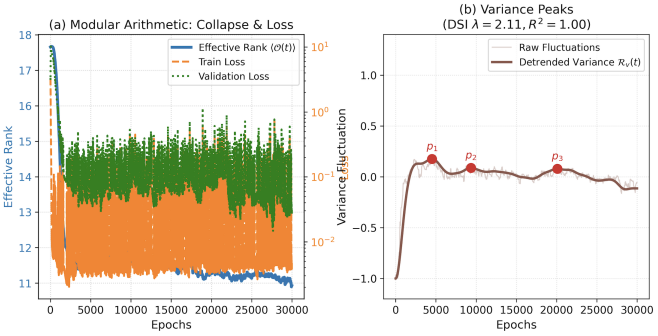


Figure 2: **Grokking (modular arithmetic).** Delayed generalization co-occurs with dimensionality collapse and a log-linear cascade of variance peaks (epochs 4500, 9500, 20000; $\lambda_t = 2.11$). Other datasets show similar cascades (Appendix F).

7 Related Work

Stochastic Dynamics and Implicit Bias: Traditionally SGD is modeled as driving networks toward low-rank manifolds (Blanc et al., 2020; Nacson et al., 2022; Chen et al., 2023) through anomalous diffusion and glassy dynamics (Baity-Jesi et al., 2018; Geiger et al., 2019; Kunin et al., 2024; Chen et al., 2020). We instead track the transient path to these manifolds through network percolation on a continuous Reeb graph.

Graph and Percolation Perspectives on Training: Recent work applies percolation tools to training, studying connectivity under dropout (Devlin & Sanders, 2025) or synaptic invariants that forecast performance (Li et al., 2022a). Our percolation graph is instead constructed from a renormalized Reeb graph (Theorem 4.4) that tracks collapse into shared invariant sets (Definition 3.3), with edges driven by architectural symmetry rather than random deletion or correlation.

Neuron and Weight Condensation: Condensation serves as a complementary lens on stochastic collapse focusing on initialization scale rather than symmetry (Luo et al., 2021; Zhou et al., 2022; 2023). Our framework gives an explicit mechanism for the transient dynamics.

8 Limitations, Conclusion and Future Work

We formalize SGD's transient dynamics as a symmetry-induced percolation process and show under assumptions that it exhibits Discrete Scale Invariance most sharply when the parameters carry the nested symmetry that deep networks inherit from composition. Future work could test whether the discontinuities persist at practical widths (Corollary C.14) and larger moduli, and relate the cascades to representation-level measures such as neural collapse. Interesting directions include an extension towards curved invariant manifolds, the usage of $\mathcal{R}_v$ peaks to schedule learning rates and uncovering deeper connections to empirical scaling laws (Remark C.25). Extending the Adam and AdamW measurements to large-scale training is left to future work.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text, mild literature review for the related work and starter code for experimentation and plot generation. We manually confirmed all citations, correctness of code and text edits. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code for reproducing the experiments is provided with the supplementary material. Full details are available in Appendices E and F.

REFERENCES

Dimitris Achlioptas, Raissa M D'Souza, and Joel Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, 2009.

David J. Aldous. Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists. *Bernoulli*, 5(1):3–48, 1999.

Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, and Gérard Ben Arous. Comparing dynamics: Deep neural networks versus glassy systems. *Journal of Statistical Mechanics*, 2018: 033301, 2018.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on Learning Theory*, pp. 483–513. PMLR, 2020.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts SGD dynamics towards simpler subnetworks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Guozhang Chen, Cheng Qu, and Pulin Gong. Anomalous diffusion dynamics of learning in deep neural networks. *arXiv preprint arXiv:2009.10588*, 2020.

Wei Chen, Malte Schröder, Raissa M D'Souza, Didier Sornette, and Jan Nagler. Microtransition cascades to percolation. *Physical review letters*, 112(15):155701, 2014.

Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9722–9732. PMLR, 2021.

Finley Devlin and Jaron Sanders. Dropout neural network training viewed from a percolation perspective. *arXiv preprint arXiv:2512.13853*, 2025.

Herbert Edelsbrunner and John Harer. Topological persistence and simplification. *Discrete & Computational Geometry*, 39(1–3):427–463, 2008.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Mark I Freidlin and Alexander D Wentzell. *Random perturbations of dynamical systems*. Springer Science & Business Media, 2012.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, and Marco Baity-Jesi. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100:012115, 2019.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pp. 2315–2357. PMLR, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2024.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2021. URL https://openreview.net/forum?id=3a0Cbtb1Hs.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, 2nd edition, 1991.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Daniel Kunin, Javier Sagastuy-Breña, Lauren Gillespie, Eshed Margalit, and Hidenori Tanaka. The limiting dynamics of SGD: Modified loss, phase-space oscillations, and anomalous diffusion. *Neural Computation*, 36:151–197, 2024.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017.

Yang Li, Kai Ma, Han Lu, Xin Chen, and Renjie Song. Neural capacitance: A new perspective of neural network selection via edge dynamics. *arXiv preprint arXiv:2201.04194*, 2022a.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss? -a mathematical framework. In *International Conference on Learning Representations*, 2022b.

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

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

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer ReLU neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.

A. A. Lushnikov. Coagulation in finite systems. *Journal of Colloid and Interface Science*, 65(2): 276–285, 1978.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Allan H. Marcus. Stochastic coalescence. *Technometrics*, 10(1):133–143, 1968.

Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Logarithmic landscape and power-law escape rate of SGD. In *International Conference on Machine Learning*, pp. 15959–15978. PMLR, 2022.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of data in deep learning. *Frontiers in Applied Mathematics and Statistics*, 6:554449, 2020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2010.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 1999.

Oliver Riordan and Lutz Warnke. Explosive percolation is continuous. *Science*, 333:322–324, 2011.

Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer-Verlag, 1996.

Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York, 1994.

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

Harshay Shah, Kaustubh Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585, 2020.

Didier Sornette. Discrete scale invariance and complex dimensions. *Physics Reports*, 297(5):239–270, 1998.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Dietrich Stauffer and Ammon Aharony. *Introduction to Percolation Theory*. Taylor & Francis, 1994.

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

P. G. J. van Dongen and M. H. Ernst. Dynamic scaling in the kinetics of clustering. *Physical Review Letters*, 54(13):1396–1399, 1985.

Marian von Smoluchowski. Drei vorträge über diffusion, brownsche molekularbewegung und koagulation von kolloidteilchen. *Physikalische Zeitschrift*, 17:557–585, 1916.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1):1–19, 2019. doi: 10.1080/00031305.2019.1583913.

Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, and Shun-ichi Amari. Dynamics of learning near singularities in layered networks. *Neural Computation*, 20(3):813–843, 2008.

Susan Wei, Daniel Murfet, Ming Gong, Huan Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10473–10486, 2022.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding the condensation of neural networks at initial training. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hanxu Zhou, Qixuan Zhou, Zhenyuan Jin, Tao Luo, Yaoyu Zhang, and Zhi-Qin John Xu. Phase diagram of initial condensation for two-layer neural networks. *arXiv preprint arXiv:2303.06561*, 2023.

Liu Ziyin, Botao Li, James B Simon, and Masahito Ueda. SGD with a constant large learning rate can converge to local maxima. In *International Conference on Learning Representations*, 2022.

CONTENTS

1 Introduction **1**

2 Stochastic Collapse in SGD **2**

3 Topological Dynamics of SGF via Attractor Equivalence **3**

4 Percolative Collapse and Discrete Scale Invariance **4**

4.1 Architectural Symmetry and Graph Renormalization 4

5 Conditional Extension to Adam and AdamW **7**

6 Empirical Validation **8**

6.1 Stochastic Condensation and Reactive Fragmentation 8

6.2 Universality, Grokking, and Benchmark Generalization 8

7 Related Work **9**

8 Limitations, Conclusion and Future Work **9**

A SDE Formulations and Invariant Sets of Learning Dynamics **16**

A.1 Deriving Stochastic Gradient Flow from SGD 16

A.2 Invariant Sets of Continuous Processes 17

B Mechanics of Topological Transitions **17**

B.1 Formalizing the Transverse Process 18

B.2 Attractor Linkage 19

B.3 Constructing the Reeb Topology 20

C Percolative Collapse and Discrete Scale Invariance **21**

C.1 Approximate Q-Symmetry, Affine Trapping, and Transverse Diffusion 21

C.2 Constructing the Percolation 22

C.3 Order Parameter and Microtransitions 24

C.4 Pairwise Condensation and Discrete Scale Invariance 25

D Conditional Extension to Adam and AdamW **28**

D.1 State Lift, Symmetry Restriction, and the Gradient Noise Model 28

D.2 Clipping, Filters, and the Step Bound 29

D.3 Temporal Tracking 31

D.4 Cross-Sectional Homogeneity 32

D.5 The Reduced Transverse Dynamics 33

E Explicit Theoretical Verification and Toy Models **37**

E.1 Generalization of the Stochastic Collapse Threshold 37
E.2 Derivation of the Scaling of Peak Times 39
E.3 Idealized Hierarchical Coalescence 39
E.4 Empirical Simulation ($N = 6$) with Task Shift and Reverse Transition 40
E.5 Block Merges Require Nested Symmetry 40

F Extended Empirical Validation Across Diverse Datasets and Architectures **42**
F.1 Methodological Extensions for Empirical Regimes 42
F.2 UCI Tabular Datasets 44
F.3 Geometric Manifolds, Vision, and Algorithmic Grokking 47

A SDE Formulations and Invariant Sets of Learning Dynamics

To analyze the transient dynamics of stochastic collapse, we formalize the relationship between discrete Stochastic Gradient Descent (SGD) and continuous Stochastic Gradient Flow (SGF). We then formally define invariant sets for these continuous processes and prove that affine invariant sets in SGD strictly translate to SGF. We adapt this section from Chen et al. (2023).

A.1 Deriving Stochastic Gradient Flow from SGD

Recall the standard SGD update rule. At iteration t, optimizing a network parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ with learning rate η over a mini-batch $\mathcal{B}_t$ of size B drawn from a dataset of size $N_{\mathcal{D}}$ yields:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{B} \sum_{i \in \mathcal{B}_t} \nabla \ell(\boldsymbol{\theta}_t; x_i, y_i). \tag{9}$$

We factorize the right-hand side to explicitly isolate the exact full-batch population gradient $\nabla \mathcal{L}(\boldsymbol{\theta}_t)$ from the zero-mean mini-batch deviation:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t) - \sqrt{\eta} \left(\sqrt{\eta} \frac{1}{B} \sum_{i \in \mathcal{B}_t} \Big(\nabla \ell(\boldsymbol{\theta}_t; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta}_t) \Big) \right).$$

Let $\xi_{\mathcal{B}}(\boldsymbol{\theta})$ represent the gradient noise at position $\boldsymbol{\theta}$. Assuming we sample mini-batches with replacement, we disentangle this batch noise into a sum of independent, identically distributed per-sample gradient deviations: $\xi_{\mathcal{B}}(\boldsymbol{\theta}) = \frac{\sqrt{\eta}}{B} \sum_{i \in \mathcal{B}_t} \xi_i$, where $\xi_i = \nabla \ell(\boldsymbol{\theta}; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta})$. By definition, the expectation $\mathbb{E}[\xi_i(\boldsymbol{\theta})] = 0$ for all $\boldsymbol{\theta}$.

Assuming these gradient noises accumulate as Gaussian random variables, we approximate the discrete step as:

$$\Delta \boldsymbol{\theta}_t = -\eta \nabla \mathcal{L}(\boldsymbol{\theta}_t) + \sqrt{2D(\boldsymbol{\theta}_t)} \Delta W,$$

where $\Delta W \sim \mathcal{N}(0, \eta I)$ and the diffusion matrix is defined as $D(\boldsymbol{\theta}) = \frac{1}{2}\text{Var}(\xi_{\mathcal{B}}) = \frac{\eta}{2B}\text{Var}(\xi_i)$. This discrete step functions as the Euler-Maruyama discretization of the following continuous-time Itô Stochastic Differential Equation (SDE), which we formally denote as Stochastic Gradient Flow (SGF):

$$d\boldsymbol{\theta}_t = -\nabla \mathcal{L}(\boldsymbol{\theta}_t) dt + \sqrt{2D(\boldsymbol{\theta}_t)} dW_t, \quad \boldsymbol{\theta}_0 = \boldsymbol{\theta}(0). \tag{10}$$

We express the position-dependent diffusion matrix exactly as:

$$D(\boldsymbol{\theta}_t) = \frac{\eta}{2B} \left(\frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N_{\mathcal{D}}} \Big(\nabla \ell(\boldsymbol{\theta}_t; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta}_t) \Big) \Big(\nabla \ell(\boldsymbol{\theta}_t; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta}_t) \Big)^\top \right).$$

This formulation explicitly decomposes the diffusion matrix into a product of a parameter-independent magnitude scalar D_m and a parameter-dependent shape matrix $D_s(\boldsymbol{\theta})$ such that $D(\boldsymbol{\theta}_t) = D_m D_s(\boldsymbol{\theta}_t)$, where:

$$D_m = \frac{\eta}{2B},$$

$$D_s(\boldsymbol{\theta}) = \frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N_{\mathcal{D}}} \Big(\nabla \ell(\boldsymbol{\theta}_t; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta}_t) \Big) \Big(\nabla \ell(\boldsymbol{\theta}_t; x_i, y_i) - \nabla \mathcal{L}(\boldsymbol{\theta}_t) \Big)^\top.$$

Optimization hyperparameters, specifically learning rate and batch size, strictly control the scalar magnitude D_m. Conversely, the network architecture, training dataset, and loss topology completely dictate the geometric shape $D_s(\boldsymbol{\theta})$.

The matrix $2D$ factors as $2D(\boldsymbol{\theta}) = G(\boldsymbol{\theta})G(\boldsymbol{\theta})^\top$ with the $d \times N_{\mathcal{D}}$ noise factor

$$G(\boldsymbol{\theta}) := \sqrt{\frac{\eta}{BN_{\mathcal{D}}}} \Big[\nabla \bar{\ell}(\boldsymbol{\theta}; x_1, y_1), \ \ldots, \ \nabla \bar{\ell}(\boldsymbol{\theta}; x_{N_{\mathcal{D}}}, y_{N_{\mathcal{D}}}) \Big], \qquad \bar{\ell}(\boldsymbol{\theta}; x_i, y_i) := \ell(\boldsymbol{\theta}; x_i, y_i) - \mathcal{L}(\boldsymbol{\theta}),$$

and we drive the SGF by an $N_{\mathcal{D}}$-dimensional Brownian motion through G, $d\boldsymbol{\theta}_t = -\nabla \mathcal{L}(\boldsymbol{\theta}_t) dt + G(\boldsymbol{\theta}_t) dW_t$. The law of the solution depends only on $GG^\top = 2D$, so the choice of factor does not affect any statement that follows. The factor itself is the reason the coefficients are Lipschitz.

A.2 Invariant Sets of Continuous Processes

To mathematically track how trajectories collapse, we formalize the topological regions that permanently trap optimization dynamics.

Definition A.1 (Invariant Sets of Continuous Processes)**.** For a given stochastic process $\{\boldsymbol{\theta}_t \in \mathbb{R}^d : t \geq 0\}$, a Borel-measurable set $A \subset \mathbb{R}^d$ acts as an invariant set if, for every initial point $\boldsymbol{\theta}_0$ strictly inside A, the probability that the process remains in A for all subsequent time $t \geq 0$ equals 1. Formally:

$$P\big[\boldsymbol{\theta}_t \in A \text{ for all } t \geq 0 \mid \boldsymbol{\theta}_0 \in A\big] = 1. \tag{11}$$

Proposition A.2 (Affine Invariance Transfer)**.** *Consider a discrete SGD process given by Equation equation 9, where the individual sample gradients $\nabla\ell(\cdot; x_i, y_i) : \mathbb{R}^d \to \mathbb{R}^d$ remain Lipschitz continuous and bounded for any $i \in [N_{\mathcal{D}}]$. Let an affine subset $A \subseteq \mathbb{R}^d$ be invariant for every sample gradient step, that is, $P\nabla\ell(\boldsymbol{\theta}; x_i, y_i) = \nabla\ell(\boldsymbol{\theta}; x_i, y_i)$ for all $i \in [N_{\mathcal{D}}]$ and all $\boldsymbol{\theta} \in A$, where P is the orthogonal projection onto the linear subspace parallel to A. Then A is a strict invariant set for the continuous SGF process given by Equation equation 10.*

Proof. By our foundational assumption, the individual sample gradients act as L-Lipschitz continuous in $\boldsymbol{\theta}$ and remain bounded such that $\|\nabla\ell(\boldsymbol{\theta}; x_i, y_i)\| \leq L$ for some constant $L > 0$. We first demonstrate that the population gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$ inherits these exact structural properties. Bounding the difference yields:

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\| = \left\| \frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N_{\mathcal{D}}} \Big(\nabla\ell(\boldsymbol{\theta}_1; x_i, y_i) - \nabla\ell(\boldsymbol{\theta}_2; x_i, y_i) \Big) \right\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Similarly, the global norm remains bounded by $\|\nabla\mathcal{L}(\boldsymbol{\theta})\| \leq \frac{1}{N_{\mathcal{D}}} \sum_{i=1}^{N_{\mathcal{D}}} \|\nabla\ell(\boldsymbol{\theta}; x_i, y_i)\| \leq L$.

The centered sample gradient $\nabla\bar{\ell}(\cdot; x_i, y_i)$ is $2L$-Lipschitz continuous and bounded by $2L$ for any index $i \in [N_{\mathcal{D}}]$, by the triangle inequality. Hence every column of the noise factor $G(\boldsymbol{\theta})$ of Section A.1 is $2L\sqrt{\eta/(BN_{\mathcal{D}})}$-Lipschitz, and

$$\|G(\boldsymbol{\theta}_1) - G(\boldsymbol{\theta}_2)\|_F \leq 2L\sqrt{\frac{\eta}{B}}\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Drift and diffusion coefficient are therefore globally Lipschitz, which guarantees the existence of a unique strong solution to the SDE in Equation equation 10 driven by G. We work with G rather than with the symmetric root $\sqrt{2D}$ because the Powers–Størmer inequality, $\|\sqrt{2D_1} - \sqrt{2D_2}\|_F^2 \leq \|2D_1 - 2D_2\|_1$, makes the symmetric root only Hölder-$\frac{1}{2}$ in $\boldsymbol{\theta}$ wherever D degenerates, and D degenerates in the transverse directions on A itself, where every sample gradient is tangent to A. Hölder-$\frac{1}{2}$ diffusion coefficients do not guarantee pathwise uniqueness in dimension $d > 1$, and the next step requires pathwise uniqueness.

It now suffices to prove that this continuous SDE possesses a solution $\{\boldsymbol{\theta}_t : t \geq 0\}$ that satisfies the invariant probability condition from Equation equation 11. We define $P : \mathbb{R}^d \to \mathbb{R}^d$ as the orthogonal projection onto the linear subspace V parallel to A and construct an auxiliary projected SDE started in A:

$$d\tilde{\boldsymbol{\theta}}_t = -P\nabla\mathcal{L}(\tilde{\boldsymbol{\theta}}_t)dt + PG(\tilde{\boldsymbol{\theta}}_t)dW_t. \tag{12}$$

Its increments lie in V, so this projected system confines its solution $\tilde{\boldsymbol{\theta}}_t$ within A, satisfying the invariance condition, and its coefficients are Lipschitz, so its solution is unique. By hypothesis every sample gradient is tangent to A on A, hence $P\nabla\mathcal{L} = \nabla\mathcal{L}$ and $PG = G$ on A, column by column. The projected SDE and the original SDE therefore have identical coefficients along any path in A, and pathwise uniqueness identifies their solutions from any $\boldsymbol{\theta}_0 \in A$ almost surely. Hence the affine subset A acts as a strict invariant set for the continuous SGF process. □

B Mechanics of Topological Transitions

Following the subnetwork partition (Definition 3.1) established in the main text, we isolate the local dynamics. The constructions below use the transverse distance $Y^{(i)}$ of each subnetwork to a target

set A and the transverse distance $Y^{(ij)}$ of each pair to its coincidence set A_{ij}. Both are well defined for any coupling between blocks; the vanishing of the pairwise transverse noise on A_{ij} comes from per-sample symmetry, which forces $\Sigma_{ij} \to \Sigma_{ii}$ there.

To formalize how these path measures structurally bind, we project the local SDEs into a shared quotient space by rigorously defining the transverse geometry.

B.1 Formalizing the Transverse Process

Definition B.1 (Transverse Distance Process)**.** Let $A \subset \mathbb{R}^{d_i}$ define an invariant manifold of the i-th block. Define the orthogonal projection operator $\pi_A : \mathbb{R}^{d_i} \to A$. We establish a local tubular neighborhood $\mathcal{N}(A, \epsilon) = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}^{(i)} - \pi_A(\boldsymbol{\theta}^{(i)})\|_2^2 < \epsilon\}$ wherein this projection remains uniquely defined. Within this boundary, we map the local dynamics into the quotient space $\mathbb{R}^{d_i}/A$ by continuously tracking the squared transverse distance functional for the i-th subnetwork:

$$Y_t^{(i)} = \|\boldsymbol{\theta}_t^{(i)} - \pi_A(\boldsymbol{\theta}_t^{(i)})\|_2^2. \tag{13}$$

Mechanically, stochastic attractivity pulls the inward deterministic gradient to overcome the transverse diffusion. We map this directly to the infinitesimal generator $\mathscr{G}$ of the SDE. Write $\boldsymbol{x}_t = \boldsymbol{\theta}_t^{(i)} - \pi_A(\boldsymbol{\theta}_t^{(i)})$ and $\Sigma_\perp = P^\perp \Sigma_{ii} P^\perp$. Applying Itô's Lemma to $Y_t^{(i)}$ and to $\phi_\alpha(Y) = Y^\alpha$, $\alpha \in (0, 1]$, expands the generators as

$$\mathscr{G} Y_t^{(i)} = -2\, \boldsymbol{x}_t^\top \nabla_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta}_t) + \eta \operatorname{tr} \Sigma_\perp(\boldsymbol{\theta}_t), \tag{14}$$

$$\mathscr{G}(Y_t^{(i)})^\alpha = \alpha (Y_t^{(i)})^{\alpha-1} \Big[\mathscr{G} Y_t^{(i)} - 2(1-\alpha)\, \eta\, \frac{\boldsymbol{x}_t^\top \Sigma_\perp(\boldsymbol{\theta}_t) \boldsymbol{x}_t}{Y_t^{(i)}} \Big], \tag{15}$$

the second term in the bracket being $\frac{1}{2}\phi''_\alpha(Y)\, \nabla Y^\top (\eta \Sigma_{ii}) \nabla Y$ with $\nabla Y = 2\boldsymbol{x}$. Only the transverse block of Σ_{ii} enters, since ∇Y and the Hessian $2P^\perp$ of Y both live in $A^\perp$.

Lemma B.2 (Boundedness of the Stopped Transverse Process)**.** *Suppose $\boldsymbol{\theta}_t$ solves the SDE in Equation equation 10 with (locally) Lipschitz coefficients, so that it admits a strong solution with almost surely continuous sample paths. Then $Y_t^{(i)} = \|\boldsymbol{\theta}_t^{(i)} - \pi_A(\boldsymbol{\theta}_t^{(i)})\|_2^2$ is a continuous functional of $\boldsymbol{\theta}_t^{(i)}$, and consequently the stopped process satisfies*

$$Y_{s\wedge\tau_\epsilon}^{(i)} \in [0, \epsilon], \qquad (Y_{s\wedge\tau_\epsilon}^{(i)})^\alpha \in [0, \epsilon^\alpha] \quad \textit{a.s. for all } s \geq t_0. \tag{16}$$

Proof. Continuity of $\boldsymbol{\theta}_t^{(i)}$ and of the projection π_A (an affine map with constant linear part P, by Theorem C.2) imply $Y_t^{(i)}$ is a.s. continuous in t. By definition $\tau_\epsilon = \inf\{t \geq t_0 : Y_t^{(i)} \geq \epsilon\}$; continuity of the path forces $Y_t^{(i)} < \epsilon$ for $t < \tau_\epsilon$ and $Y_{\tau_\epsilon}^{(i)} = \epsilon$ whenever $\tau_\epsilon < \infty$ — a continuous path cannot jump over the boundary. Hence $Y_{s\wedge\tau_\epsilon}^{(i)} \leq \epsilon$ for every s, and non-negativity of $Y^{(i)}$ gives the lower bound. Monotonicity of $y \mapsto y^\alpha$ transfers both bounds to the power. □

Theorem B.3 (Local Supermartingale Trapping)**.** *Given a stochastically attractive invariant set A of order α, there exists an $\epsilon > 0$ defining the basin $\mathcal{N}(A, \epsilon)$ such that the stopped transverse process $(Y_{t\wedge\tau_\epsilon}^{(i)})^\alpha$ operates as a non-negative supermartingale, where $\tau_\epsilon = \inf\{t \geq t_0 : Y_t^{(i)} \geq \epsilon\}$ marks the exact boundary of escape.*

Proof. Let A be stochastically attractive of order α. By Definition 2.3, the bracket in Equation equation 15 is non-positive on $\mathcal{N}(A, \epsilon) \setminus A$, so $\mathscr{G}(Y_t^{(i)})^\alpha \leq 0$ prior to the stopping time τ_ϵ wherever $Y_t^{(i)} > 0$. Since A is invariant (Proposition A.2), a path that reaches $Y = 0$ stays there. For $\delta > 0$ let $\sigma_\delta := \inf\{t \geq t_0 : Y_t^{(i)} \leq \delta\}$. On $[t_0, \tau_\epsilon \wedge \sigma_\delta]$ the map ϕ_α is C^2, so we argue for the process stopped at $\tau_\epsilon \wedge \sigma_\delta$ and let $\delta \downarrow 0$ at the end.

Applying Itô's formula to the process stopped at $\tau_\epsilon \wedge \sigma_\delta$ decomposes the trajectory into a Lebesgue drift integral and a stochastic integral:

$$(Y_{t\wedge\tau_\epsilon}^{(i)})^\alpha = (Y_{t_0}^{(i)})^\alpha + \int_{t_0}^{t\wedge\tau_\epsilon} \mathscr{G}(Y_s^{(i)})^\alpha\, ds + \int_{t_0}^{t\wedge\tau_\epsilon} 2\alpha (Y_s^{(i)})^{\alpha-1}\, \boldsymbol{x}_s^\top G_\perp(\boldsymbol{\theta}_s)\, dW_s, \tag{17}$$

with $G_\perp = P^\perp G_i$ the transverse part of the rows G_i of the noise factor of Appendix A that belong to $\boldsymbol{\theta}^{(i)}$. Because $\mathscr{G}(Y_s^{(i)})^\alpha \le 0$, the Lebesgue integral accumulates non-positive drift. On A every sample gradient is tangent to A, so $P^\perp \nabla\bar{\ell}(\boldsymbol{\theta}) = P^\perp\big(\nabla\bar{\ell}(\boldsymbol{\theta}) - \nabla\bar{\ell}(\pi_A\boldsymbol{\theta})\big)$ and $\|G_\perp(\boldsymbol{\theta})\| \le 2L\sqrt{\eta/B}\,\sqrt{Y}$. The transverse noise amplitude therefore vanishes on A at Lipschitz rate. The integrand of the stochastic integral is therefore bounded by $2\alpha Y^{\alpha-1}\sqrt{Y}\cdot 2L\sqrt{\eta/B}\sqrt{Y} = 4\alpha L\sqrt{\eta/B}\,Y^\alpha$, which is at most $4\alpha L\sqrt{\eta/B}\,\epsilon^\alpha$ by Lemma B.2. A stochastic integral with a bounded, progressively measurable integrand is a genuine L^2-martingale rather than only a local one (Karatzas & Shreve, 1991); taking the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ annihilates it without any further localization argument. This leaves only the non-positive accumulated drift, enforcing

$$\mathbb{E}\left[(Y^{(i)}_{(t+s)\wedge\tau_\epsilon})^\alpha \,\middle|\, \mathcal{F}_t\right] \le (Y^{(i)}_{t\wedge\tau_\epsilon})^\alpha. \tag{18}$$

Since $Y_t^{(i)}$ measures a squared distance, it is bounded below by 0. Letting $\delta \downarrow 0$, bounded convergence ($0 \le Y^\alpha \le \epsilon^\alpha$) carries the inequality over to the process stopped at τ_ϵ alone. Thus $(Y^{(i)}_{t\wedge\tau_\epsilon})^\alpha$ constitutes a non-negative supermartingale. □

Corollary B.4 (Probabilistic Non-Escape)**.** *For any stochastically attractive invariant set A of order α, the probability of a trajectory initialized at $\boldsymbol{\theta}_0$ escaping the ϵ-neighborhood is bounded by $\epsilon^{-\alpha}\,\mathbb{E}[(Y_{t_0}^{(i)})^\alpha \mid \mathcal{F}_{t_0}]$.*

Proof. By Theorem B.3, $(Y^{(i)}_{s\wedge\tau_\epsilon})^\alpha$ is a continuous, non-negative supermartingale for all $s \ge t_0$. Doob's maximal inequality for continuous supermartingales (Revuz & Yor, 1999) bounds the tail probability of its supremum:

$$\mathbb{P}\left(\sup_{s\ge t_0} Y^{(i)}_{s\wedge\tau_\epsilon} \ge \epsilon \,\middle|\, \mathcal{F}_{t_0}\right) = \mathbb{P}\left(\sup_{s\ge t_0} (Y^{(i)}_{s\wedge\tau_\epsilon})^\alpha \ge \epsilon^\alpha \,\middle|\, \mathcal{F}_{t_0}\right) \le \frac{1}{\epsilon^\alpha}\mathbb{E}\left[(Y_{t_0}^{(i)})^\alpha \mid \mathcal{F}_{t_0}\right]. \tag{19}$$

The event $\left\{\sup_{s\ge t_0} Y^{(i)}_{s\wedge\tau_\epsilon} \ge \epsilon\right\}$ identifies the escape event $\{\tau_\epsilon^{(i)} < \infty\}$ by Lemma B.2. □

B.2 Attractor Linkage

To map the structural binding of decoupled subnetworks, we define the conditions under which independent stochastic trajectories synchronize. For a pair (i, j) of subnetworks with identical architecture the shared manifold is the coincidence set $A_{ij} = \{\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(j)}\}$, an affine subspace with $Y_t^{(ij)} = \frac{1}{2}\|\boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^{(j)}\|_2^2 = d(\boldsymbol{\theta}_t, A_{ij})^2$ and escape time $\tau_\epsilon^{(ij)} = \inf\{s \ge t : Y_s^{(ij)} \ge \epsilon\}$. Every statement of Section B applies to A_{ij} verbatim.

Definition B.5 (Attractor Link Probability)**.** We define the link probability $\mathcal{P}^{(i,j)}_{\text{link}}(t)$ as the conditional probability that the pair never breaches the boundary of the local basin $\mathcal{N}(A_{ij}, \epsilon)$ from time t onward:

$$\mathcal{P}^{(i,j)}_{\text{link}}(t) := \mathbb{P}\left(\tau_\epsilon^{(ij)} = \infty \mid \mathcal{F}_t\right) = \mathbb{P}\left(\sup_{s\ge t} Y_s^{(ij)} < \epsilon \,\middle|\, \mathcal{F}_t\right). \tag{20}$$

Lemma B.6 (Maximal Link Bound)**.** *Let A_{ij} be stochastically attractive of order α. Given an instantaneous state inside the local basin ($Y_t^{(ij)} < \epsilon$), the link probability satisfies*

$$\mathcal{P}^{(i,j)}_{\text{link}}(t) \ge 1 - \frac{(Y_t^{(ij)})^\alpha}{\epsilon^\alpha}. \tag{21}$$

Proof. By Theorem B.3 applied to A_{ij}, the stopped process $(Y^{(ij)}_{s\wedge\tau_\epsilon})^\alpha$ is a continuous, non-negative supermartingale for all $s \ge t$. Doob's maximal inequality (Revuz & Yor, 1999) gives

$$\mathbb{P}\left(\sup_{s\ge t} Y^{(ij)}_{s\wedge\tau_\epsilon} \ge \epsilon \,\middle|\, \mathcal{F}_t\right) \le \frac{(Y_t^{(ij)})^\alpha}{\epsilon^\alpha}. \tag{22}$$

The event on the left is the escape event $\{\tau_\epsilon^{(ij)} < \infty\}$, and its complement is the link event, which yields the bound. □

Definition B.7 (Attractor Linkage, $\sim_A$)**.** Two subnetworks are linked at time t, $\boldsymbol{\theta}^{(i)} \sim_A \boldsymbol{\theta}^{(j)}$, if and only if some $s \leq t$ has $Y_s^{(ij)} < \epsilon_{\text{in}}$ and $Y_u^{(ij)} < \epsilon$ for all $u \in [s, t]$. A link forms when $Y^{(ij)}$ first falls below an inner radius $\epsilon_{\text{in}} < \epsilon$ and breaks at the next escape from the tube of radius ϵ. By Lemma B.6 a link survives forever with probability at least $1 - (\epsilon_{\text{in}}/\epsilon)^\alpha$. Without the inner radius the bound at formation would be $1 - \epsilon^{-\alpha}\epsilon^\alpha = 0$, since a continuous path enters the tube at $Y = \epsilon$ and a nondegenerate diffusion recrosses the level immediately.

Proposition B.8 (Linkage Classes)**.** *$\sim_A$ is reflexive and symmetric. The class $[i]_t$ of i is the connected component of i in the graph G_t^ϵ with edge set $\{\{i, j\} : \boldsymbol{\theta}^{(i)} \sim_A \boldsymbol{\theta}^{(j)}\}$. The classes partition $\mathcal{S}$ at every time, and two subnetworks in the same class are joined by a chain of links, each surviving with the probability of Lemma B.6.*

Proof. $Y_t^{(ii)} = 0 < \epsilon$ and $Y_t^{(ij)} = Y_t^{(ji)}$ give reflexivity and symmetry. Connected components of a graph partition its vertex set. They form the transitive closure of $\sim_A$, and the closure is all that the constructions below use. A single-pair relation cannot be transitive at fixed ϵ, since $Y^{(ij)}, Y^{(jk)} < \epsilon$ only gives $Y^{(ik)} < 4\epsilon$ by the triangle inequality, and components resolve this without weakening the tube. □

B.3 Constructing the Reeb Topology

Now, we map the links onto a discrete spatial graph, using $\sim_A$ as an adjacency function.

Definition B.9 (The ϵ-Fixed Topological Network)**.** At any fixed training time $t \in [0, \infty)$ and spatial boundary $\epsilon > 0$, define the instantaneous spatial network $G_t^\epsilon = (\mathcal{S}, E_t^\epsilon)$. The node set $\mathcal{S} = \{1, \ldots, N\}$ indexes the decoupled subnetworks. The relation $\sim_A$ of Definition B.7 is the adjacency function, so an undirected edge $e_{ij} \in E_t^\epsilon$ is present if and only if the pair is linked at time t.

This ϵ-fixed network captures instantaneous connectivity. To model stochastic collapse, we track how this network rewires across training time by elevating the discrete spatial networks into a continuous topological space.

Definition B.10 (Continuous Reeb Graph)**.** We define the macroscopic topological state of the neural network as the Reeb graph $\mathcal{R}$, constructed as the quotient space over the product of the subnetwork indices and time. The quotient is taken with respect to the components of a time-dependent graph rather than the level sets of a scalar function, so $\mathcal{R}$ is a Reeb-type quotient graph, and we keep the shorter name throughout:

$$\mathcal{R} := (\mathcal{S} \times [0, \infty)) \,/\, \sim_{\mathcal{R}} \tag{23}$$

where the identification $(i, t) \sim_{\mathcal{R}} (j, t)$ holds if and only if $j \in [i]_t$, the connected component of i in G_t^ϵ. We take $\sim_{\mathcal{R}}$ to be the smallest closed equivalence relation containing these identifications, so that at a split time both branches share the split vertex and $\mathcal{R}$ is Hausdorff.

Theorem B.11 (Non-Equilibrium Condensation and Fragmentation)**.** *The Reeb graph $\mathcal{R}$ maps the structural phase transitions of the SDE via two topological mechanisms:*

- ***Condensation*** *($\cup$): As a pair's transverse distance falls below ϵ_{in}, an edge forms in G_t^ϵ. If the two classes were disjoint, the quotient map fuses them into a single topological merge node on $\mathcal{R}$; otherwise the edge closes a cycle and $\mathcal{R}$ is unchanged.*
- ***Fragmentation*** *($\emptyset$): If an injected variance deviation breaches the basin $\mathcal{N}(A_{ij}, \epsilon)$, the edge in G_t^ϵ deletes. If it was a bridge of its component, the quotient map splits the node into decoupled branches on $\mathcal{R}$.*

Proof. We map the SDE dynamics to the topological level sets of $\mathcal{R}$ (Edelsbrunner & Harer, 2008).

For **condensation**, let $Y_t^{(ij)}$ drop below ϵ_{in}. The adjacency function updates to $e_{ij} = 1$ in G_t^ϵ, and Lemma B.6 bounds the probability that this edge ever deletes by

$$\mathbb{P}(\tau_\epsilon^{(ij)} < \infty \mid \mathcal{F}_t) \leq \frac{(Y_t^{(ij)})^\alpha}{\epsilon^\alpha} \leq \left(\frac{\epsilon_{\text{in}}}{\epsilon}\right)^\alpha, \tag{24}$$

which is small when the inner radius is small against the tube. The quotient projection $\pi_{\mathcal{R}} : \mathcal{S} \times [0, \infty) \to \mathcal{R}$ identifies $(i, t) \sim_{\mathcal{R}} (j, t)$; when $[i]_{t^-} \cap [j]_{t^-} = \emptyset$, this fuses the two components into a single merge vertex $v \in \mathcal{R}$ with in-degree$(v) \geq 2$.

For **fragmentation**, let a stochastic large deviation (Freidlin & Wentzell, 2012) breach the local basin $\mathcal{N}(A_{ij}, \epsilon)$. This exit realizes the stopping time:

$$\exists s \in (t - \delta, t] \text{ s.t. } Y_s^{(ij)} \geq \epsilon \implies \tau_\epsilon^{(ij)} \leq t. \tag{25}$$

The edge e_{ij} deletes from G_t^ϵ. When e_{ij} was a bridge, the component disconnects ($[i]_t \cap [j]_t = \emptyset$) and the quotient map partitions the trajectory into a critical split vertex $v \in \mathcal{R}$ with out-degree$(v) \geq 2$. □

C Percolative Collapse and Discrete Scale Invariance

We first bridge the geometric symmetries of the non-convex loss landscape with the statistical mechanics of percolation. We then show how the invariant sets generated by network symmetries restrict collapse, breaking continuous scale invariance and causing the macroscopic network to show Discrete Scale Invariance (DSI).

C.1 Approximate Q-Symmetry, Affine Trapping, and Transverse Diffusion

In deep neural networks, intrinsic architectural symmetries, specifically permutation invariance among functionally equivalent hidden neurons, naturally partition the parameter space into distinct geometric subspaces (Chen et al., 2023). Under stochastic gradient dynamics, these subspaces become stochastically attractive and trap the learning trajectory. To ground the subsequent macroscopic topology in the physical architecture of the network, we formalize this geometric property.

Definition C.1 (Approximate Q-Symmetry)**.** Let $Q \in \mathbb{R}^{d \times d}$ be an orthogonal matrix, $QQ^\top = I_d$, which covers the coordinate permutations used throughout. This operator represents a parameter transformation, such as the permutation of weights between identical hidden subnetworks. A loss functional $\mathcal{L}(\boldsymbol{\theta})$ exhibits approximate Q-symmetry around the affine subspace $A = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid Q\boldsymbol{\theta} = \boldsymbol{\theta}\}$ if, for any arbitrary tolerance $\epsilon > 0$, there exists a spatial radius $\delta > 0$ such that:

$$|\mathcal{L}(Q\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| < \epsilon \cdot d(\boldsymbol{\theta}, A) \quad \forall \boldsymbol{\theta} \text{ s.t. } d(\boldsymbol{\theta}, A) < \delta, \tag{26}$$

where $d(\boldsymbol{\theta}, A) = \inf_{\boldsymbol{a} \in A} \|\boldsymbol{\theta} - \boldsymbol{a}\|_2$ defines the minimal Euclidean distance from the parameter vector to the subspace.

Treating A as an arbitrary stochastic attractor omits the structural mechanics of the merge process. Deriving the geometry of A directly from the network's linear permutation operations guarantees that the merge geometry is affine and that transverse projections are constant.

Theorem C.2 (Affine Generation from Architectural Permutations)**.** *If the invariant set A is generated by the architectural permutation of identical sub-components (the symmetric group S_n), A is strictly a flat, affine subspace, characterized by a constant orthogonal projection operator $P^\perp$.*

Proof. The transposition or permutation of n indistinguishable neural components corresponds to swapping specific blocks of parameters within the weight vector $\boldsymbol{\theta}$. This operation is isomorphic to left-multiplication by a constant block-permutation matrix Q. The set of configurations invariant under this structural symmetry constitutes the eigenspace of Q corresponding to the eigenvalue $\lambda = 1$. The null space $\ker(I - Q)$ forms a linear subspace by definition, in particular an affine one. Consequently, the normal bundle is globally parallel, and the orthogonal projection $P^\perp$ mapping any state to its transverse distance vector is globally constant and independent of the local coordinate $\boldsymbol{\theta}$. □

Theorem C.3 (Symmetry-Induced Trapping)**.** *If the differentiable loss function $\mathcal{L}$ satisfies approximate Q-symmetry around the affine subspace A, then A constitutes a strict invariant set for the continuous deterministic gradient dynamics. (Chen et al., 2023)*

Proof. Let $\boldsymbol{\theta} \in A$, so $Q\boldsymbol{\theta} = \boldsymbol{\theta}$, and let $\boldsymbol{n}$ be a unit vector orthogonal to A. Then $Q(\boldsymbol{\theta} + h\boldsymbol{n}) = \boldsymbol{\theta} + hQ\boldsymbol{n}$ and $d(\boldsymbol{\theta} + h\boldsymbol{n}, A) = h$, so approximate Q-symmetry gives, for every $\epsilon > 0$ and $0 < h < \delta$,

$$\left| \frac{\mathcal{L}(\boldsymbol{\theta} + hQ\boldsymbol{n}) - \mathcal{L}(\boldsymbol{\theta} + h\boldsymbol{n})}{h} \right| < \epsilon. \tag{27}$$

Letting $h \to 0$ gives $|\nabla\mathcal{L}(\boldsymbol{\theta})^\top (Q - I)\boldsymbol{n}| \leq \epsilon$ for every $\epsilon > 0$, hence $\nabla\mathcal{L}(\boldsymbol{\theta})^\top (Q - I)\boldsymbol{n} = 0$. Since Q is orthogonal, $I - Q$ is normal, so its range is $\ker(I - Q)^\perp = A^\perp$ and it maps $A^\perp$ onto $A^\perp$. Hence $\nabla\mathcal{L}(\boldsymbol{\theta})$ is orthogonal to $A^\perp$, that is $P^\perp \nabla\mathcal{L}(\boldsymbol{\theta}) = 0$ and $Q\nabla\mathcal{L}(\boldsymbol{\theta}) = \nabla\mathcal{L}(\boldsymbol{\theta})$. The deterministic drift is tangent to A on A. □

The affine geometry of A (Theorem C.2) removes every curvature term from the transverse dynamics.

Theorem C.4 (Curvature-Free Transverse Diffusion)**.** *Because the permutation-invariant set A is an affine subspace, the Itô correction of the transverse distance $Y_t = d(\boldsymbol{\theta}_t, A)^2$ carries no geometric curvature term, and its generator is the flat expression of Equation equation 15, and Y_t decouples from the tangential coordinates up to the $\boldsymbol{\theta}$-dependence of $\nabla\mathcal{L}$ and $\Sigma_\perp$.*

Proof. Let the parameter update follow the stochastic differential equation $d\boldsymbol{\theta}_t = -\nabla\mathcal{L}(\boldsymbol{\theta}_t)dt + G(\boldsymbol{\theta}_t)dW_t$. Applying Itô's Lemma to the transverse distance function Y_t, the differential dY_t picks up a correction governed by the Hessian of the distance function. For a general manifold M, the Laplacian of the distance function encodes the mean curvature trace $\mathrm{Tr}(H_M)$, injecting a geometric drift of the form $\propto \frac{1}{2}D \cdot \mathrm{Tr}(H_M)dt$.

Because A is an affine subspace (Theorem C.2), all principal curvatures are identically zero ($H_A = 0$) and the Hessian of Y is the constant $2P^\perp$. The normal bundle is globally flat, and the generator of Y_t reduces to Equation equation 15, which contains only the gradient attraction $-2\boldsymbol{x}^\top\nabla\mathcal{L}$ and the transverse noise trace. The radial part of the transverse process is then a Bessel-type process driven by the transverse noise projection and the gradient attraction. A one-dimensional Fokker–Planck description of Y_t alone additionally requires the transverse coefficients to depend on $\boldsymbol{\theta}$ only through Y, a closure that Remark C.7 states. □

Remark C.5 (Generalization to Highly Curved Invariant Manifolds). If one generalizes this framework to neural architectures possessing continuous symmetries (e.g., continuous rotation invariances generating highly curved, non-linear invariant manifolds M), the non-zero mean curvature $H_M \neq 0$ introduces a geometric drift proportional to the local concavity of M. This drift couples the transverse fluctuations to the tangential motion along the manifold, modifying the local stationary distribution. Consequently, generalizing this topological approach to curved invariant manifolds requires incorporating these geometric potential terms to accurately model the edge formation rates.

C.2 Constructing the Percolation

While the deterministic gradient aligns parallel to A, the continuous injection of stochastic gradient noise drives the parameters into these lower-dimensional geometric spaces.

Definition C.6 (The Macroscopic Topological Graph)**.** Index the N initially decoupled subnetworks as nodes in a dynamic spatial graph $G_\tau = (\mathcal{S}, E_\tau^\epsilon)$. An undirected edge e_{ij} exists at time τ if and only if $\boldsymbol{\theta}^{(i)} \sim_A \boldsymbol{\theta}^{(j)}$ in the sense of Definition B.7, so that the pair has entered the inner tube and has not since left the tube around its shared invariant set A_{ij}.

To map stochastic network dynamics to percolation, we require a monotonically increasing control parameter. Continuous Brownian noise injects transient fragmentations (edges breaking) and condensations (edges forming) across the basin boundary ϵ, which breaks monotonicity for any single realization, and monotonicity is carried instead by the ensemble. Assumption 4.3 asks that the expected number of components $\mathbb{E}[M(\tau)]$ be nonincreasing on the training interval under study, which holds when link formation dominates link breaking, and Lemma B.6 bounds the breaking probability of a formed link by $(\epsilon_{\mathrm{in}}/\epsilon)^\alpha$. This interval is the condensation phase. A task shift or a learning-rate drop ends it (Appendix E), and under Adam with weight decay so does a slingshot instability (Remark D.30).

Remark C.7 (The Fokker–Planck Picture). Let $p(y, t)$ denote the density of $Y_t^{(ij)}$, so that $\mathcal{P}_e(t) = \int_0^\epsilon p(y, t)dy$ and, with $J(y, t)$ the probability current of the one-dimensional Fokker–Planck equation

(Risken, 1996) and a reflecting origin, $\frac{d}{dt}\mathcal{P}_e(t) = -J(\epsilon, t)$. If the transverse dynamics relax on a fast timescale τ_{fast} to a quasi-stationary law $p_{ss}(y; \sigma(t))$ that deforms on a slow timescale $\tau_{slow} \gg \tau_{fast}$ through a contracting scale $\sigma(t)$, then $\frac{d}{dt}\mathcal{P}_e = \dot{\sigma}(t) \int_0^\epsilon \partial_\sigma p_{ss}(y; \sigma) dy$, and this is positive exactly when the family is stochastically ordered on the tube, that is, when the distribution function $F_\sigma(y)$ rises as σ falls for every $y \leq \epsilon$. Two facts fix the status of this picture. First, zero stationary current at ϵ balances fragmentations against condensations in the mean and does not stop individual crossings, and for this reason Assumption 4.3 concerns the ensemble occupancy rather than individual realizations. Second, a falling mean and a falling variance do not by themselves order the family. In the toy model of Appendix E, Y is Gamma-distributed with shape $\mu/\zeta^2 - \frac{1}{2}$ and scale ζ^2. Lowering μ at fixed ζ lowers the shape at fixed scale, and a Gamma family is stochastically monotone in its shape, so the distribution function rises at every y and the ordering holds everywhere. Index the family instead by ζ, which the learning rate sets through $\zeta^2 \propto \eta$, at fixed μ. Raising ζ lowers the shape and raises the scale together, the mean $\mu - \zeta^2/2$ falls, and the two distribution functions cross at exactly one point, since the density ratio $y^{k_2-k_1} e^{y(1/\theta_1 - 1/\theta_2)}$ decreases and then increases. The distribution function rises below a crossing point $y_\times(\mu, \zeta)$ and falls above it. For $\mu = 1$ and ζ^2 moving from 1 to 1.5 the mean halves, $y_\times \approx 4.2$, and $F(5)$ drops from 0.99843 to 0.99805. The ordering therefore holds on the tube when $\epsilon < y_\times$ and fails beyond it, which is the regime the tube condition asks for. The quasi-stationary picture itself dissolves at the threshold $\mu \to \zeta^2/2$, where the log-coordinate $x = \ln w$ diffuses in the potential $V(x) = -(\mu - \zeta^2/2)x + \frac{1}{2}e^{2x}$, whose left wall flattens and whose mixing time diverges. The monotone occupancy is therefore an assumption, and the ensemble fraction of seeds inside the tube verifies it. The spectral geometry of deep SGD supplies the timescale separation behind the picture (Li et al., 2022b; HaoChen et al., 2021; Blanc et al., 2020), with fast drift along the large Hessian eigenvalues into a basin followed by slow noise-driven diffusion along the flat invariant manifold.

Definition C.8 (Merge Fraction, p)**.** Let $M(\tau)$ denote the number of connected components of G_τ^ϵ. We define the control parameter as the merge fraction

$$p(\tau) := 1 - \frac{\mathbb{E}[M(\tau)]}{N} \in \Big[0, 1 - \frac{1}{N}\Big], \tag{28}$$

which counts the expected number of merges per subnetwork. When the components are M blocks of equal size i, $1 - p = M/N = 1/i$, so the merge fraction records the component size directly, which the pairwise edge density $\mathbb{E}|E_\tau^\epsilon|/\binom{N}{2}$ does not, since M cliques of size i have edge density $(i-1)/(N-1)$.

Proposition C.9 (Reeb Graph Reduction to Percolation Graph)**.** *Under Assumption 4.3, $p(\tau)$ is nondecreasing in τ, and $p(\tau) = 1 - 1/N$ once every realization is a single component. Set $\tau_p := \inf\{\tau : p(\tau) \geq p\}$ and $G_p := G_{\tau_p}$. This yields a percolation process $G_p = (\mathcal{V}, \mathcal{E}_p)$ with control parameter $p \in [0, 1 - 1/N]$. The transient Reeb graph $\mathcal{R}_t$ and the process G_p record the same edges, and only the index differs.*

Proof. p is nondecreasing because $\mathbb{E}[M]$ is nonincreasing by assumption, and $M = 1$ in every realization gives $p = 1 - 1/N$. The supermartingale property alone does not make the transverse moments decay, so the progress of p toward $1 - 1/N$ rests on links forming, which Proposition C.10 supplies under strict attractivity, and on links persisting, which Lemma B.6 bounds through $(\epsilon_{\text{in}}/\epsilon)^\alpha$. The edge indicator of a realization is the same object in $\mathcal{R}_t$ and in G_τ, and only the index changes from τ to p. A nondecreasing $p(\tau)$ can be constant on an interval, so $\tau \mapsto p(\tau)$ need not be injective. The generalized inverse τ_p picks the first time the density reaches p and makes G_p well defined without asking for strict monotonicity. □

Single realizations may delete edges. The percolation process is monotone in its control parameter, not in its realizations, and the relative variance of the next subsection measures those deletions. As the transverse moments collapse, stochastic attractivity binds path measures and p advances from 0 (fully decoupled) to $1 - 1/N$ (complete collapse of the N symmetric subnetworks onto the invariant set A).

Proposition C.10 (Exponential Decay under Strict Attractivity)**.** *If A is strictly attractive with rate κ, so that $\mathscr{G}Y^\alpha \leq -\kappa Y^\alpha$ on $\mathcal{N}(A, \epsilon) \setminus A$, then for every $t \geq t_0$*

$$\mathbb{E}\big[(Y_t^{(i)})^\alpha \mathbb{1}\{\tau_\epsilon > t\} \,\big|\, \mathcal{F}_{t_0}\big] \leq e^{-\kappa(t-t_0)} (Y_{t_0}^{(i)})^\alpha.$$

Proof. Apply Itô's lemma to $e^{\kappa t}(Y_t^{(i)})^\alpha$ on $\{t < \tau_\epsilon\}$. The drift is $e^{\kappa t}(\kappa Y^\alpha + \mathscr{G}Y^\alpha) \le 0$, and the stochastic integral is a martingale by the argument of Theorem B.3, so $e^{\kappa(t\wedge\tau_\epsilon)}(Y_{t\wedge\tau_\epsilon}^{(i)})^\alpha$ is a non-negative supermartingale. Restricting the expectation to $\{\tau_\epsilon > t\}$, where $t \wedge \tau_\epsilon = t$, and using non-negativity of the discarded part gives $e^{\kappa t}\,\mathbb{E}[(Y_t^{(i)})^\alpha \mathbb{1}\{\tau_\epsilon > t\} \mid \mathcal{F}_{t_0}] \le e^{\kappa t_0}(Y_{t_0}^{(i)})^\alpha$. □

C.3 Order Parameter and Microtransitions

To detect macroscopic phase transitions driven by this percolation process, we formalize the physical observables tracking the connectivity of G_p.

Definition C.11 (Structural Order Parameter)**.** Let $C_{\max}(p)$ denote the largest connected component of vertices existing on the realization graph $G_p = G_{\tau_p}$ at density p. We define the macroscopic order parameter $\mathcal{O}(p)$ as its fractional size relative to the entire network:

$$\mathcal{O}(p) = \frac{|C_{\max}(p)|}{N}. \tag{29}$$

This formulation isolates the relative size of the largest macroscopic component, mirroring C_1/N in explosive percolation physics (Chen et al., 2014).

Definition C.12 (Merge Multiplicity)**.** A condensation event has multiplicity n if it joins n components of equal size i into one component of size ni. Proposition C.18 makes generic condensation pairwise, so $n = 2$ in the generic case. The balanced cascade, in which merged blocks merge again at equal sizes, is a further assumption on the order of the merges, since pairwise merging alone also admits the sequence $1+1 \to 2$, $1+2 \to 3$, $1+3 \to 4$. Nested symmetry permits the balanced order (Appendix E.5), and we keep n as a free parameter of the cascade.

In Erdős-Rényi random graphs, the order parameter grows by absorbing components of vanishing relative size. The binding of symmetric subnetworks forces block jumps. We formalize this divergence from Erdős-Rényi dynamics via the following theorem.

Proposition C.13 (Non-Erdős-Rényi Discontinuity)**.** *Let $G_{ER}(N,p)$ denote an Erdős-Rényi graph governed by uniform random edge selection. Every jump of its order parameter $\mathcal{O}_{ER}(p)$ at finite N is $o(1)$ with high probability, so $\mathcal{O}_{ER}$ is asymptotically continuous. In contrast, a condensation event of multiplicity n on components of size i in G_p produces a jump $\Delta\mathcal{O} = (n-1)i/N$, and whether this discontinuity survives the thermodynamic limit $N \to \infty$ depends on the size of the merging components relative to N at the moment of the merge (Corollary C.14).*

Proof. In $G_{ER}(N,p)$, edges arrive one at a time. An edge joining two components of sizes $a \le b$ raises the largest component by at most a, since if b is the largest component the new largest is $a+b$, and otherwise $a + b - |C_{\max}| \le a$. Below the critical window all components are $O(\log N)$, inside it the largest are $O(N^{2/3})$, and above it a unique giant component absorbs components of size $O(\log N)$ (Stauffer & Aharony, 1994). Every jump of $|C_{\max}|$ is therefore $o(N)$ with high probability, so every jump of $\mathcal{O}_{ER}$ is $o(1)$, which is the statement that $\mathcal{O}_{ER}$ is asymptotically continuous.

Conversely, a condensation event of multiplicity n binds n components of size i simultaneously into a single component of size ni. The order parameter experiences a step-like increment at finite N:

$$\Delta\mathcal{O}(p) = \frac{ni - i}{N} = \frac{(n-1)i}{N} > 0. \tag{30}$$

Whether $\Delta\mathcal{O}(p)$ remains bounded away from 0 as $N \to \infty$ is governed by the scaling of i with N; see Corollary C.14. □

Corollary C.14 (Regime Dependence of the Discontinuity)**.** *Let $i = i(N)$ denote the size of the merging components at a given microtransition.*

- ***Macroscopic regime:*** *If $i(N) = \Theta(N)$, i.e. $i(N)/N \to c \in (0,1]$ as $N \to \infty$, then $\Delta\mathcal{O}(p) \to (n-1)c > 0$: the jump is a genuine discontinuity that survives the thermodynamic limit.*
- ***Microscopic regime:*** *If $i(N) = O(1)$, i.e. the size of each merging component is bounded independently of N, then $\Delta\mathcal{O}(p) = O(1/N) \to 0$: the jump vanishes as $N \to \infty$, and the transition is asymptotically continuous.*

Proof. Immediate from $\Delta\mathcal{O}(p) = (n-1)i(N)/N$ (Proposition C.13) by taking $N \to \infty$ under each scaling assumption on $i(N)$. □

Remark C.15 (Relation to Achlioptas-Type Explosive Percolation and Finite-Size Scaling). Riordan & Warnke (2011) show that all Achlioptas processes in which an edge is chosen competitively among a bounded number of candidates at each step have continuous phase transitions in the thermodynamic limit, contrary to earlier numerical claims of discontinuity (Achlioptas et al., 2009). Corollary C.14 is consistent with this result, since the microscopic regime ($i = O(1)$) reproduces exactly the finite-N illusion-of-discontinuity behavior that Achlioptas-type processes are now known to exhibit. Our mechanism differs in the macroscopic regime, where merges are driven by symmetry-forced identification of already-macroscopic components. While determining whether practical architectures reside in the macroscopic $\Theta(N)$ or microscopic $O(1)$ regime remains an active theoretical frontier, this distinction ultimately governs only the true asymptotic limit ($N \to \infty$).

To detect these discontinuities across ensemble trajectories, we utilize the relative variance of the order parameter following (Chen et al., 2014).

Definition C.16 (Relative Variance and Microtransitions)**.** Define the relative variance $R_v(p)$ of the order parameter $\mathcal{O}(p)$ over the ensemble of stochastic trajectories as the variance normalized by the squared expectation:

$$R_v(p) = \frac{\mathbb{E}\left[(\mathcal{O}(p) - \mathbb{E}[\mathcal{O}(p)])^2\right]}{\mathbb{E}[\mathcal{O}(p)]^2}. \tag{31}$$

Discontinuous block merges manifest as sharp, localized peaks in $R_v(p)$. We define these peaks as *microtransitions*, which precede complete collapse at $p = 1 - 1/N$.

Theorem C.17 (Bounded Peak of Relative Variance at Microtransitions)**.** *If the ensemble law of the order parameter near a microtransition density p_i is a two-point mixture of the pre-jump value $\mathcal{O}_1$ and the post-jump value $\mathcal{O}_1 + \Delta\mathcal{O}$, with weight w on the merged state, then*

$$R_v(w) = \frac{w(1-w)\,\Delta\mathcal{O}^2}{(\mathcal{O}_1 + w\,\Delta\mathcal{O})^2}, \tag{32}$$

with $R_v(0) = R_v(1) = 0$ and a unique maximum at $w^* = \mathcal{O}_1/(2\mathcal{O}_1 + \Delta\mathcal{O})$ of height $\Delta\mathcal{O}^2/\big(4\mathcal{O}_1(\mathcal{O}_1 + \Delta\mathcal{O})\big)$. For a block merge of multiplicity n we have $\Delta\mathcal{O} = (n-1)\mathcal{O}_1$, so the peak occurs at $w^* = 1/(n+1)$ with height $(n-1)^2/(4n)$, independent of i and N. For pairwise merges the height equals $1/8$.

Proof. Near p_i, stochastic fluctuations cause different network realizations to either undergo the block jump or remain in the unmerged state. Model the ensemble distribution as a mixture of two states with values $\mathcal{O}_1$ and $\mathcal{O}_2 = \mathcal{O}_1 + \Delta\mathcal{O}$, weighted by probabilities $1 - w$ and w respectively. Then $\mathbb{E}[\mathcal{O}] = \mathcal{O}_1 + w\Delta\mathcal{O}$ and

$$\mathrm{Var}(\mathcal{O}) = w(1-w)(\Delta\mathcal{O})^2, \tag{33}$$

which gives the stated $R_v(w)$. Set $a = \mathcal{O}_1$ and $b = \mathcal{O}_1 + \Delta\mathcal{O}$. Then

$$R_v(w) = \frac{w(1-w)(b-a)^2}{(a + w(b-a))^2}, \qquad R_v'(w) = \frac{(b-a)^2\,[a - (a+b)w]}{(a + w(b-a))^3}, \tag{34}$$

which vanishes at $w^* = a/(a+b)$, where $R_v(w^*) = (b-a)^2/(4ab)$. Substituting $b = na$ gives $w^* = 1/(n+1)$ and $(n-1)^2/(4n)$. Since $\mathcal{O} \geq 1/N$, the peak is always finite: the first merge out of isolated vertices has $\mathcal{O}_1 = 1/N$ and $\Delta\mathcal{O} = (n-1)/N$, and gives the same height. The peak flags the microtransition, and its height does not. In the fixed-multiplicity model every peak of a cascade has the same height, and each peak is located at the w^*-quantile of the hitting law. □

C.4 Pairwise Condensation and Discrete Scale Invariance

In standard continuous percolation, cluster sizes follow universal power laws in $p_c - p$ and edges are drawn uniformly. Permutation symmetry restricts which merges can happen.

Proposition C.18 (Generic Pairwise Condensation)**.** *Let A be generated by the symmetric group S_n acting on n identical subnetworks of dimension d_0. The fixed subspaces of the subgroups $S_k \subset S_n$, $2 \leq k \leq n$, are nested, $A_{S_2} \supset A_{S_3} \supset \cdots \supset A_{S_n} = A$, with codimension $(k-1)d_0$. A trajectory*

reaching A_{S_k} from a generic initial condition passes through $A_{S_{k-1}}$ first, so condensation events are pairwise for generic initializations, and a simultaneous n-body merge requires the $n-1$ independent pairwise transverse distances to vanish at the same time, a codimension-$(n-1)d_0$ coincidence.

Proof. $A_{S_k} = \{\boldsymbol{\theta}^{(1)} = \cdots = \boldsymbol{\theta}^{(k)}\}$ has codimension $(k-1)d_0$, and $A_{S_k} \subset A_{S_{k-1}}$ since equality of k blocks implies equality of the first $k-1$. Each pairwise coincidence set A_{ij} is invariant and, under stochastic attractivity, trapping for its pair by Theorem B.3, so the pairwise mechanism operates independently of the higher coincidences. Reaching A_{S_k} means $Y^{(ij)} \to 0$ for $k-1$ independent pairs. Their hitting times are distinct for generic transverse dynamics, and for $n \geq 3$ the event that they coincide has codimension $(n-1)d_0 > d_0$ in the space of transverse configurations. The largest component therefore grows by pairwise merges, and the hierarchy $1 \to 2 \to 4 \to \cdots$ of Definition C.12 is the idealization in which merged pairs merge pairwise again at equal sizes. □

Traditional scale invariance models self-similarity through a continuous Lie group of dilations, where observables remain invariant or scale homogeneously under arbitrary continuous rescaling transformations $x \to \lambda x$ for any real $\lambda > 0$. The infinitesimal generators of these scaling transformations form a continuous Lie algebra, enforcing continuous translation symmetry in logarithmic space ($\ln x \to \ln x + \ln \lambda$).

However, when underlying microscopic constraints or discrete architectural symmetries govern the system, continuous translation invariance in log-space is broken down to a discrete subgroup. The system ceases to be scale-invariant for arbitrary continuous factors, preserving self-similarity only under a discrete geometric sequence of preferred magnification ratios. This gives rise to Discrete Scale Invariance (DSI) (Sornette, 1998).

Definition C.19 (Discrete Scale Invariance (DSI))**.** A system property or observable $f(x)$ exhibits Discrete Scale Invariance if it is invariant under scaling by a discrete set of preferred magnification factors rather than a continuous Lie group, satisfying scaling relations under discrete ratios λ^k (Sornette, 1998).

Definition C.20 (Hitting Laws)**.** Let $\{\omega\}$ index independent noise realizations of the SGF process, and let $T_i^{(\omega)}$ denote the time at which the mean component size $N/M(t)$ of realization ω first reaches i. We write T_i for the random variable, call its law the hitting law of size i, and write $p_i = 1 - 1/i$ for the corresponding merge fraction. In the balanced cascade of Definition C.12 every component at a level has the same size, so T_i is the time at which the largest component reaches i and $\mathcal{O}$ jumps. In asynchronous coalescence with the constant kernel the size distribution at time t is geometric with mean $N/M(t)$, so the largest of M components exceeds the mean by a factor of order $\log M$, which varies slowly and cancels from the ratio T_{ni}/T_i in the limit.

Remark C.21 (Why the Scaling Concerns the Hitting Laws). No functional of the ensemble mean growth curve $m(p) := \mathbb{E}_\omega[|C^{(\omega)}_{\max}(p)|]$ determines p_i, since inverting m at level i does not return $\mathbb{E}[T_i]$. Take $C(p) = 1 + \mathbb{1}\{T \leq p\}$ with $F_T(p) = 1/(1-p) - 1$ on $[0, \frac{1}{2}]$. Then $m(p) = 1/(1-p)$ and $m^{-1}(2) = \frac{1}{2}$, while $\mathbb{E}[T] = 1 - \log 2$. Inverting a mean and averaging an inverse are different operations, and the cascade is a statement about the second.

Definition C.22 (Symmetric Coalescence Process)**.** Read the components of G_τ as clusters with mass equal to their size. Two components of sizes a and b merge when the transverse distance of one of their constituent pairs enters the tube, and Assumption 4.7 states that this happens at rate $K(a,b)/N$ with K homogeneous of degree $\gamma_K < 1$. This is the Marcus–Lushnikov process with kernel K (Marcus, 1968; Lushnikov, 1978), whose mean-field limit is the Smoluchowski coagulation equation (von Smoluchowski, 1916; Aldous, 1999).

Proof of Lemma 4.8. For the constant kernel, the number $M(t)$ of components decreases by one at total rate $M(M-1)/(2N)$, and in the mean-field limit $N \to \infty$ with M/N fixed the law of large numbers gives $\dot{M} = -M(M-1)/(2N)$ (Aldous, 1999, Section 2). The mean component size $m = N/M$ then satisfies $\dot{m} = \frac{1}{2}(1 - 1/M) \to \frac{1}{2}$, so $m(t) = 1 + t/2$ grows linearly, the hitting time of size i is $T_i \to 2(i-1)$, and $T_{ni}/T_i \to n$. For a kernel homogeneous of degree $\gamma_K < 1$ that admits a dynamical scaling solution, which is proved for the constant and additive kernels and conjectured in general (Aldous, 1999), the Smoluchowski equation has solutions $c(x,t) = s(t)^{-2}\varphi(x/s(t))$ with $s(t) \propto t^{1/(1-\gamma_K)}$ (van Dongen & Ernst, 1985; Aldous, 1999), so the size scale reaches i at

$T_i \propto i^{1-\gamma_K}$ and $T_{ni}/T_i \to n^{1-\gamma_K} = \lambda_t$. In the hierarchical cascade, level k has $M_k = Nn^{-k}$ components of size n^k, each linking at rate $M_k K(n^k, n^k)/N = n^{k(\gamma_K - 1)}$, so the level duration is $D_k = n^{k(1-\gamma_K)}\tilde{D}_k = \lambda_t^k \tilde{D}_k$, where $\tilde{D}_k$ is the completion time of a level with unit link rate, a sum of $M_k(1-1/n)$ independent exponential waiting times whose relative fluctuation is $O(M_k^{-1/2})$. Summing the levels, $T_{n^{k+1}} = \sum_{j=0}^{k} D_j = \tilde{D}_0 + \lambda_t \sum_{j=0}^{k-1} \lambda_t^j \tilde{D}_{j+1} = \tilde{D}_0 + \lambda_t T'_{n^k}$ with T'_{n^k} equal in law to T_{n^k} as $N \to \infty$ at fixed k (at finite N the level durations differ in their numbers of summands), and $\tilde{D}_0$ is a fraction of order λ_t^{-k} of the total. The merge fraction is $p = 1 - M/N$, and at the hitting time of size i in the hierarchy $M = N/i$, so $1 - p_i = 1/i$ and $1 - p_{ni} = (1-p_i)/n$ exactly. The mean-field statements concern the typical component size, and the largest component follows the same scaling up to logarithmic factors in M_k, which the hierarchy removes. □

Lemma 4.8 gives a single factor $\lambda_t > 1$ with $T_{ni} \overset{d}{=} \lambda_t T_i$ for every i, so that the hitting law of size ni is the hitting law of size i dilated by λ_t, up to the stated corrections, and in the density variable $1 - p_{ni} = (1-p_i)/n$.

Theorem C.23 (Discrete Scale Invariance)**.** *Assume the coalescence process of Assumption 4.7 has the self-similar scaling limit $s(t) \propto t^{1/(1-\gamma_K)}$ that its dynamical scaling clause supplies, and that the merges are balanced. Then for every i and every functional Φ of a law that commutes with affine maps of the line, so the mean, every quantile, and the w^*-quantile q_i that locates the R_v peak of Theorem C.17,*

$$\frac{\Phi(T_{ni})}{\Phi(T_i)} = \lambda_t = n^{1-\gamma_K}, \tag{35}$$

up to the corrections of Lemma 4.8. The peak times lock into the geometric cascade $q_{n^k i} = \lambda_t^k q_i$ asymptotically in k (Corollary C.24 gives the finite-level ratios), so the R_v peaks of the microtransitions are equally spaced in $\log t$ with spacing $\log \lambda_t$, and the peak densities satisfy $1 - q^{(p)}_{n^k i} = n^{-k}(1 - q^{(p)}_i)$. The peaks precede the macroscopic transition and are equally spaced in $\log(1-p)$ with spacing $\log n$ (Chen et al., 2014).

Proof. Lemma 4.8 gives $T_{ni} \overset{d}{=} \lambda_t T_i$. A functional that commutes with affine maps satisfies $\Phi(\lambda_t T) = \lambda_t \Phi(T)$, and means and quantiles do, since $\mathbb{E}[\lambda_t T] = \lambda_t \mathbb{E}[T]$ and $\mathbb{P}(\lambda_t T \le \lambda_t z) = \mathbb{P}(T \le z)$. Hence $\Phi(T_{ni}) = \lambda_t \Phi(T_i)$, and induction on k gives the cascade. The density statement is $p = 1 - M/N$ with $M = N/i$ at the hitting time of size i. The R_v peak of Theorem C.17 occurs where the merged fraction $w = \mathbb{P}(T_i \le t)$ equals w^*, so q_i is a quantile of T_i and the identity covers it. Continuous translational invariance in $\log t$ is broken to the discrete subgroup generated by $\log \lambda_t$, which is DSI. □

Corollary C.24 (Geometric Peak Times)**.** *Under Assumption 4.7, consecutive R_v peak times satisfy $t_{ni}/t_i = \lambda_t = n^{1-\gamma_K}$ up to the finite-size corrections of Lemma 4.8, so $\log t_{n^k i}$ is linear in k with slope $\log \lambda_t$, and the measured ratio fixes $\gamma_K = 1 - \log_n \lambda_t$ once n is known. For the constant kernel the exact finite-size ratios are $(ni-1)/(i-1)$.*

Proof. The peak time t_i is the w^*-quantile of T_i, and Lemma 4.8 scales every quantile by λ_t. For the constant kernel, $T_{n^k} = 2(n^k - 1)$ from the proof of the lemma. □

It is essential to note that DSI generalizes traditional continuous scale invariance by capturing hierarchical, multi-scale organizational structures. While standard continuous power laws assume scale-free behavior across all intervals, DSI systems reveal preferred magnification scales governed by underlying discrete microscopic symmetries, such as pairwise permutation constraints.

Remark C.25 (Empirical Neural Scaling Laws and Stagewise Development). Empirical scaling laws in deep learning, such as power-law relationships governing test loss relative to compute budgets and parameter counts (Kaplan et al., 2020; Hoffmann et al., 2022), are conventionally modeled as smooth, continuous power functions. However, finer-grained analyses of training dynamics reveal that network development is inherently stagewise and non-monotonic, characterized by sudden behavioral shifts and localized phase transitions corresponding to changes in loss landscape degeneracy (Wei et al., 2022; Hoogland et al., 2024). Interpreting parameter collapse through symmetry-induced percolation suggests that empirical scaling laws might capture coarse-grained averages over an underlying DSI sequence of symmetry-breaking microtransitions.

D CONDITIONAL EXTENSION TO ADAM AND ADAMW

The percolation and DSI results of Appendix C are derived for vanilla SGD. We state conditions under which the trapping argument carries over to Adam (Kingma & Ba, 2015) and AdamW (Loshchilov & Hutter, 2019). The extension requires three departures from the SGD setting. The state must be lifted to include the first- and second-moment estimates (m_t, v_t). The admissible symmetry group narrows from general orthogonal Q to coordinate permutations, since elementwise squaring in v_t does not commute with sign flips or general rotations. The gradient noise model must accommodate the heavy tails documented for attention-based architectures (Zhang et al., 2020). We handle the last point by analyzing the clipped recursion that practitioners run, and separate the resulting approximation error into a deterministic cross-sectional component, controlled by equivariance and the Lipschitz geometry of Proposition A.2, and a temporal component, controlled by martingale-difference variance bounds for the two exponential filters. When the trapping inequality equation 37 and the confinement condition of Corollary D.27 hold, trapping and the cascade of Theorem C.23 carry over on an annulus around the joint invariant set up to the block escape time. Throughout, ϵ_{opt} is Adam's numerical constant and ϵ the tube radius.

Standing assumption for this appendix. The sample gradients $\nabla\ell(\cdot; x_i, y_i)$ are uniformly L-Lipschitz in $\boldsymbol{\theta}$. No bound on their magnitude is assumed, since Assumption D.4 permits none, and the boundedness hypothesis of Proposition A.2 belongs to the SGD sections and is not used here. For AdamW we assume in addition that $\|\boldsymbol{\theta}_t\|_\infty \le \Theta$ on every coordinate over the horizon, so that the decoupled decay step is bounded by $\eta\lambda\Theta$ and every step bound below holds for AdamW with $\Delta_{\max}$ replaced by $\Delta_{\max} + \eta\lambda\Theta$. Lipschitzness of the mini-batch gradient and of $\nabla\mathcal{L}$ follow by averaging, exactly as in the first display of the proof of Proposition A.2, which uses Lipschitzness alone. We fix a horizon of T steps and a confidence level $\delta \in (0, 1)$, and every high-probability statement below holds uniformly over $t_0 \le t \le T$ and over the n coordinates of the block under analysis.

D.1 STATE LIFT, SYMMETRY RESTRICTION, AND THE GRADIENT NOISE MODEL

At iteration t, Adam and AdamW maintain the joint state $\xi_t = (\boldsymbol{\theta}_t, m_t, v_t) \in \mathbb{R}^{3d}$, updated as

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^{\circ 2}, \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon_{\text{opt}}} + \lambda \boldsymbol{\theta}_{t-1} \mathbb{1}_{\text{AdamW}} \right), \end{aligned}$$

where $g_t^{\circ 2}$ denotes the elementwise square and $\hat{m}_t, \hat{v}_t$ are the bias-corrected estimates. This recursion is not Markov in $\boldsymbol{\theta}_t$ alone, so any invariant set must be stated jointly over $(\boldsymbol{\theta}, m, v)$.

Definition D.1 (Joint Invariant Set)**.** Let $P_\pi \in \mathbb{R}^{d\times d}$ be a coordinate permutation matrix. Define the joint invariant set

$$\tilde{A} := \{(\boldsymbol{\theta}, m, v) \in \mathbb{R}^{3d} \mid P_\pi \boldsymbol{\theta} = \boldsymbol{\theta},\ P_\pi m = m,\ P_\pi v = v\}.$$

Lemma D.2 (Joint Equivariance of the Adam Recursion)**.** *If the gradient map is P_π-equivariant, $g_t(P_\pi\boldsymbol{\theta}) = P_\pi g_t(\boldsymbol{\theta})$, then for every t,*

$$m_t(P_\pi\boldsymbol{\theta}) = P_\pi m_t(\boldsymbol{\theta}), \qquad v_t(P_\pi\boldsymbol{\theta}) = P_\pi v_t(\boldsymbol{\theta}).$$

The AdamW decoupled weight-decay term imposes no further restriction.

Proof. We proceed by induction on t. The claim holds trivially at $t = 0$ since $m_0 = v_0 = 0$. Assume $m_{t-1}(P_\pi\boldsymbol{\theta}) = P_\pi m_{t-1}(\boldsymbol{\theta})$ and $v_{t-1}(P_\pi\boldsymbol{\theta}) = P_\pi v_{t-1}(\boldsymbol{\theta})$. For the first moment, linearity of the recursion and P_π-equivariance of g_t give $m_t(P_\pi\boldsymbol{\theta}) = \beta_1 P_\pi m_{t-1}(\boldsymbol{\theta}) + (1-\beta_1) P_\pi g_t(\boldsymbol{\theta}) = P_\pi m_t(\boldsymbol{\theta})$. This step holds for any Q with $QQ^\top = I$. For the second moment, write $(P_\pi g)_i = g_{\pi(i)}$ for the permutation π. Then $\big((P_\pi g)^{\circ 2}\big)_i = g_{\pi(i)}^2 = \big(g^{\circ 2}\big)_{\pi(i)} = \big(P_\pi(g^{\circ 2})\big)_i$, since permuting coordinates and squaring elementwise commute exactly, because squaring acts identically on every coordinate regardless of label. Consequently $v_t(P_\pi\boldsymbol{\theta}) = \beta_2 P_\pi v_{t-1}(\boldsymbol{\theta}) + (1-\beta_2) P_\pi\big(g_t(\boldsymbol{\theta})^{\circ 2}\big) = P_\pi v_t(\boldsymbol{\theta})$. This commutation fails for a general orthogonal Q, since $(Qg)_i^2 = \big(\sum_j Q_{ij} g_j\big)^2$ mixes coordinates

before squaring. We therefore work with the coordinate-permutation group throughout, which is also the group generated by S_n used in Theorem C.2. The bias correction scales by a deterministic scalar and preserves equivariance. The AdamW term $\lambda\boldsymbol{\theta}_{t-1}$ is linear in $\boldsymbol{\theta}$ and imposes no additional restriction. Clipping acts coordinatewise by the same map on every coordinate, so it commutes with P_π as well, and the clipped recursion below inherits the lemma. □

Remark D.3. Lemma D.2 restricts the admissible symmetry class of the Approximate Q-Symmetry condition (Definition C.1) from general orthogonal or symmetric Q to coordinate permutations. Neuron-permutation invariance, the source of the S_n-symmetry used throughout Section 4, already lies in this restricted class, so Theorem C.2 and Proposition C.18 require no modification.

Having fixed the admissible symmetry class, we turn to the statistical model of the gradient noise itself.

Assumption D.4 (Heavy-Tailed Gradient Noise)**.** There exist $p \in (1, 2]$ and $\sigma > 0$ such that $\mathbb{E}[|g_{t,i}|^p \mid \mathcal{F}_{t-1}] \le \sigma^p$ almost surely for every coordinate i and every t. The conditional form is the one the clipping-bias and tracking bounds use.

Assumption D.4 is the standard model for gradient noise in attention-based architectures (Zhang et al., 2020), and permits infinite variance when $p < 2$. The following representation grounds σ^p as a computable functional of the gradient's characteristic function, making Assumption D.4 checkable in practice.

Lemma D.5 (Fractional Moment via the Characteristic Function)**.** *Let* $\phi_t(u) := \mathbb{E}[e^{iug_{t,i}}]$. *For* $p \in (1, 2)$,

$$\mathbb{E}|g_{t,i}|^p = \frac{2\,\Gamma(p+1)\sin(\pi p/2)}{\pi} \int_0^\infty \frac{1 - \operatorname{Re}\phi_t(u)}{u^{p+1}}\, du.$$

Proof. For $p \in (0, 2)$ and any $x \in \mathbb{R}$, $|x|^p = c_p^{-1} \int_0^\infty u^{-p-1}(1 - \cos(ux))\, du$ with $c_p := \pi/(2\Gamma(p+1)\sin(\pi p/2))$ (Samorodnitsky & Taqqu, 1994). Taking expectations and exchanging with the u-integral by Tonelli's theorem, valid since the integrand is non-negative, gives $\mathbb{E}|g_{t,i}|^p = c_p^{-1} \int_0^\infty u^{-p-1}(1 - \mathbb{E}[\cos(ug_{t,i})])\, du$, and $\mathbb{E}[\cos(ug_{t,i})] = \operatorname{Re}\phi_t(u)$ gives the stated form. □

Remark D.6. The restriction $p \in (1, 2)$ excludes the finite-variance case $p = 2$, where the identity of Lemma D.5 degenerates as $\sin(\pi p/2) \to 0$. There, $\mathbb{E}[g_{t,i}^2] = -\phi_t''(0)$ directly. Convergence of the integral requires $1 - \operatorname{Re}\phi_t(u) = O(u^p)$ as $u \to 0$, the standard local regularity condition replacing an ordinary second derivative when $p < 2$.

D.2 Clipping, Filters, and the Step Bound

Under Assumption D.4 with $p < 2$, g_t may have infinite variance, so neither the momentum nor the second-moment recursion admits moment estimates directly. Practice resolves this by clipping, and we analyze the clipped recursion.

Definition D.7 (Clipped Gradient)**.** For a fixed threshold $\tau > 0$ define $\operatorname{clip}_\tau(x) := \operatorname{sign}(x)\min(|x|, \tau)$ coordinatewise and $\hat{g}_t := \operatorname{clip}_\tau(g_t)$. The map clip_τ is 1-Lipschitz, fixes $[-\tau, \tau]$ pointwise, and satisfies $|x - \operatorname{clip}_\tau(x)| = (|x| - \tau)_+ \le |x|\,\mathbb{1}\{|x| > \tau\}$.

Lemma D.8 (Clipping Bias and Coincidence)**.** *Under Assumption D.4,* $\big|\mathbb{E}[g_{t,i} \mid \mathcal{F}_{t-1}] - \mathbb{E}[\hat{g}_{t,i} \mid \mathcal{F}_{t-1}]\big| \le \sigma^p \tau^{1-p}$ *for every coordinate, and on a run of* T *steps the clipped and unclipped recursions produce identical iterates on an event of probability at least* $1 - dT\sigma^p\tau^{-p}$.

Proof. The bias equals $\mathbb{E}[(g_{t,i} - \operatorname{clip}_\tau(g_{t,i})) \mid \mathcal{F}_{t-1}]$, bounded in absolute value by $\mathbb{E}[|g_{t,i}|\mathbb{1}\{|g_{t,i}| > \tau\} \mid \mathcal{F}_{t-1}]$. By Hölder's inequality with exponents $p, p/(p-1)$ and Markov's inequality applied to $|g_{t,i}|^p$,

$$\mathbb{E}\big[|g_{t,i}|\mathbb{1}\{|g_{t,i}| > \tau\}\big] \le (\mathbb{E}|g_{t,i}|^p)^{1/p}\, \mathbb{P}(|g_{t,i}| > \tau)^{(p-1)/p} \le \sigma \left(\frac{\sigma^p}{\tau^p}\right)^{(p-1)/p} = \sigma^p \tau^{1-p}.$$

Clipping is inactive when $|g_{s,i}| \le \tau$ for all $s \le T$ and all i. Markov's inequality and a union bound over the dT coordinates bound the complementary probability by $dT\sigma^p\tau^{-p}$, and on that event the two recursions receive identical inputs and produce identical iterates. □

The coincidence bound is informative for $\tau \gg (dT)^{1/p}\sigma$. The analysis concerns clipped Adam, the algorithm that is run in practice, and Lemma D.8 transfers its conclusions to unclipped Adam on the event that clipping is inactive.

Since $\hat{g}_t \in [-\tau, \tau]^d$ almost surely, $\hat{m}_t, \hat{v}_t$ have finite moments of every order, with $|\hat{m}_{t,i}| \le \tau$ and $\hat{v}_{t,i} \in [0, \tau^2]$. We write the bias-corrected filters as explicit weighted sums,

$$\hat{m}_t = \sum_{s=1}^{t} w_{t-s}^{(1,t)}\, \hat{g}_s, \quad w_k^{(1,t)} := \frac{(1-\beta_1)\beta_1^k}{1-\beta_1^t}, \qquad \hat{v}_t = \sum_{s=1}^{t} w_{t-s}^{(2,t)}\, \hat{g}_s^{\circ 2}, \quad w_k^{(2,t)} := \frac{(1-\beta_2)\beta_2^k}{1-\beta_2^t},$$

each family of weights summing to one over $0 \le k \le t-1$. The transfer-function view of these filters fixes their memory scales.

Definition D.9 (Shift Operator and z-Transform)**.** For a sequence $\{x_t\}_{t\ge 0}$, define the shift operator $(Sx)_t := x_{t-1}$, with $x_{-1} := 0$, and the one-sided z-transform (Oppenheim & Schafer, 2010)

$$X(z) := \mathcal{Z}\{x_t\} := \sum_{t=0}^{\infty} x_t z^{-t},$$

defined on the region $|z| > \rho$ for the smallest $\rho \ge 0$ for which the series converges absolutely.

Lemma D.10 (Shift Property)**.** $\mathcal{Z}\{Sx\}(z) = z^{-1}X(z)$.

Proof. $\mathcal{Z}\{Sx\}(z) = \sum_{t\ge 0} x_{t-1} z^{-t} = \sum_{t\ge 1} x_{t-1} z^{-t}$, since the $t = 0$ term vanishes because $x_{-1} = 0$, giving $z^{-1}\sum_{s\ge 0} x_s z^{-s} = z^{-1}X(z)$ after reindexing $s = t-1$. □

Let $\hat{G}(z) := \mathcal{Z}\{\hat{g}_t\}$. Since $|\hat{g}_t| \le \tau$, the series converges for all $|z| > 1$.

Lemma D.11 (Momentum and Second-Moment Filters)**.** $\hat{M}(z) = \dfrac{1-\beta_1}{1-\beta_1 z^{-1}}\hat{G}(z)$ *and* $\hat{V}(z) = \dfrac{1-\beta_2}{1-\beta_2 z^{-1}}\mathcal{Z}\{\hat{g}_t^{\circ 2}\}$*, with poles at* $z = \beta_1$ *and* $z = \beta_2$ *and effective memories* $\tau_1 := 1/(1-\beta_1)$*,* $\tau_2 := 1/(1-\beta_2)$.

Proof. Applying $\mathcal{Z}\{\cdot\}$ to $\hat{m}_t = \beta_1\hat{m}_{t-1} + (1-\beta_1)\hat{g}_t$ and using Lemma D.10, $\hat{M}(z) = \beta_1 z^{-1}\hat{M}(z) + (1-\beta_1)\hat{G}(z)$. Solving for $\hat{M}(z)$ gives the stated form. The second recursion is linear in the sequence $\hat{g}_t^{\circ 2}$, whose transform exists since the sequence is bounded by τ^2, and the same computation applies. The poles correspond to impulse responses $(1-\beta_j)\beta_j^k$ with decay times $1/(1-\beta_j)$. □

Lemma D.12 (Step Bound)**.** *Assume* $\beta_1^2 < \beta_2$ *and* $t \ge t_0$ *with* $\beta_1^{t_0}, \beta_2^{t_0} \le \frac{1}{2}$*. Then every coordinate of every Adam step satisfies*

$$|\Delta\theta_{t,i}| \le \Delta_{\max} := 2\eta C_\beta, \qquad C_\beta := \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\beta_1^2/\beta_2)}},$$

with no lower bound on $\hat{v}$.

Proof. By Cauchy–Schwarz,

$$|m_{t,i}| = \Big|\sum_k (1-\beta_1)\beta_1^k\, \hat{g}_{t-k,i}\Big| \le \Big(\sum_k \frac{(1-\beta_1)^2\beta_1^{2k}}{(1-\beta_2)\beta_2^k}\Big)^{1/2}\Big(\sum_k (1-\beta_2)\beta_2^k\, \hat{g}_{t-k,i}^2\Big)^{1/2} = C_\beta\sqrt{v_{t,i}},$$

the first factor summing a geometric series with ratio $\beta_1^2/\beta_2 < 1$. Bias correction multiplies the ratio $m/\sqrt{v}$ by $\sqrt{1-\beta_2^t}/(1-\beta_1^t) \le 2$ for $t \ge t_0$. Hence $|\Delta\theta_{t,i}| = \eta|\hat{m}_{t,i}|/(\sqrt{\hat{v}_{t,i}} + \epsilon_{\text{opt}}) \le \eta|\hat{m}_{t,i}|/\sqrt{\hat{v}_{t,i}} \le 2\eta C_\beta$. The weight-decay term is handled separately below. □

D.3 Temporal Tracking

Appendix A samples mini-batches with replacement, so the clipped gradient at step s is conditionally independent of the past given $\boldsymbol{\theta}_s$. Define the martingale differences

$$\xi_s := \hat{g}_s - \mathbb{E}[\hat{g}_s \mid \mathcal{F}_{s-1}], \qquad \zeta_s := \hat{g}_s^{\circ 2} - \mathbb{E}[\hat{g}_s^{\circ 2} \mid \mathcal{F}_{s-1}],$$

with $|\xi_{s,i}| \le 2\tau$ and $|\zeta_{s,i}| \le \tau^2$, and the targets

$$m_i^*(\boldsymbol{\theta}) := \mathbb{E}[\hat{g}_{t,i}^2 \mid \boldsymbol{\theta}_t = \boldsymbol{\theta}], \qquad \mathbb{E}[\hat{g}_s \mid \mathcal{F}_{s-1}] = \nabla\mathcal{L}(\boldsymbol{\theta}_s) + b_s, \quad |b_{s,i}| \le \sigma^p \tau^{1-p},$$

the bias bound coming from Lemma D.8.

Lemma D.13 (Lipschitz Targets). *m_i^* is L_m-Lipschitz with $L_m = 2\tau L$, and $\nabla\mathcal{L}$ is L-Lipschitz with $|\partial_i \mathcal{L}| \le \sigma$.*

Proof. For a fixed mini-batch realization, $|\mathrm{clip}_\tau(g_i(\boldsymbol{\theta}_1))^2 - \mathrm{clip}_\tau(g_i(\boldsymbol{\theta}_2))^2| \le 2\tau\,|\mathrm{clip}_\tau(g_i(\boldsymbol{\theta}_1)) - \mathrm{clip}_\tau(g_i(\boldsymbol{\theta}_2))| \le 2\tau L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, using $|a^2 - b^2| \le 2\tau|a-b|$ on $[-\tau,\tau]$, the 1-Lipschitz clip, and the L-Lipschitz mini-batch gradient, itself an average of L-Lipschitz sample gradients. Averaging over the mini-batch law preserves the bound. Lipschitzness of $\nabla\mathcal{L}$ follows from the standing assumption by averaging, and $|\partial_i\mathcal{L}| = |\mathbb{E}g_{t,i}| \le (\mathbb{E}|g_{t,i}|^p)^{1/p} \le \sigma$. □

Lemma D.14 (Window Drift Bound). *For $t \ge t_0$,*

$$\Big|\sum_s w_{t-s}^{(2,t)}\big(m_i^*(\boldsymbol{\theta}_s) - m_i^*(\boldsymbol{\theta}_t)\big)\Big| \le 2L_m\sqrt{d}\,\Delta_{\max}\,\frac{\beta_2}{1-\beta_2},$$

$$\Big\|\sum_s w_{t-s}^{(1,t)}\big(\nabla\mathcal{L}(\boldsymbol{\theta}_s) - \nabla\mathcal{L}(\boldsymbol{\theta}_t)\big)\Big\| \le 2L\sqrt{d}\,\Delta_{\max}\,\frac{\beta_1}{1-\beta_1}.$$

Proof. Lemma D.12 and the telescoping sum of steps give $\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_t\| \le \sqrt{d}\,(t-s)\Delta_{\max}$, the factor $\sqrt{d}$ converting the coordinatewise bound into a Euclidean one. Lipschitz continuity (Lemma D.13) turns this into $|m_i^*(\boldsymbol{\theta}_s) - m_i^*(\boldsymbol{\theta}_t)| \le L_m\sqrt{d}\,(t-s)\Delta_{\max}$. Summing against the weights, $\sum_k k\, w_k^{(2,t)} = \frac{(1-\beta_2)}{1-\beta_2^t}\sum_k k\beta_2^k \le \frac{\beta_2}{(1-\beta_2)(1-\beta_2^t)} \le \frac{2\beta_2}{1-\beta_2}$ for $t \ge t_0$. The momentum bound is the same computation with (L, β_1). □

Lemma D.15 (Filtered Noise Variance). *For $t \ge t_0$,*

$$\mathrm{Var}\Big[\sum_s w_{t-s}^{(2,t)}\,\zeta_{s,i}\Big] \le 4\,\frac{1-\beta_2}{1+\beta_2}\,\tau^4, \qquad \mathrm{Var}\Big[\sum_s w_{t-s}^{(1,t)}\,\xi_{s,i}\Big] \le 16\,\frac{1-\beta_1}{1+\beta_1}\,\tau^2.$$

Proof. Martingale differences are orthogonal in L^2, so the variance of the weighted sum equals $\sum_k (w_k^{(2,t)})^2\,\mathbb{E}[\zeta_{t-k,i}^2] \le \tau^4\frac{(1-\beta_2)^2}{(1-\beta_2^t)^2}\sum_k \beta_2^{2k} \le 4\tau^4\,\frac{1-\beta_2}{1+\beta_2}$. The same computation with $|\xi_{s,i}| \le 2\tau$ gives the momentum bound. Neither stationarity nor decorrelation of the noise is used, since the martingale structure of fresh mini-batches replaces the correlation-time hypothesis that would otherwise be required. □

The variance bounds hold at a fixed t. The supermartingale argument below needs the tracking errors controlled at every step up to the exit time, so we pass from variances to exponential tail bounds and take a union over the horizon.

Lemma D.16 (Uniform Tracking over the Horizon). *Set*

$$\delta' := \tau^2\sqrt{8\,\tfrac{1-\beta_2}{1+\beta_2}\,\log\tfrac{2nT}{\delta}}, \qquad \delta'' := \tau\sqrt{32\,\tfrac{1-\beta_1}{1+\beta_1}\,\log\tfrac{2nT}{\delta}}.$$

With probability at least $1-\delta$, simultaneously for every $t_0 \le t \le T$ and every coordinate i of the block,

$$\Big|\sum_s w_{t-s}^{(2,t)}\,\zeta_{s,i}\Big| \le \delta',$$

and, on a second event of probability at least $1-\delta$, $\Big|\sum_s w_{t-s}^{(1,t)}\,\xi_{s,i}\Big| \le \delta''$.

Proof. For fixed t and i, $\sum_s w_{t-s}^{(2,t)} \zeta_{s,i}$ is a sum of martingale differences with increments bounded by $w_{t-s}^{(2,t)} \tau^2$, so the Azuma–Hoeffding inequality gives $\mathbb{P}(|\sum_s w_{t-s}^{(2,t)} \zeta_{s,i}| \geq x) \leq 2\exp\big(-x^2/(2\tau^4 \sum_k (w_k^{(2,t)})^2)\big) \leq 2\exp\big(-x^2(1+\beta_2)/(8\tau^4(1-\beta_2))\big)$, using $\sum_k (w_k^{(2,t)})^2 \leq 4\frac{1-\beta_2}{1+\beta_2}$ from the proof of Lemma D.15. Setting the right side to $\delta/(nT)$ gives $x = \delta'$, and a union bound over the nT pairs (t,i) gives the first event. The same computation with increments bounded by $2\tau w_{t-s}^{(1,t)}$ and $\sum_k (w_k^{(1,t)})^2 \leq 4\frac{1-\beta_1}{1+\beta_1}$ gives the second with $x = \delta''$. Boundedness of the clipped gradients makes the exponential bound available, and the horizon enters only through $\sqrt{\log(nT/\delta)}$. □

Corollary D.17 (Temporal Tracking Bound)**.** *On the first event of Lemma D.16, for every $t_0 \leq t \leq T$ and every coordinate i of the block,*

$$|\hat{v}_{t,i} - m_i^*(\boldsymbol{\theta}_t)| \leq 2L_m \sqrt{d}\, \Delta_{\max} \tfrac{\beta_2}{1-\beta_2} + \delta'.$$

Proof. Since $\mathbb{E}[\hat{g}_{s,i}^2 \mid \mathcal{F}_{s-1}] = m_i^*(\boldsymbol{\theta}_s)$,

$$\hat{v}_{t,i} - m_i^*(\boldsymbol{\theta}_t) = \sum_s w_{t-s}^{(2,t)} \big(m_i^*(\boldsymbol{\theta}_s) - m_i^*(\boldsymbol{\theta}_t)\big) + \sum_s w_{t-s}^{(2,t)}\, \zeta_{s,i},$$

and the weights sum to one. Combine Lemma D.14 with Lemma D.16. □

Lemma D.18 (Momentum Tracking)**.** *On the second event of Lemma D.16, for every $t_0 \leq t \leq T$ and every coordinate i of the block,*

$$|\hat{m}_{t,i} - \partial_i \mathcal{L}(\boldsymbol{\theta}_t)| \leq r_{\max} + \delta'', \qquad r_{\max} := 2L\sqrt{d}\, \Delta_{\max} \frac{\beta_1}{1-\beta_1} + \sigma^p \tau^{1-p}.$$

Proof. Decompose

$$\hat{m}_t - \nabla \mathcal{L}(\boldsymbol{\theta}_t) = \sum_s w_{t-s}^{(1,t)} \big(\nabla \mathcal{L}(\boldsymbol{\theta}_s) - \nabla \mathcal{L}(\boldsymbol{\theta}_t)\big) + \sum_s w_{t-s}^{(1,t)}\, b_s + \sum_s w_{t-s}^{(1,t)}\, \xi_s,$$

and bound the three terms by Lemma D.14, by Lemma D.8 with weights summing to one, and by Lemma D.16. □

Remark D.19 (Momentum Lag). The momentum is the numerator of the update, and controlling $\hat{v}$ alone leaves it untouched. The exact recursion lets $\hat{m}_t$ point against the current gradient after a sign change. From $m_0 = 0$ with $g_1 = -1$ and $g_2 = a \in (0, \beta_1)$, $\hat{m}_2 = (\beta_1 g_1 + g_2)/(1+\beta_1) = (a - \beta_1)/(1+\beta_1) < 0$ while $g_2 > 0$. Lemma D.18 bounds this lag by the gradient variation over one memory window, $L\sqrt{d}\, \Delta_{\max} \beta_1/(1-\beta_1)$, and this quantity enters the transverse drift below.

D.4 Cross-Sectional Homogeneity

Corollary D.17 bounds the error accrued from a single coordinate's own second-moment estimate drifting away from its target over time. A second source of error arises across coordinates within a symmetric block. Let A be generated by the full permutation group on an S_n-symmetric block, as in Theorem C.2. Every coordinate in such a block projects onto the same point of A; for units with several coordinates this holds role by role, over the coordinates the permutation exchanges.

Lemma D.20 (Common Projection Point)**.** *For every coordinate i in the block, $\pi_A(\boldsymbol{\theta})_i = \bar{\theta}^{\mathrm{blk}} := \frac{1}{n} \sum_{k \in \mathrm{block}} \theta_k$.*

Proof. A restricted to the block is $\{\theta \in \mathbb{R}^n : \theta_i = \theta_j\ \forall i,j\}$, since A is the fixed-point set of the full permutation group on the block (Theorem C.2). The orthogonal projection of θ onto this subspace minimizes $\sum_k (\theta_k - c)^2$ over the diagonal point $c\mathbf{1}$. Differentiating in c gives $c = \bar{\theta}^{\mathrm{blk}}$. □

Assumption D.21 (Non-Degeneracy)**.** $\hat{v}_i(\boldsymbol{\theta}_t) \geq v_{\min} > 0$ for every coordinate i under analysis, for all $t \leq \tau_\epsilon^{\mathrm{blk}}$.

Lemma D.22 (Pathwise Block Homogeneity). *Let $\tau_\epsilon^{\mathrm{blk}} := \min_{k\in\mathrm{block}} \tau_\epsilon^{(k)}$. For $t \le \tau_\epsilon^{\mathrm{blk}}$ and any coordinates i, j of the block exchanged by the permutation, almost surely,*

$$|\hat v_{t,i} - \hat v_{t,j}| \le 4\sqrt{2}\, L\tau\sqrt{\epsilon}, \qquad \|E_t\|_\infty \le \kappa \cdot 4\sqrt{2}\, L\tau\sqrt{\epsilon}, \quad \kappa := \frac{1}{2\sqrt{v_{\min}}(\sqrt{v_{\min}} + \epsilon_{\mathrm{opt}})^2},$$

where E_t denotes the deviation of $P(\boldsymbol{\theta}_t)|_{\mathrm{block}} = \mathrm{diag}(1/(\sqrt{\hat v_t} + \epsilon_{\mathrm{opt}}))$ from c_t, the diagonal matrix that assigns to each coordinate the average of $1/(\sqrt{\hat v_t} + \epsilon_{\mathrm{opt}})$ over the coordinates the permutation exchanges with it (one value per coordinate role, common to both units), and κ is evaluated using Assumption D.21. If the pair enters the tube at step t_{enter}, gradients from before entry add at most $2\tau^2\beta_2^{t-t_{\mathrm{enter}}}$ to the first bound, which falls below it after a burn-in of $\tau_2 \ln\big(\tau/(2\sqrt{2}\, L\sqrt{\epsilon})\big)$ steps.

Proof. By Lemma D.20 and Lemma B.2, $|\theta_{t,i} - \theta_{t,j}| \le \sqrt{Y_t^{(i)}} + \sqrt{Y_t^{(j)}} \le 2\sqrt{\epsilon}$ for $t \le \tau_\epsilon^{\mathrm{blk}}$. Let P_{ij} swap coordinates i and j. Equivariance gives $g_{s,i}(\boldsymbol{\theta}) = g_{s,j}(P_{ij}\boldsymbol{\theta})$ for every mini-batch, and $\|P_{ij}\boldsymbol{\theta} - \boldsymbol{\theta}\| = \sqrt{2}\,|\theta_i - \theta_j|$, so

$$|g_{s,i} - g_{s,j}| = |g_{s,j}(P_{ij}\boldsymbol{\theta}_s) - g_{s,j}(\boldsymbol{\theta}_s)| \le L\sqrt{2}\,|\theta_{s,i} - \theta_{s,j}| \le 2\sqrt{2}\, L\sqrt{\epsilon}.$$

Clipping is 1-Lipschitz, so the same bound holds for $\hat g$ with no boundary case. A hard truncation $x\,\mathbb{1}\{|x| \le \tau\}$ would separate the inputs $\tau \pm h$ by $\tau - h$, while clip_τ separates them by h. Then $|\hat g_{s,i}^2 - \hat g_{s,j}^2| \le 2\tau|\hat g_{s,i} - \hat g_{s,j}| \le 4\sqrt{2}\, L\tau\sqrt{\epsilon}$ for every $s \le t$, and averaging with the weights $w^{(2,t)}$ gives the bound on $\hat v$. Each entry of c_t is a convex combination of the $1/(\sqrt{\hat v_{t,k}} + \epsilon_{\mathrm{opt}})$ over exchanged coordinates. The map $v \mapsto 1/(\sqrt{v} + \epsilon_{\mathrm{opt}})$ has derivative bounded by κ on $[v_{\min}, \infty)$, and the mean value theorem transfers the bound to $\|E_t\|_\infty$. □

D.5 The Reduced Transverse Dynamics

The update on the block is $\Delta\boldsymbol{\theta}_t = -\eta(c_t + E_t)\hat m_t - \eta\lambda\boldsymbol{\theta}_{t-1}\mathbb{1}_{\mathrm{AdamW}}$, and the transverse coordinate $\boldsymbol{x}_t = P^\perp\boldsymbol{\theta}_t$ follows by applying $P^\perp$. The term $E_t = O(\sqrt{\epsilon})$, bounded through the history of $|\theta_i - \theta_j|$ on the tube, multiplies $\hat m_t = O(1)$, which is the order of $c_t\nabla_\perp\mathcal{L}$ at the tube boundary, and its sign follows that of $(\hat v_{t,i} - \hat v_{t,j})\,\partial_j\mathcal{L}$. We therefore freeze it.

Assumption D.23 (Quasi-Static Preconditioner). $1 - \beta_2$ is small enough that $\hat v$ on the block is constant over the trapping horizon, so $c_t = c$ and $E_t = E$.

Assumption D.23 describes the regime in which τ_2 exceeds the transverse relaxation time. The quantity that checks it on a run is the relative variation $\max_{t_0\le t\le T} |\hat v_{t,k} - \hat v_{t_0,k}|/\hat v_{t_0,k}$ over the block, and under it E is a constant to be measured rather than bounded. Remark D.30 reports this relative variation on the grokking task.

Proposition D.24 (Preconditioned SGF Reduction). *Under Assumptions D.4, D.21 and D.23, and reading the exponentially filtered momentum noise as white noise with its long-run covariance, for $t_0 \le t \le \tau_\epsilon^{\mathrm{blk}}$ the transverse coordinate of the block evolves as the Euler–Maruyama step, at the resolution of one iteration, of*

$$d\boldsymbol{x}_t = -\eta\, c\,\nabla_\perp\mathcal{L}(\boldsymbol{\theta}_t)\,dt - \eta\lambda\,\boldsymbol{x}_t\,\mathbb{1}_{\mathrm{AdamW}}\,dt + \boldsymbol{f}\,dt + \boldsymbol{\varrho}_t\,dt + \eta\, c\,(\Sigma_\perp^\tau(\boldsymbol{\theta}_t))^{1/2}\,dW_t, \tag{36}$$

where $\boldsymbol{f} := -\eta P^\perp E\,\nabla\mathcal{L}(\boldsymbol{\theta}_t)$ satisfies $\|\boldsymbol{f}\| \le \eta\sqrt{n}\,\|E\|_\infty\,\sigma$, $\Sigma_\perp^\tau := P^\perp\,\mathrm{Cov}(\hat g_t \mid \boldsymbol{\theta}_t)\,P^\perp$ is the transverse covariance of the clipped mini-batch gradient, and, on the second event of Lemma D.16, for every $t_0 \le t \le T$,

$$\|\boldsymbol{\varrho}_t\| \le \rho_{\max} := \eta\sqrt{n}\,\big(\|c\|_\infty + \|E\|_\infty\big)\,(r_{\max} + \delta'').$$

Proof. Substitute $\hat m_t = \nabla\mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{r}_t + \boldsymbol{\xi}_t^m$ from Lemma D.18, with $\boldsymbol{r}_t$ the deterministic lag-plus-bias and $\boldsymbol{\xi}_t^m := \sum_s w_{t-s}^{(1,t)}\xi_s$ the filtered martingale noise, into $P^\perp\Delta\boldsymbol{\theta}_t = -\eta P^\perp(c + E)\hat m_t - \eta\lambda\boldsymbol{x}_{t-1}\mathbb{1}_{\mathrm{AdamW}}$. Since c is common to both units it commutes with the swap and with $P^\perp$, so the drift $-\eta c\nabla_\perp\mathcal{L}$ and the forcing $\boldsymbol{f}$ come out of $\nabla\mathcal{L}$, and the bound on $\boldsymbol{f}$ uses $|\partial_i\mathcal{L}| \le \sigma$ (Lemma D.13). The residual $\boldsymbol{\varrho}_t := -\eta P^\perp(c + E)\boldsymbol{r}_t - \eta P^\perp E\boldsymbol{\xi}_t^m$ collects the deterministic errors and the part of the filtered noise that the deviation E multiplies, and Lemma D.18 bounds $\|\boldsymbol{r}_t + \boldsymbol{\xi}_t^m\|$ coordinatewise by $r_{\max} + \delta''$ on the stated event, which gives $\rho_{\max}$. The remaining noise $-\eta c P^\perp\boldsymbol{\xi}_t^m$ enters as the diffusion of equation 36 through the white-noise reading stated in the proposition. Weight decay contributes $-\eta\lambda\boldsymbol{x}$ exactly, since $P^\perp$ is linear and $P^\perp\boldsymbol{\theta} = \boldsymbol{x}$. □

Remark D.25 (The White-Noise Reading). The reading is a modeling step, since the recursion does not imply it, and Theorem D.26 is stated for the reduced SDE while the tracking bounds are discrete, so the Euler–Maruyama error between the two is not controlled here. The filtered noise $\boldsymbol{\xi}_t^m = \sum_s w_{t-s}^{(1,t)} \xi_s$ is an exponential average of martingale differences, so most of it at step t is carried over from step $t-1$ and only the fresh increment $w_0^{(1,t)}\,\xi_t$ is unpredictable given $\mathcal{F}_{t-1}$. Its per-step variance is smaller than its long-run variance by the factor $\frac{1-\beta_1}{1+\beta_1}$, and on time scales beyond τ_1 its increments decorrelate and the long-run covariance equals that of its input, $\eta^2 c\,\Sigma_\perp^\tau\, c$, which is the covariance equation 36 carries. The step that turns SGD into the SGF of Appendix A swaps independent increments for Gaussian ones, while this one swaps correlated increments for independent ones, so it is a modeling step of the same kind and a stronger one. The statements below take it as a hypothesis. The filtered noise also enters $\rho_{\max}$ through $c\,\delta''$, so it is counted once as forcing and once as diffusion, which errs on the safe side. A route that drops the reading altogether bounds the whole increment pathwise on the annulus, using Lemma D.16, and expands Y^α to second order with a remainder. It yields a discrete supermartingale inequality without the concavity term on the right of equation 37, with the pathwise second-order piece $\|\Delta \boldsymbol{x}\|^2 \le n\Delta_{\max}^2$ for Adam, and $\le (\sqrt{n}\,\Delta_{\max} + \eta\lambda\|\boldsymbol{x}\|)^2$ for AdamW since Lemma D.12 excludes the decay step, in place of $\eta^2 \operatorname{tr}(c\,\Sigma_\perp^\tau\, c)$. Dropping the concavity term discards a favorable contribution, so the condition it produces is stronger than equation 37 and needs no white-noise reading.

This reduced transverse dynamics has the structural form of the vanilla-SGD flow, with the drift and the diffusion preconditioned by c and hence the covariance by $c(\cdot)c$, a forcing $\boldsymbol{f}$ off A, and a residual. We now state the condition under which the trapping argument survives. Every statement below carries the white-noise reading of Remark D.25 as a hypothesis.

Theorem D.26 (Conditional Trapping Criterion for the Reduced Adam Dynamics). *Fix $\alpha \in (0,1]$ and $0 < r_0 < \sqrt{\epsilon}$, let $\rho_\parallel \le \rho_{\max}$ be a constant with $|\boldsymbol{x}_t^\top \boldsymbol{\varrho}_t| \le \rho_\parallel \sqrt{Y_t}$ for $t_0 \le t \le T$ on the tracking event of Lemma D.16, and suppose that on the annulus $\mathcal{N}_{r_0,\epsilon} := \{r_0^2 \le Y < \epsilon\}$*

$$-2\eta\,\boldsymbol{x}^\top c\,\nabla_\perp \mathcal{L} - 2\eta\lambda Y\,\mathbb{1}_{\mathrm{AdamW}} + 2\,\boldsymbol{x}^\top \boldsymbol{f} + 2\sqrt{Y}\,\rho_\parallel + \eta^2 \operatorname{tr}(c\,\Sigma_\perp^\tau\, c) \;\le\; 2(1-\alpha)\,\eta^2\,\frac{\boldsymbol{x}^\top c\,\Sigma_\perp^\tau\, c\,\boldsymbol{x}}{Y}. \tag{37}$$

The conclusion also holds with $2\,\boldsymbol{x}^\top \boldsymbol{f} + 2\sqrt{Y}\,\rho_\parallel$ replaced by $2\sqrt{Y}\,(\|\boldsymbol{f}\| + \rho_{\max})$. Let $\tau_{\mathrm{ann}} := \inf\{t \ge t_0 : Y_t \notin \mathcal{N}_{r_0,\epsilon}\}$. Then $Y_{t\wedge\tau_{\mathrm{ann}}}^\alpha$ is a non-negative supermartingale under equation 36.

Proof. Itô's lemma for $\phi_\alpha(Y) = Y^\alpha$ under equation 36 gives

$$\mathscr{G}Y^\alpha = \alpha Y^{\alpha-1}\Big[-2\eta\,\boldsymbol{x}^\top c\,\nabla_\perp\mathcal{L} - 2\eta\lambda Y\,\mathbb{1}_{\mathrm{AdamW}} + 2\boldsymbol{x}^\top(\boldsymbol{f}+\boldsymbol{\varrho}_t) + \eta^2\operatorname{tr}(c\,\Sigma_\perp^\tau\, c) - 2(1-\alpha)\,\eta^2\,\frac{\boldsymbol{x}^\top c\,\Sigma_\perp^\tau\, c\,\boldsymbol{x}}{Y}\Big].$$

By the definition of $\rho_\parallel$, $2\boldsymbol{x}^\top\boldsymbol{\varrho}_t \le 2\sqrt{Y}\rho_\parallel$, so equation 37 makes the bracket non-positive on the annulus, and Cauchy–Schwarz bounds $2\boldsymbol{x}^\top(\boldsymbol{f}+\boldsymbol{\varrho}_t) \le 2\sqrt{Y}(\|\boldsymbol{f}\| + \rho_{\max})$ for the cruder form. Preconditioning drift and noise amplitude by the same c scales the drift at first order and the diffusion at second order in c, so the sign of the bracket is decided by equation 37, and $c > 0$ alone does not determine it. For $dX = -cX\,dt + cX\,dW$ the generator of X^2 equals $(c^2 - 2c)X^2$, which is positive for $c > 2$, while the generator of $\log X^2$ equals $-(2c + c^2)$, which is negative for every c. The concavity term on the right of equation 37 carries the c^2 contribution to the favorable side. In transverse dimension one with $\Sigma_\perp^\tau = \zeta^2 Y$ the two c^2 terms combine to $(2\alpha - 1)\,\eta^2 c^2 \zeta^2 Y$, so for $\alpha \le \frac{1}{2}$ the noise helps for every c, while at $\alpha = 1$ the condition reads $c \le 2$ in the example above and excludes $c = 3$. On the annulus $Y \ge r_0^2$, the integrand of the stochastic integral, $2\alpha Y^{\alpha-1}\eta c\,\boldsymbol{x}^\top(\Sigma_\perp^\tau)^{1/2}$, is bounded, so the integral is a genuine martingale (Karatzas & Shreve, 1991) and the conditional expectation annihilates it, leaving the non-positive drift. □

Corollary D.27 (Conditional Non-Escape and Confinement). *Over the horizon $t_0 \le t \le T$,*

$$\mathbb{P}\big(Y \text{ exits the annulus through } \epsilon \text{ before reaching the core and before } T \mid \mathcal{F}_{t_0}\big) \le \epsilon^{-\alpha}\,\mathbb{E}[Y_{t_0}^\alpha \mid \mathcal{F}_{t_0}] + 2\delta.$$

Let $\tilde{Y} := \|\boldsymbol{x} - \boldsymbol{x}^*\|^2$*, where* $\boldsymbol{x}^*$ *is the zero of the transverse drift, and suppose that* $\mathscr{G}\tilde{Y} \le -2\kappa\tilde{Y} + C$ *on* $\{\tilde{Y} < \epsilon\}$ *for constants* $\kappa > 0$ *and* $C \ge 0$*. Then*

$$\mathbb{P}\big(\tilde{Y} \text{ exits the tube before } T \mid \mathcal{F}_{t_0}\big) \le \frac{\tilde{Y}_{t_0} + C\,(T - t_0)}{\epsilon} + 2\delta.$$

Proof. Let τ_{bad} be the first step $t \ge t_0$ at which either tracking bound of Lemma D.16 fails. Both bounds are $\mathcal{F}_t$-measurable, so τ_{bad} is a stopping time, and Lemma D.16 gives $\mathbb{P}(\tau_{\mathrm{bad}} \le T) \le 2\delta$. Before $\tau_{\mathrm{ann}} \wedge \tau_{\mathrm{bad}}$ the residual bound $\rho_{\max}$, and with it $\rho_{\|} \le \rho_{\max}$, holds at every step, so equation 37 holds and Theorem D.26 makes $Y^{\alpha}_{t\wedge\tau_{\mathrm{ann}}\wedge\tau_{\mathrm{bad}}}$ a non-negative supermartingale without conditioning on any event. Doob's maximal inequality (Revuz & Yor, 1999) bounds $\mathbb{P}(\sup_{t\le T} Y^{\alpha}_{t\wedge\tau_{\mathrm{ann}}\wedge\tau_{\mathrm{bad}}} \ge \epsilon^{\alpha})$ by the first term. An exit through ϵ before T either happens before τ_{bad}, which the stopped process records, or after it, which costs $\mathbb{P}(\tau_{\mathrm{bad}} \le T) \le 2\delta$. A single-step tail bound does not give uniformity over the horizon, and with Chebyshev in place of Azuma–Hoeffding the union bound would cost a factor T rather than $\log T$. For the confinement bound, the generator of $\tilde{Y}_t + C(T - t)$ is $\mathscr{G}\tilde{Y} - C \le -2\kappa\tilde{Y} \le 0$ on the tube, and $\tilde{Y}_t + C(T - t) \ge 0$ for $t \le T$. On the tube the integrand of the stochastic integral, $2\eta c\,(\boldsymbol{x} - \boldsymbol{x}^*)^{\top}(\Sigma^{\tau}_{\perp})^{1/2}$, is bounded, so the integral is a martingale and the stopped process is a non-negative supermartingale on $[t_0, T]$. Doob's maximal inequality bounds $\mathbb{P}(\sup_{t_0\le t\le T} \tilde{Y}_{t\wedge\tau} \ge \epsilon)$ by $(\tilde{Y}_{t_0} + C(T - t_0))/\epsilon$, and the tracking failure adds 2δ. □

The first bound covers the first exit of the annulus. A path that enters the core $\{Y < r_0^2\}$ can return to the annulus and exit through ϵ, each such excursion costs at most $(r_0^2/\epsilon)^{\alpha}$, and Theorem D.26 does not bound their number. The second bound covers the whole horizon. Its constant C is the supremum over the tube of $\eta^2\,\mathrm{tr}(c\,\Sigma^{\tau}_{\perp}\,c) + 2\sqrt{\epsilon}(\|\boldsymbol{f}\| + \rho_{\|})$, the terms that do not vanish at $\boldsymbol{x}^*$, and κ is the inward rate of the terms linear in $\tilde{Y}$. A nondegenerate diffusion is confined near $\boldsymbol{x}^*$ rather than trapped at it, so the second bound supplies the link survival of Lemma B.6, Theorem D.29 assumes it, and Remark D.30 measures C.

Trajectories that reach the core $\{Y < r_0^2\}$ lie at a distance $O(r_0)$ from A, where the forcing $\boldsymbol{f}$ and the residual set the scale. The transverse drift $-\eta c\nabla_{\perp}\mathcal{L} + \boldsymbol{f}$ vanishes at a point off A, and the trap is that point rather than A itself. This is the direct analogue, under Adam and AdamW, of Corollary B.4, with the invariant set replaced by a shifted trap and the tube by an annulus. Remark D.30 reports the measurement of these quantities on the grokking task.

Assumption D.28 (Macroscopic Merges)**.** The merging components at the observed microtransitions satisfy $i(N) = \Theta(N)$.

Theorem D.29 (Conditional Cascade Transfer to Adam and AdamW)**.** *Under Assumptions D.4, D.21, D.23, D.28, 4.3 and 4.7, and the confinement condition* $\mathscr{G}\tilde{Y} \le -2\kappa\tilde{Y} + C$ *of Corollary D.27, on* $t \le \tau_{\epsilon}^{\mathrm{blk}}$ *the Adam- or AdamW-trained trajectory admits a percolation process* G_p *whose* R_v *peaks form the cascade of Theorem C.23 with the same factor* λ_t*, with escape probability bounded as in Corollary D.27.*

Proof. The confinement bound of Corollary D.27 supplies the link-survival probability that Lemma B.6 provides in the SGD case. Assumption 4.3 supplies the monotone control parameter of Proposition C.9. Proposition C.13 and Corollary C.14 apply verbatim, and Assumption D.28 places the merges in the macroscopic regime. Assumption 4.7 is a statement about the link rates of the process as run, so the frozen preconditioner c, which is a constant of the block's dynamics, never enters the ratio of Theorem C.23. □

Remark D.30 (Measurement of the Adam Conditions on the Grokking Task). We measured the quantities of Theorem D.26 on the AdamW-trained modular-arithmetic Transformer of Section 6, the grokking task (Appendix F), for two fixed permutation pairs of feed-forward hidden units, $(3, 7)$ and $(11, 42)$, and for the closest pair among all 512 units, at every 100 epochs over 30000 epochs, which is 301 checkpoints and 60 multiples of τ_2 at two steps per epoch. The block of a unit consists of its input weights, its bias and its output weights (257 coordinates), and Y is the transverse distance of the pair. At each checkpoint we computed the exact full-batch gradient (the training set has 202 examples), $\Sigma^{\tau}_{\perp}$ from 32 minibatch gradients, c and E from the bias-corrected $\hat{v}$, and the deterministic lag $\boldsymbol{r}_t$ as the difference between the β_1-average of the exact gradient over the 30 steps preceding the

checkpoint and the gradient at the checkpoint, so that ρ is $\eta\|P^{\perp}[(cI+E)\boldsymbol{r}_t + E\,\boldsymbol{\xi}_t^m]\|$ as Proposition D.24 defines it. We recorded $\hat{v}$ on the block every 10 steps and flagged an epoch as a loss spike when the training loss exceeded five times the median of the preceding 50 epochs. Two further realizations with different seeds were measured in the same way, one with the same two pairs and one with 50 pairs of units drawn at random, together with the distribution of Y over all pairs of live units, defined as units whose norm exceeds half the median norm. Where the realizations differ, the numbers of the other realizations are given in parentheses. The condition, drift and tube statistics are from three further realizations (seeds 1000 to 1002), computed with the role-wise preconditioner c of Lemma D.22 and with $\Sigma_{\perp}^{\tau}$ and $\nabla\mathcal{L}$ measured with dropout active, as in training.

Preconditioner. Relative to a single scalar, the deviation $\|E\|_{\infty}/c$ has median 1.46 and 1.61 for the two pairs and reaches 15 and 35 at the spikes. Most of it is role-wise: split as in Lemma D.22, the spread of c across coordinate roles is 0.80 to 0.91 of its mean, while the cross-unit part E is 0.75 to 1.03 of it and nearly independent of $\sqrt{Y}$ (log-log slope 0.07 to 0.14), since slingshots recur before the burn-in of Lemma D.22 completes. E is therefore measured rather than bounded (Assumption D.23), and the forcing $\boldsymbol{f}$ has the order of the drift at the tube boundary (Proposition D.24). The medians of $\|\boldsymbol{f}\|$ and ρ are 4.1×10^{-4} and 4.8×10^{-4} for $(3,7)$ and 7.8×10^{-4} and 7.6×10^{-4} for $(11,42)$. Replacing $\boldsymbol{r}_t$ by the full lag $\hat{m}-\nabla\mathcal{L}$ raises ρ by a factor of only 1.1 to 1.6, so the lag on this run is deterministic, since minibatches of 128 out of 202 examples carry little noise, and Lemma D.18 is the operative bound while the noise term of Lemma D.16 is small. The relative variation of $\hat{v}$ over windows of length τ_2 has median 1.6 and 2.4 on the two blocks, which is a factor of 1.6 to 1.8 in c over a typical window, and 90th percentile 290 and 600, which is a factor of 17 to 25 in c on the windows that contain a spike. Assumption D.23 therefore holds to within a factor of two inside a condensation phase, each spike opens a new window, and $\hat{v}$ on the block spans $10^{-9.2}$ to $10^{-3.6}$ over the run. The empirical autocorrelation time of $\hat{g}^{\circ 2}$ is one step, $\sqrt{\hat{v}}$ on the block exceeds ϵ_{opt} by two orders of magnitude, and $\max|g|/\mathrm{std}(g)\approx 77$.

Condition. Condition equation 37 holds at 50 to 62% of checkpoints for the pairs $(3,7)$ and $(11,42)$, at 52 to 58% averaged over the 50 random pairs and at 62 to 65% for the closest pair, with phase averages between 0.47 and 0.62 from memorization to the end of training. With the noise measured with dropout off, the same seeds give 64 to 78%, and replacing the role-wise c by a scalar changes every fraction by at most 0.02. The cruder form with $2\sqrt{Y}(\|\boldsymbol{f}\|+\rho_{\max})$ holds at 14 to 23%, because $\boldsymbol{f}$ is nearly orthogonal to $\boldsymbol{x}$ and the Cauchy–Schwarz step discards most of the margin. The fractions are the same for every α between 0.05 and 1, since the diffusion terms lie about two and a half orders of magnitude below the drift. Dropout raises the transverse noise 31- to 46-fold, and the noise still vanishes at coincidence, scaling as $Y^{0.96}$ to $Y^{1.04}$ for live pairs.

Drift. $\cos(\boldsymbol{x},\boldsymbol{D})$ with $\boldsymbol{D}=-\eta c\nabla_{\perp}\mathcal{L}-\eta\lambda\boldsymbol{x}+\boldsymbol{f}$ is negative at 63 to 74% of checkpoints for the two pairs and the 50 random pairs. The inward pull comes from weight decay: evaluated without dropout, the gradient term alone points outward, $\boldsymbol{x}^{\top}\nabla_{\perp}\mathcal{L}<0$, at 89 to 95% of checkpoints on contracting and expanding segments alike, while $\boldsymbol{D}$ points inward at 96 to 97% of checkpoints on contracting segments and at 57 to 65% on expanding ones.

Contraction and restarts. The training loss shows 44 spike events over the run, with mean spacing 679 epochs, which is $1.4\,\tau_2$, and median duration 28 epochs. These are the slingshot instabilities of Adam with weight decay (Thilak et al., 2022), and each one starts a new condensation phase. Within a phase Y evolves geometrically. Of the 45 inter-spike segments per pair, 60% contract (53% averaged over the 50 random pairs), and the segments that expand are those on which the full drift $\boldsymbol{D}$ points inward less often. Among the contracting segments the median e-fold time of Y is 1226 steps for $(3,7)$ and 802 steps for $(11,42)$ (1390 steps over the 50 pairs), against $1/(2\eta\lambda)=333$ steps for weight decay alone, and the fastest phases reach e-fold times of 400 steps. The largest single-phase contractions are 1.7 decades for each pair ($2.5\to0.053$ over epochs 16863 to 17642 and $5.7\to0.11$ over epochs 24519 to 25293), and the second realization contains a phase of 3.6 decades. Over the whole run contraction leads expansion by 2.5 decades for $(3,7)$ (13.0 against 10.5) and by 1.2 decades for $(11,42)$ (17.2 against 16.0), a net drift toward coincidence that the restarts distribute across the run. Among the 344 live units at the end of the 50-pair realization, 114 live pairs lie below $Y=10^{-2}$, the 0.1% quantile of Y over live pairs is 7×10^{-3}, and random live pairs reach $Y\approx10^{-3}$ within a single phase (pair $(54,325)$ at epochs 5163 and 11930, with unit norms 0.4 and 2.8). The closest pair among all units at every checkpoint consists of two units with norms 2×10^{-3} to 5×10^{-2} against a median unit norm of 0.3 to 0.8, so the closest-pair statistic tracks units that

weight decay has driven toward the origin, and the live-pair statistics above are the ones that measure coincidence.

Tube occupancy. Over the three further realizations, the sampled pairs enter the tube $\{Y < 10^{-2}\}$ progressively: 0 to 12% by epoch 2000, 25 to 77% by the second variance peak and 62 to 100% by the end of training, fastest between the first two peaks, with a median stay of 80 to 605 epochs per entry. Most of these entries are units that weight decay drives together toward the origin; among live units the ϵ-graph reaches a merge fraction of 0.17 to 0.24, with the largest cluster holding 4 to 17% of them. Condition equation 37 holds more often the closer the pair: at 0.62 to 0.66 of checkpoints inside the tube and 0.49 to 0.50 for $Y \geq 1$, with the drift inward at 0.71 to 0.76 and 0.61 to 0.62. Within 100 epochs after a spike the condition holds at 0.47 to 0.56 of checkpoints, against 0.52 to 0.59 in quiet windows.

Confinement constant. The constant C of Corollary D.27 is the sum of the diffusion term and the forcing-plus-residual term at the tube boundary, and at typical checkpoints the latter dominates, $2\sqrt{\epsilon}(\|\boldsymbol{f}\| + \rho) \approx 2\sqrt{\epsilon} \times 10^{-3}$, with the diffusion term four orders of magnitude smaller. The bound certifies confinement over horizons $T \lesssim \epsilon/C$, which at this value of C is a few hundred steps for ϵ of order 10^{-2}. The reinflation of Y at each spike occurs beyond this horizon.

Reading. Condition equation 37 holds at most checkpoints inside phases of length about $1.4\,\tau_2$, at 0.3 to 0.4 of the weight-decay contraction rate in the median and at 0.8 in the fastest phases, and the slingshot instabilities restart the condensation phase on that timescale. We treat each inter-spike window as a condensation phase in the sense of Assumption 4.3, with the spikes as its boundaries. The repulsion phases, on which $\boldsymbol{D}$ points outward, are the complement of the annulus in time. The order parameter of the grokking task is the effective rank of the embedding, and the feed-forward block measured here is the composition through which the network reads the embedding, so its coincidence sets are the ones that nest above the embedding's.

E Explicit Theoretical Verification and Toy Models

We construct two distinct experimental environments. First, we impose the topological mechanics in a controlled kinematic setting and check that the detection pipeline recovers the imposed cascade. Second, we deploy an empirical neural network optimized via Stochastic Gradient Descent (SGD) to verify that these topological phase transitions—both condensation and reactive fragmentation—govern non-convex empirical loss landscapes.

E.1 Generalization of the Stochastic Collapse Threshold

To derive the exact mathematical condition under which a subnetwork irreversibly collapses into an invariant set, we analyze the local dynamics of continuous stochastic gradient flow (SGF) (Li et al., 2017).

Consider a network with multiplicative parameter interactions, such as adjacent layers connected by weights w_{in} and w_{out}. Assuming the subnetwork learns a target signal μ under a Mean Squared Error (MSE) objective, the local loss is $\mathcal{L}(w_{in}, w_{out}) = \frac{1}{4}(w_{in}w_{out} - \mu)^2$. Under continuous gradient descent, the difference of squared weights is a conserved quantity: $\frac{d}{dt}(w_{out}^2 - w_{in}^2) = 0$. Assuming a balanced initialization, the parameters satisfy $w_{in} = w_{out} = w$. The transverse dynamics consequently collapse onto a one-dimensional effective loss landscape:

$$\mathcal{L}(w) = \frac{1}{4}(w^2 - \mu)^2. \tag{38}$$

Near the structurally collapsed state at the origin ($w = 0$), the gradient is $\mathcal{L}'(w) = w^3 - \mu w$, establishing $w = 0$ as a stationary invariant manifold where the first derivative vanishes. The local curvature is given by the second derivative $\mathcal{L}''(0) = -\mu$, defining a deterministic repulsive drift away from the manifold.

Simultaneously, the gradient noise covariance matrix $D(w)$ inherently scales with the magnitude of the active weights (Blanc et al., 2020). Near the manifold, the transverse diffusion term scales quadratically, $D(w) \approx \frac{1}{2}\zeta^2 w^2$, yielding purely multiplicative noise. Substituting the exact gradient

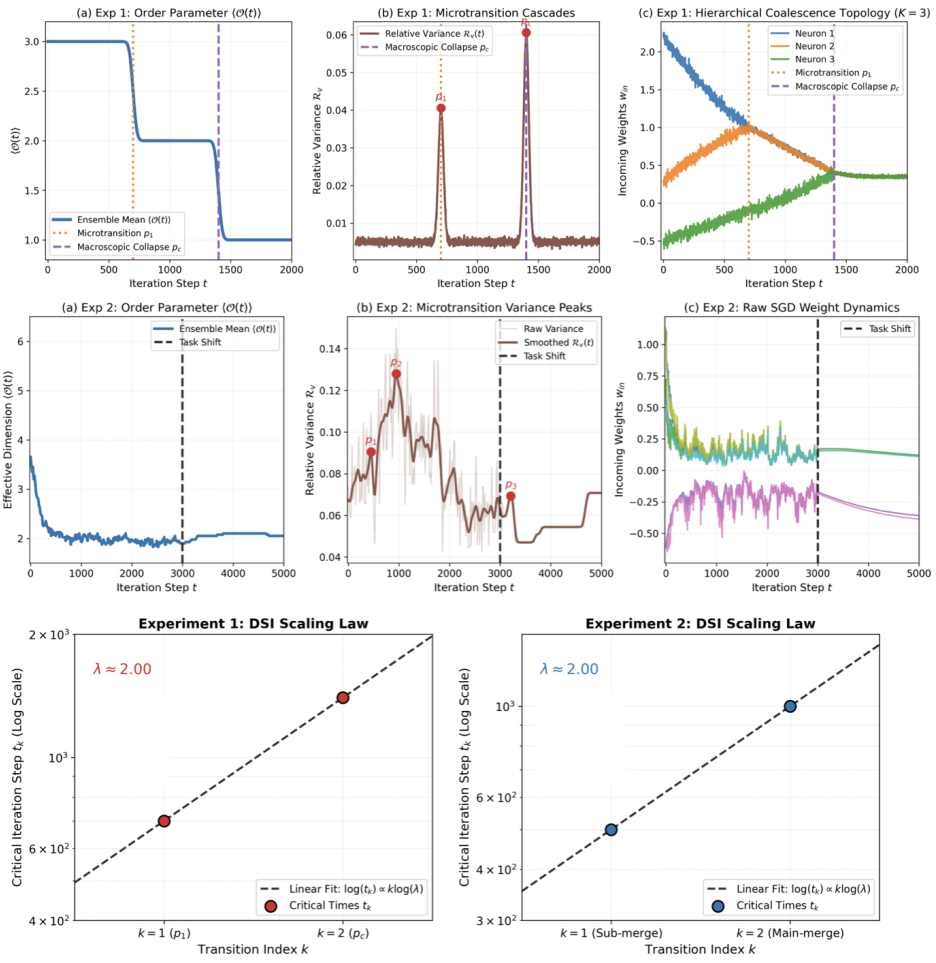


Figure 3: Experimental validation of symmetry-induced topological collapse and the scaling of peak times. **Top Row (Experiment 1):** Kinematic simulation of hierarchical coalescence ($N = 3$), displaying discrete jumps in the order parameter (a), peaks in relative variance exactly at the imposed merge points (b), and forced continuous spline trajectories (c). **Middle Row (Experiment 2):** Empirical SGD dynamics ($N = 6$), showing macroscopic dimensionality collapse (a), distinct precursory variance peaks p_1, p_2 during the condensation phase, and a reactive fragmentation peak p_3 following a task shift at $t = 3000$ (b), alongside the raw stochastic weight trajectories (c). **Bottom Row:** Log-linear progressions of peak times: Experiment 1 (Left) recovers the imposed ratio $\lambda_t = 2.00$, and Experiment 2 (Right) gives the two-peak ratio $\lambda_t \approx 2.00$; the finite-N prediction for the first two merges of $N = 6$ units is 2.5 (Section E).

and the local noise approximation into the continuous SDE $dw_t = -\nabla_w \mathcal{L}(w_t)dt + \sqrt{2D(w_t)}dW_t$ yields:

$$dw_t = -(w_t^3 - \mu w_t)dt + \zeta w_t dW_t. \tag{39}$$

To determine whether the network undergoes collapse, we evaluate the Fokker-Planck equation for localized stationarity ($\partial_t p = 0$). Forcing the probability current to vanish yields the stationary density (Chen et al., 2023)

$$p_{ss}(w) \propto w^{2\mu/\zeta^2 - 2}\, e^{-w^2/\zeta^2}. \tag{40}$$

The global normalization partition function diverges when the exponent of w forces a non-integrable singularity at the origin, requiring $\frac{2\mu}{\zeta^2} - 2 \leq -1$.

Theoretical Threshold: The invariant manifold becomes strictly stochastically attractive—collapsing the transverse probability mass into a Dirac delta distribution $\delta(w)$—if and only if the multiplicative noise variance overcomes the deterministic drift:

$$\mu \leq \frac{\zeta^2}{2}. \tag{41}$$

Because the empirical SGD diffusion amplitude scales proportionally with the learning rate η, maintaining a sufficiently high learning rate explicitly satisfies this threshold, permanently trapping the parameters within the lower-dimensional invariant subspace (Chen et al., 2023).

Definition 2.3 reads this threshold off directly. With $\boldsymbol{x} = w$, $\nabla\mathcal{L} = w^3 - \mu w$ and $\eta\Sigma_\perp = \zeta^2 w^2$, condition equation 2 becomes $2\mu w^2 - 2w^4 + \zeta^2 w^2 \leq 2(1-\alpha)\zeta^2 w^2$, that is $\mu \leq \frac{\zeta^2}{2}(1-2\alpha) + w^2$, and $\alpha \downarrow 0$ recovers $\mu \leq \zeta^2/2$ near the origin. At $\alpha = 1$ the condition reads $2\mu + \zeta^2 \leq 0$ and fails for every $\mu > 0$, because the generator of w^2 is $(2\mu + \zeta^2)w^2 - 2w^4 > 0$ near the origin even as $w \to 0$ almost surely, and for this reason the definition is stated for Y^α rather than for Y. Above the threshold, $Y = w^2$ is Gamma-distributed with shape $\mu/\zeta^2 - \frac{1}{2}$ and scale ζ^2 (Remark C.7).

E.2 Derivation of the Scaling of Peak Times

Once parameters are trapped in the invariant manifold, Definition C.12 and Proposition C.18 make condensation pairwise, $n = 2$, in the generic case. Lemma 4.8 then gives $t_{2i}/t_i = \lambda_t = 2^{1-\gamma_K}$ for the R_v peak times (Theorem C.23, Corollary C.24). For the constant kernel the mean component size grows linearly in time, $\bar{C} = 1 + t/2$, so $T_{2^k} = 2(2^k - 1)$ and $\lambda_t = 2$ asymptotically.

Prediction under the constant kernel: For pairwise merges the sequential peak times t_k detected by the macroscopic relative variance $\mathcal{R}_v(t)$ satisfy the log-linear progression $t_{k+1} = 2t_k$ asymptotically, with exact finite-size ratios $(2i-1)/(i-1)$ for the times at which the largest component reaches sizes i and $2i$. This follows from Assumption 4.7 with $\gamma_K = 0$ through Lemma 4.8. For N units the expected waiting time between the merges that take the component count from M to $M-1$ is $2N/(M(M-1))$ in the same units, so for $N = 6$ the expected times of the first five merges are $0.4, 1.0, 2.0, 4.0, 10.0$ and the first two merges are in ratio 2.5. A measured $\lambda_t \neq 2$ at large N reflects $\gamma_K \neq 0$, $n \neq 2$, or both, and the ratio alone does not separate them.

E.3 Idealized Hierarchical Coalescence

Purpose: In non-convex empirical loss landscapes, topological phase transitions are tightly coupled with geometric distortions, complicating the isolation of pure structural scaling. We impose the topological variable in a controlled environment and check that the detection pipeline recovers it.

Construction: We construct a synthetic kinematic system of $N = 3$ decoupled parameters. We impose sequential, binary hierarchical coalescence ($n = 2$) by routing continuous spline trajectories injected with decaying Gaussian noise, with merge times fixed at $t_1 = 700$ and $t_c = 1400$. This tests whether the $\mathcal{R}_v$ pipeline locates imposed merges and returns the imposed ratio of their times.

Results: The binding of parameters 1 and 2 at $t_1 = 700$, followed by parameter 3 at $t_c = 1400$ (mapped in Figure 3, Top Row, Panel c), causes the macroscopic order parameter $\langle \mathcal{O}(t) \rangle$ (Figure 3, Top Row, Panel a) to collapse via discrete jumps. Driven by the two-point ensemble states at these thresholds, the relative variance $\mathcal{R}_v(t)$ (Figure 3, Top Row, Panel b) registers localized peaks exactly at the imposed merge points. Plotting these transition times on a logarithmic scale against

their sequential index k (Figure 3, Bottom Left) returns $\lambda_t = 1400/700 = \mathbf{2.00}$, the imposed ratio. This checks the pipeline, and the theoretical content of the prediction is tested by the SGD simulation below.

E.4 Empirical Simulation ($N = 6$) with Task Shift and Reverse Transition

Theoretical Justification: Having checked the pipeline, we show that the noise characteristics generated by standard neural network optimization satisfy these topological constraints without external kinematic forcing. Furthermore, the theoretical threshold ($\mu \leq \zeta^2/2$) dictates that the collapse is dynamically reversible: dropping the noise variance while increasing the signal curvature should render the invariant manifold repulsive, forcing reactive fragmentation, which ends the condensation phase of Assumption 4.3.

Empirical Verification: We deploy a PyTorch environment, initializing a $N = 6$ shallow neural network with a GELU activation function, optimized on an MSE objective by label-noise SGD (Blanc et al., 2020): each step uses the full batch of 32 inputs $x \sim \mathcal{N}(0, 1)$ with fresh Gaussian label noise of standard deviation 2.0. The network computes $\sum_{k=1}^{6} w_{out,k}\,\mathrm{gelu}(w_{in,k}x)$ for the target $0.5 \sin 2x$, with learning rate 0.2, weight decay 10^{-3}, gradient-norm clipping at 5 and initialization standard deviation 0.8, and the order parameter counts average-linkage clusters of the $w_{in,k}$ at distance 0.25. We manipulate the theoretical stochastic collapse condition by separating the training into two distinct phases. In the condensation phase ($t < 3000$), a high learning rate injects multiplicative noise to satisfy $\mu \leq \zeta^2/2$. At $t = 3000$, we execute a task shift: we inject a high-frequency target signal (increasing the deterministic curvature μ) while simultaneously dropping the learning rate (decreasing the noise amplitude ζ^2): the target gains $0.8 \cos 5x$, the learning rate falls to 0.01 and the label-noise standard deviation to 0.1. This tests both forward condensation and reverse reactive fragmentation.

Results: During the initial high learning rate phase, the aggressive stochastic diffusion neutralizes the deterministic gradient drift (Figure 3, Middle Row, Panel c). This forces the independent parameters to collapse into shared, lower-dimensional invariant subspaces. The SGD noise shifts the merge thresholds across ensemble realizations, creating a distinct multi-phase precursory cascade. We detect sequential peaks in the smoothed relative variance $\mathcal{R}_v(t)$ (Figure 3, Middle Row, Panel b) strictly within the high learning rate regime. Extracting these transition indices and computing their scaling progression (Figure 3, Bottom Right) yields the two-peak ratio $\boldsymbol{\lambda_t \approx 2.00}$. Two facts qualify this number. With a tube radius of $\epsilon = 0.25$ on weights initialized with standard deviation 0.8, about a third of the pairs lie inside the tube at initialization, so the component count starts near 4 and the first detected peak partly records links that were present at initialization. And for six units the constant kernel predicts a ratio of 2.5 for the first two merges, with 2 the asymptotic value at large N (Section E, prediction paragraph), so the two peaks of this run measure the asymptotic ratio only approximately. The large-N test of the prediction is the cascade over many levels, which the constant kernel fixes at $(2i-1)/(i-1)$ level by level.

Following the task shift at $t = 3000$, the drastic reduction in learning rate and the addition of the new target signal immediately violate the trapping threshold. The invariant manifold becomes stochastically repulsive. The macroscopic topology registers a reverse transition: the relative variance $\mathcal{R}_v(t)$ peaks as the previously bound parameters fragment to map the new target dimensionality.

Separating the two interventions. Past the threshold the escape rate from the manifold scales as $\eta(\mu - \zeta^2/2)$, so the curvature and the noise act on different clocks. We repeat the protocol over 40 realizations with common random numbers in four arms (Table 1). Raising the curvature at the original learning rate fragments 30% of seeds and raises $\mathcal{R}_v$ 1.61-fold within a few hundred steps. Lowering the noise through the learning rate fragments seeds in the same direction on a 20-fold slower clock, and the combined protocol, which runs on that clock, is still building its response at the end of the 2000-step window.

E.5 Block Merges Require Nested Symmetry

Proposition 4.5 separates two ways in which the largest component can grow, namely block merges, in which two already-macroscopic components coincide at once, and single attachment, in which the largest component absorbs one subnetwork at a time. The cascade of Lemma 4.8 is a property of the hitting laws and holds under either mode. What the two modes decide is the sharpness of the

Table 1: Task-shift ablation ($N = 6$, 40 realizations, common random numbers; interventions at $t = 3000$). Fragmenting seeds end with at least half a cluster more, on average over the last 300 steps, than just before the shift.

Arm	Clusters before → after	Fragmenting seeds	Post/pre $\max \mathcal{R}_v$	Maximum at
Target and noise (Fig. 3)	$1.93 \to 2.00$	0.12	1.19	$t = 4995$
Noise only	$1.93 \to 2.08$	0.10	0.87	$t = 3045$
Target only	$1.93 \to 2.11$	0.30	1.61	$t = 4155$
Control	$1.93 \to 1.95$	0.03	1.05	$t = 4155$

individual peaks, since the two-point mixture of Theorem 4.6 needs a macroscopic jump, which block merges produce and single attachment, the Erdős–Rényi regime, does not. We test which of the two SGD produces, and what the architecture has to supply for block merges to occur, with two networks built from the same units, the same target, and the same optimizer.

Construction. Both networks are built from units $a\,\mathrm{gelu}(bx + c)$ with scalar parameters, trained on the target $\mu\,\mathrm{gelu}(x)$ with $\mu = 1$ on 256 standard normal inputs, by SGD with learning rate $\eta = 0.1$, weight decay $\lambda = 10^{-2}$, batch size 8, for 6000 SGD steps and 32 seeds. The weight decay supplies the inward drift of Definition 2.3 along the directions in which the gradient is conserved between coincident units. The *flat* layer sums N such units and carries only the pairwise permutation symmetry of Proposition C.18. The *binary tree* of depth d has $N = 2^d$ leaf units, every internal node computes $\mathrm{gelu}(u_L h_L + u_R h_R + c)$ from its two children, and the root is linear. Swapping any two sibling subtrees, together with all of their parameters, is an exact symmetry at every level of the tree, so the coincidence set of two siblings at level k is an affine subspace of the form of Theorem C.2, and a merge at level k joins two identical blocks of 2^{k-1} leaves. Two leaves are linked when the parameters along their paths to the root, ϕ_i and ϕ_j, satisfy $\frac{1}{2}\|\phi_i - \phi_j\|_2^2 < \epsilon$ with $\epsilon = 10^{-2}$, so that a link between leaves in different subtrees requires the subtrees and their ancestors to coincide, as a block merge requires. Both graphs start with at most one or two links out of $\binom{N}{2}$, with initial mean cluster counts 15.9, 15.5, 31.9 and 30.3 for tree and flat at $N = 16$ and $N = 32$. We record the largest component of the ϵ-graph at every step and report the growth of its running maximum per seed, so that each level is counted once and crossings of the tube boundary are not counted as repeated merges.

Results. Figure 1 and Table 2 summarize the two architectures at $N = 16$ and $N = 32$. Only the tree reaches full collapse, in 47% of seeds at $N = 32$ and 19% at $N = 16$, against 0% for the flat layer at either size, whose cluster count plateaus near 2.3. The tree produces merges that the flat layer does not. At $N = 32$ its record growth contains three jumps of $+16$, the coincidence of two half-trees, and twelve of $+8$, the coincidence of two quarter-trees. The flat layer at $N = 32$ shows a handful of jumps between $+8$ and $+12$ (nine events in 498), which are attachments of a group of units that have already collapsed to a common point rather than merges of symmetric blocks, and none reaches $+16$. The share of record-growth events that exactly double the largest component, the signature of a merge of two equal blocks, is 0.22 and 0.11 for the tree at $N = 16$ and $N = 32$, against 0.11 and 0.03 for the flat layer. In both architectures the majority of record-growth events are single attachments (0.60 to 0.74), because a subtree need not finish collapsing internally before its ancestors align with those of its sibling, so block merges occur in the tree but do not dominate its growth. Nested symmetry forces every merge across subtrees to be a block merge, since leaves in different subtrees link only when their shared ancestors coincide; within a subtree, leaves can still attach one at a time. Nested symmetry therefore makes macroscopic block merges possible, and by Proposition 4.5 the jump survives only when the merging components are already macroscopic, which the architecture permits and the dynamics realize in part. At the first level the ensemble is two-point (all seeds at $|C_{\max}| \in \{1, 2\}$ at $N = 32$), and the $\mathcal{R}_v$ peak of $|C_{\max}|/N$ has height 0.125 (95% bootstrap interval $[0.124, 0.169]$) at merged weight $w = 0.34$, at the w^*-quantile of the hitting law, as Theorem 4.6 predicts. At higher levels the ensemble occupies more than two states, reflecting the single attachments, and the peaks are correspondingly taller (0.21 to 0.37).

Clock. Panel (c) reports the per-seed hitting times $\mathbb{E}[T_{2^k}]$ of $|C_{\max}| \geq 2^k$ by level. For the tree at $N = 32$ they are 748, 1659, 1922, 2200 and 2701 steps, with consecutive ratios 2.2, 1.16, 1.14 and 1.23, against 3, 2.33, 2.14 and 2.07 for a constant kernel under a time-homogeneous clock (Lemma 4.8). The first level is slow because the leaves must converge from a random

initialization, and every higher level is paced by the same quantity, the contraction $e^{-\eta\lambda t}$ of the conserved sibling differences under weight decay, so that successive levels close at a nearly constant spacing $\Delta t \approx \ln(\Delta_0/\sqrt{2\epsilon})/(\eta\lambda)$ rather than at geometrically growing intervals. The level structure is the pairwise hierarchy of Definition C.12, and in the balanced hierarchy the cascade is geometric with factor 2 in the density variable, since $1-p$ halves at each level. The clock, however, is exponential rather than the power law of a time-homogeneous kernel, so the ratios in t reflect the clock rather than the multiplicity. A kernel exponent fitted to these ratios through $\gamma_K = 1 - \log_2 \lambda_t$ would read 0.74, attributing to component size a variation that comes from training time. This is the time-inhomogeneity that Remark F.2 names, realized in a setting where the level structure and the clock can both be read off, and it places the weight-decay-paced toy outside the time-homogeneous hypothesis of Theorem 4.9 while confirming its level structure.

Table 2: Growth of the record largest component under SGD for the two architectures (32 seeds, 6000 SGD steps). "Exact doubling" is the share of record-growth events with $|C_{\max}|_{\text{after}} = 2\,|C_{\max}|_{\text{before}}$, and "macroscopic" is the share with jump $\geq N/4$. Hitting ratios omit the first level for the flat layer at $N = 32$, where one pair is inside the tube at initialization ($T_2 = 18$).

	Tree $N = 16$	**Flat** $N = 16$	**Tree** $N = 32$	**Flat** $N = 32$
Full collapse (share of seeds)	19%	0%	47%	0%
Final mean cluster count	3.6	2.3	1.7	2.3
Single-attachment share	0.74	0.68	0.65	0.60
Exact-doubling share	0.22	0.11	0.11	0.03
Macroscopic jumps ($\geq N/4$) share	0.09	0.05	0.05	0.02
Largest jump observed	+4	+5	+16	+12
Hitting ratios $\mathbb{E}[T_{2i}]/\mathbb{E}[T_i]$	1.85, 1.22	3.8, 1.5	2.2, 1.16, 1.14, 1.23	1.7, 1.2

F Extended Empirical Validation Across Diverse Datasets and Architectures

Having recovered the pairwise cascade in controlled kinematic and low-dimensional SGD environments, we now generalize the framework to standard deep learning paradigms. We deploy unconstrained neural networks across a diverse suite of datasets, spanning tabular data, geometric manifolds, vision tasks, and algorithmic grokking.

F.1 Methodological Extensions for Empirical Regimes

Transitioning from isolated theoretical toy models to highly non-convex, unconstrained deep learning environments requires adapting our measurement observables. All empirical experiments were run on a Tesla T4 GPU and require about one day of compute. Architectures and optimization settings are listed in Table 3.

Soft Dimensionality Collapse (Spectral Effective Rank): In high-dimensional empirical networks, parameter constraints rarely force weights to absolute zero (hard topological collapse). Instead, structural merges manifest as spectral concentration, where the energy of the weight matrix collapses into a smaller subset of principal components. We map the macroscopic order parameter $\langle \mathcal{O}(t) \rangle$ to the Effective Rank of the target layer's weight matrix. Computing the Singular Value Decomposition (SVD), we extract the singular values σ_k, normalize them into a probability distribution $p_k = \sigma_k / \sum \sigma_i$, and compute the exponential of the Shannon entropy:

$$\mathcal{O}(t) = \exp\left(-\sum p_k \ln p_k\right). \tag{42}$$

This continuous metric reliably tracks the macroscopic dimensionality collapse as the top singular values diverge and absorb the network's capacity.

Signal Processing, Detrended Log-Variance, and Ablation Controls: Real-world SGD exhibits massive baseline heteroskedasticity; the raw variance is known to decay as the network settles into a basin (Mandt et al., 2017; Mori et al., 2022). To isolate the microscopic variance peaks ($\mathcal{R}_v(t)$) from this global decay, we operate strictly in log-variance space, $\log(\mathrm{Var}[\mathcal{O}(t)])$. We subsequently subtract

Table 3: Settings of the empirical suite. Categorical features are one-hot encoded (first level dropped) and tabular features standardized; models are trained on the full dataset, since the observable is the training dynamics. UCI runs use 20 initializations and the other tasks 15. UCI peak detection uses macro $\sigma = 20$, local $\sigma = 1.5$ and minimum peak distance 8; the other tasks use macro $\sigma = E/5$, local $\sigma = E/50 + 1$ and distance $E/20$ for E epochs; the prominence is 0.05 throughout.

Task	Model	Monitored matrix	Optimization	Data
Heart Disease, German Credit, Abalone, Digits	$d_{\text{in}} \to 64 \to 32 \to C$, GELU	first layer ($64 \times d_{\text{in}}$)	SGD, lr 0.05, batch 32, 100 epochs	OpenML `heart-statlog`, `credit-g`, `abalone` (rings ≤ 8, 9–10, ≥ 11); scikit-learn digits
Moons, Swiss Roll	$d_{\text{in}} \to 64 \to 32 \to 2$, GELU	first layer ($64 \times d_{\text{in}}$)	SGD, lr 0.1, batch 64, 100 epochs	2000 samples, noise 0.1
MNIST, FashionMNIST	$784 \to 128 \to 64 \to 10$, GELU	first layer (128×784)	SGD, lr 0.05, batch 128, 50 epochs	5000-image subset
Modular arithmetic	1 layer, $d = 128$, 4 heads, FF 512, GELU, dropout 0.1, no LayerNorm	token embedding (18×128)	AdamW, lr 3×10^{-3}, $\lambda = 0.5$, batch 128, 30000 epochs	$(a + b) \bmod 17$, input $[a, b, =]$, 70/30 split (202/87)

a global linear secant to detrend the signal, preventing artificial boundary artifacts during macroscopic smoothing. This isolates the structural fluctuations from the standard optimization noise. To establish a baseline, we construct a spectral null model testing the null hypothesis that observed cascades are artifacts of this optimization noise. We compute 1000 phase-randomized spectral surrogates which preserve the empirical power spectrum of the variance trajectory while destroying localized temporal structures to calculate empirical False Positive Rates (FPR), the fraction of surrogates whose peak sequence fits a log-linear law at least as well as the data; with 1000 surrogates the smallest attainable value is $1/1001$. Additionally, we compute 500 bootstrap resamples per configuration to ensure statistical stability, aggregating variance across 20 independent initializations for the UCI tabular suite and 15 independent initializations for high-dimensional vision and algorithmic grokking datasets. A log-linear fit through k peaks has $k - 2$ residual degrees of freedom, so for three peaks the fit has one and the FPR is the primary evidence for the cascade. We report R^2 for the longer cascades only.

Graded evidence: We report the FPR as a continuous measure of evidence rather than applying a fixed significance level, following the ASA statements on p-values (Wasserstein & Lazar, 2016; Wasserstein et al., 2019). A p-value does not measure the size or importance of an effect, and conclusions should not rest on whether it crosses a specific threshold. Heart Disease (4.9%) and FashionMNIST (10.7%) therefore differ by a factor of two in how often spectrally matched noise reproduces the fit, not by a qualitative change in evidence, and the negative control German Credit (80.2%) marks the other end of the scale.

Spectral Amplitude of Block Merges: In the idealized $N = 3$ kinematic model (Appendix E), exact counting gives an integer jump in $|C_{\max}|$. Mapping unconstrained networks through the continuous Effective Rank changes the amplitude of each jump and leaves its location alone: the cascade factor λ is a property of the hitting laws (Theorem C.23), the measured ratio λ_t a property of the time map (Corollary C.24), and neither depends on the observable used to detect the peaks, since two monotone observables of the same process share every transition time. What the observable sets is the size of the step. The effective rank therefore serves as a detector of transition times; the peak height of Theorem C.17 is a property of $|C_{\max}|/N$, and Appendix E.5 tests it there.

Proposition F.1 (Spectral Step of a Coincidence Merge)**.** *Let the order parameter be the Spectral Effective Rank $\mathcal{O} = \exp(H)$ with $H = -\sum_k p_k \ln p_k$ and $p_k = \sigma_k / \sum_j \sigma_j$. Let the weight matrix have N mutually orthogonal columns of equal norm, and let n of them coincide. Before the merge $\mathcal{O}_{pre} = N$. After it the singular values are $N - n$ of the original size and one of $\sqrt{n}$ times that size,*

so with $S := N - n + \sqrt{n}$

$$\mathcal{O}_{post} = \exp\Big(\frac{(N-n)\ln S + \sqrt{n}\,\ln(S/\sqrt{n})}{S}\Big), \qquad \frac{\mathcal{O}_{pre}}{\mathcal{O}_{post}} \neq n^{n/N}.$$

With the squared normalization $p_k = \sigma_k^2 / \sum_j \sigma_j^2$ *the spectral mass is conserved through the merge and the step is exactly* $\mathcal{O}_{pre}/\mathcal{O}_{post} = n^{n/N}$.

Proof. Orthogonal columns of equal norm $\|c\|$ give N singular values equal to $\|c\|$ and $p_k = 1/N$, so $\mathcal{O}_{pre} = N$. When n columns become identical, their span has dimension one and the single nonzero singular value on it is $\sqrt{n}\|c\|$, while the remaining $N-n$ singular values are unchanged. The σ-mass $\sum_k \sigma_k$ falls from $N\|c\|$ to $S\|c\|$, so it is not conserved, and evaluating H at the post-merge spectrum gives the displayed formula. Under the squared normalization the masses are σ_k^2, which sum to $N\|c\|^2$ before and after, the merged component carries $p = n/N$ and each other component $1/N$, and $H_{pre} - H_{post} = (n/N)\ln n$, which is the stated step. For $N = 16$ and $n = 2$ the two normalizations give steps of 1.072 and 1.091. Discrete counting registers a component reduction by the factor n, the spectral observable registers the step above, and the transition times are those of the underlying process and do not move, so n is read from the height of the effective-rank step given N, and the peak spacing is unaffected by the choice of observable. □

Remark F.2 (Reading the Measured Ratios). Corollary C.24 gives $\lambda_t = n^{1-\gamma_K}$, so the measured ratio constrains n and the kernel exponent γ_K jointly, and $\gamma_K = 1 - \log_n \lambda_t$ once n is fixed; the fitted γ_K is therefore descriptive rather than a test of the kernel. It treats training time as the coalescence clock, so a link rate that varies with training time enters it alongside the size dependence of the kernel. For λ_t near 1 the corrections λ_t^{-k} of Lemma 4.8 are large at the levels observed, about 0.3 at $k = 5$ for $\lambda_t = 1.28$, so those ratios carry finite-level corrections of that size.

- **Modular Arithmetic (Transformer Grokking):** $\lambda_t = 2.11$ is pairwise ($n = 2$) with $\gamma_K = -0.08$, a constant kernel within the resolution of three peaks, or three-body ($n = 3$) with $\gamma_K = 0.32$. The ratio alone does not separate the two.
- **UCI Abalone:** $\lambda_t = 1.28$, which for pairwise merges gives $\gamma_K = 0.64$.
- **UCI Heart Disease:** $\lambda_t = 1.71$, which for pairwise merges gives $\gamma_K = 0.23$.
- **FashionMNIST (Vision Benchmark):** $\lambda_t = 1.57$, which for pairwise merges gives $\gamma_K = 0.35$.
- **Conjecture on Network and Task Complexity:** We cautiously hypothesize that highly deep neural models trained on complex representations form distinct macroscopic sub-network symmetries, yielding clean log-linear cascades. Conversely, shallower networks on low-dimensional tabular data generate variance fluctuations frequently indistinguishable from baseline stochastic gradient noise.

F.2 UCI Tabular Datasets

We evaluate shallow Multi-Layer Perceptrons (MLPs) optimized via SGD on a suite of standard tabular datasets from the UCI Machine Learning Repository. Across all datasets, the macroscopic order parameter (Effective Rank) and the training loss exhibit smooth, continuous decay. However, the detrended variance breaks this continuity, revealing discrete microtransitions. Table 4 aggregates the empirical peak-time fits and spectral null-model evaluations across the suite.

Heart Disease gives the strongest tabular evidence (FPR $= 4.9\%$). Conversely, German Credit serves as a critical negative control. Although the network exhibits a 4-peak variance sequence ($\lambda_t = 1.99, R^2 = 0.86$), spectral null testing yields a False Positive Rate of 80.2%. This negative result establishes the necessity of the signal processing pipeline; without it, standard optimization noise mimics geometric scaling artifacts.

Table 4: Empirical peak-time cascades and statistical null-model evaluations for UCI tabular datasets (aggregated over 20 independent initializations). A cascade fit needs at least three peaks, so none is reported for Digits.

Dataset	**Peaks**	λ_t	R^2	**FPR (%)**
Heart Disease	4	1.71	0.98	4.9%
Abalone	6	1.28	0.97	15.6%
German Credit	4	1.99	0.86	80.2%
Digits	1	N/A	N/A	N/A

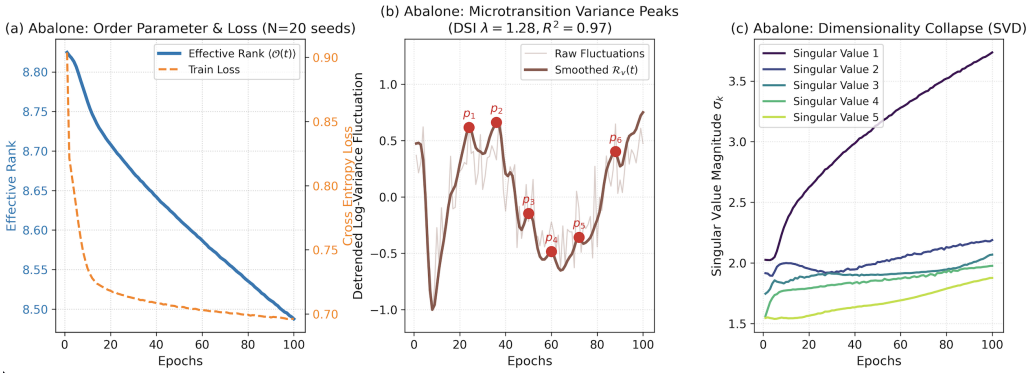


Figure 4: **Abalone:** The macroscopic continuous decay of effective rank (a) is underscored by a structured 6-peak cascade yielding a log-linear fit of $R^2 = 0.97$ with a fractional ratio $\lambda_t = 1.28$ (b). The collapse is driven by a dominant top singular value (c).

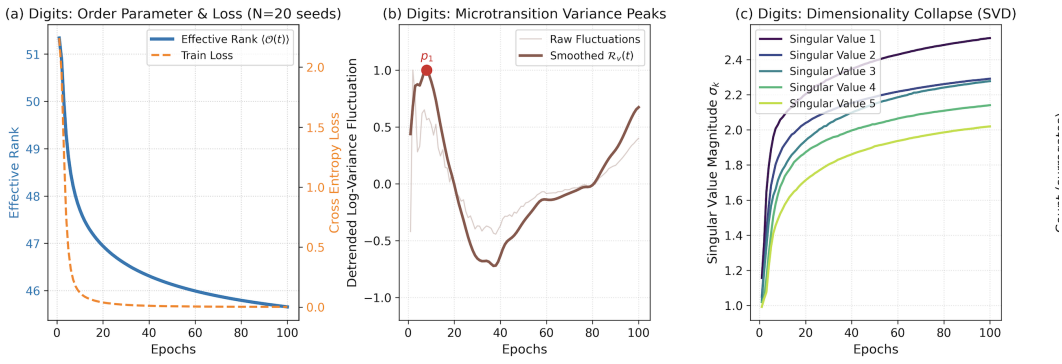


Figure 5: **Digits:** Variance fluctuations map the topological condensation of the network during early-stage SGD optimization.

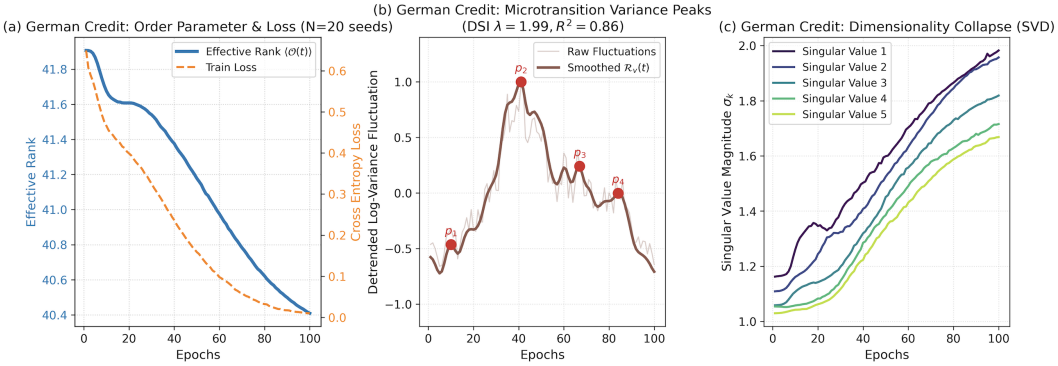


Figure 6: **German Credit:** The network exhibits a 4-peak cascade ($\lambda_t = 1.99, R^2 = 0.86$), but yields a spectral null False Positive Rate of 80.2%, acting as a negative control.

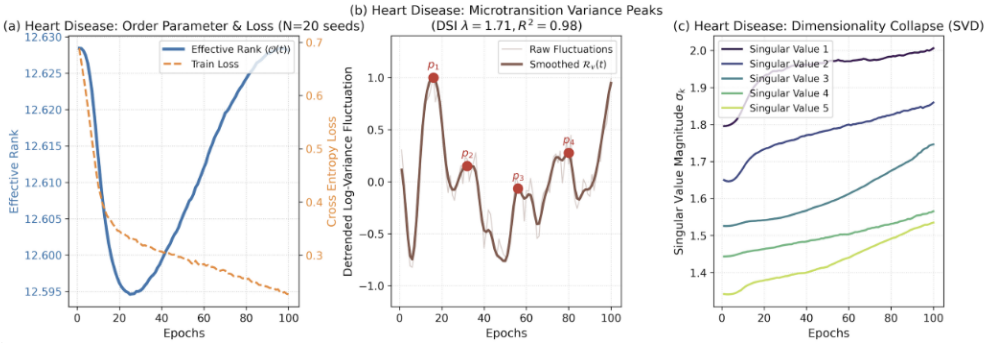


Figure 7: **Heart Disease:** Dimensionality collapse (c) is mirrored by a sequential 4-peak variance cascade ($\lambda_t = 1.71, R^2 = 0.98, \text{FPR} = 4.9\%$) mapping the internal topological constraints.

F.3 Geometric Manifolds, Vision, and Algorithmic Grokking

To verify that these scaling laws are not strictly an artifact of tabular data or shallow MLPs, we extend the empirical validation to geometric manifolds (Moons, Swiss Roll), vision benchmarks (MNIST, FashionMNIST), and algorithmic tasks (Modular Arithmetic).

Critically, the Modular Arithmetic task utilizes a Transformer architecture optimized via AdamW with explicit weight decay. This algorithmic setting inherently induces grokking, a dynamic phenomenon wherein deep neural networks initially overfit the training distribution before abruptly acquiring perfect generalization after extensive continued optimization (Power et al., 2022). Mechanistic analyses reveal that this delayed performance spike masks a representational phase transition; the network progressively dismantles high-rank, dense memorization circuits to construct low-dimensional, structured algorithmic manifolds Nanda et al. (2023); Liu et al. (2022). Driven by explicit regularization, specifically the weight decay utilized in our AdamW optimizer, and inherent stochastic gradient noise, the optimization trajectory continuously penalizes complex parameterizations, forcing the weights to systematically escape sharp memorizing minima Thilak et al. (2022). Weight decay enters condition equation 37 of Theorem D.26 through the inward term $-\eta\lambda\boldsymbol{x}$, and Remark D.30 reports the measurements. The transverse drift toward permutation coincidence of feed-forward units is inward at 63 to 74% of checkpoints across three realizations, the trapping condition holds at 50 to 62%, and Y contracts by up to 3.6 decades between slingshot instabilities, which reset the condensation phase about once per $1.4\,\tau_2$. Consequently, we expect this prolonged macroscopic dimensionality reduction to happen through a sequence of discrete topological collapses. The Transformer's subsequent delayed generalization is consistent with this hypothesis.

Across 15 independent initializations, the final train loss converges to 0.005 ± 0.004, validation loss drops to 0.030 ± 0.033, and effective rank collapses to 10.93 ± 0.99. The detrended variance isolates a log-linear 3-peak cascade with $\lambda_t = 2.11$. With three peaks the fit has one degree of freedom, so the evidence for the cascade is the spectral null, which gives a False Positive Rate of at most 0.1% (no surrogate matched the data). We test on addition modulo 17 for 30000 epochs with AdamW weight-decay coefficient $\lambda = 0.5$ and learning rate 0.003; the three peaks lie at epochs 4500, 9500 and 20000. Fourier pruning of the embedding, the progress measure of Nanda et al. (2023), completes by epoch 2850 in three probe seeds, so the cascade runs after it at fixed Fourier support. In the same seeds the embedding columns coalesce at a tube radius of about half their norm in halving levels, the first reached at epochs 3960 to 5470, while coincidence of the units on which the permutation symmetry acts is measured on the feed-forward block (Remark D.30).

The high-dimensional vision benchmarks mirror this discretized capacity reduction. FashionMNIST exhibits consecutive variance peaks highlighting structural scaling, yielding a 3-peak cascade with $\lambda_t = 1.57$ and a spectral null FPR of 10.7%. Conversely, MNIST produces a 2-peak sequence (p_1, p_2) mapping early-stage topological condensation without forming a full extended cascade.

Evaluations on geometric manifolds map the collapse phase as the MLP parameterizes spatial embeddings. Moons isolates a single primary microtransition (p_1) navigating the non-linear 2D geometry. Swiss Roll similarly identifies a primary structural microtransition (p_1) charting the 3D spatial embedding. These initial peaks capture the epoch of spatial alignment consistent with discrete topological condensation during early-stage optimization.

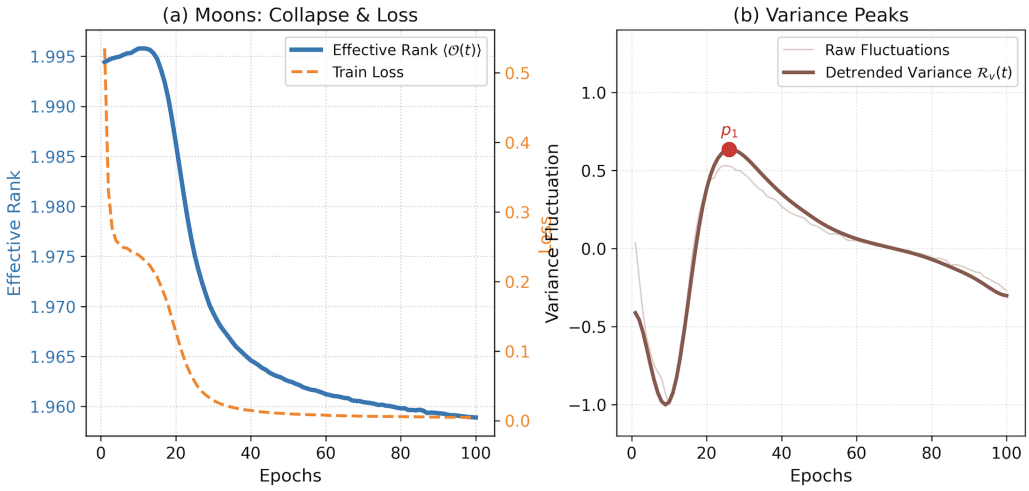


Figure 8: **Moons (Geometric Manifold):** Collapse phase mapped by variance fluctuations as the MLP parameterizes the 2D non-linear manifold.

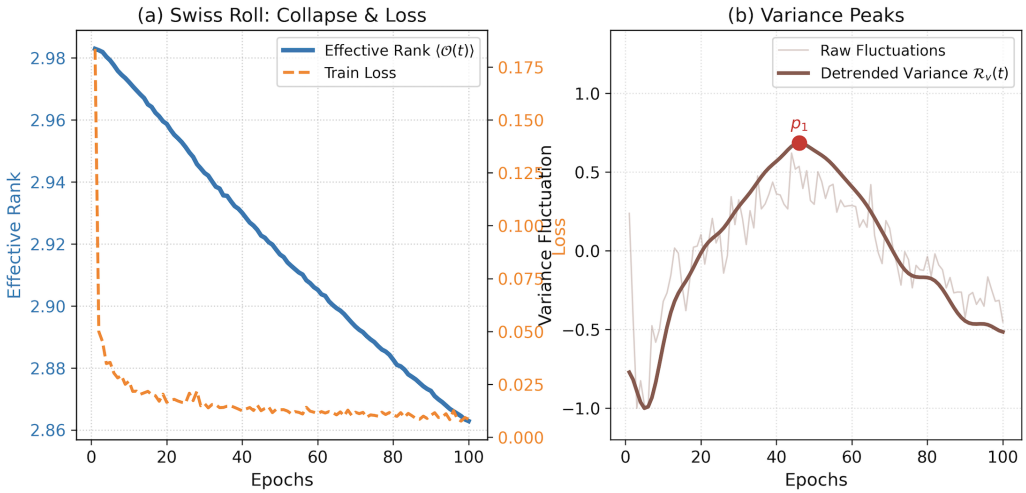


Figure 9: **Swiss Roll (Geometric Manifold):** Sequential structural microtransitions navigate the 3D spatial embedding of the data.

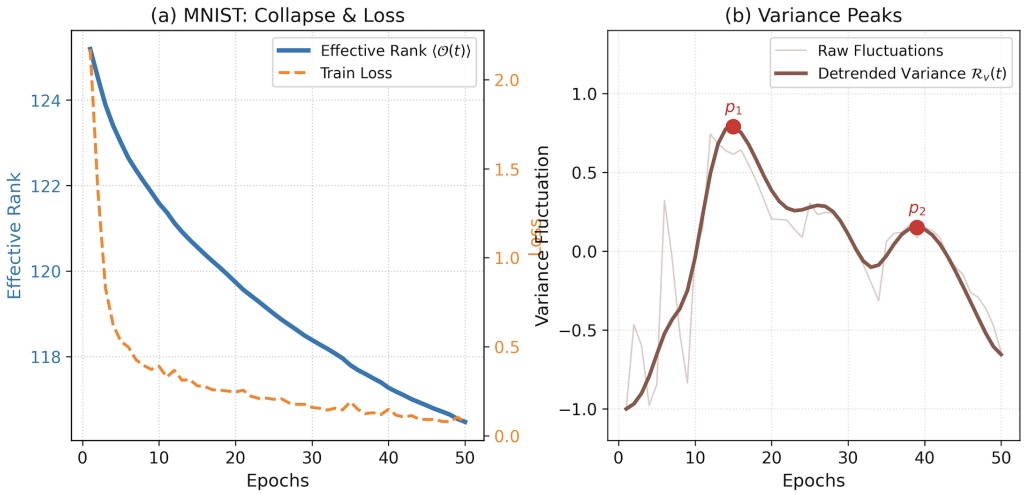


Figure 10: **MNIST (Vision):** Variance fluctuations produce a 2-peak sequence (p_1, p_2) mapping the topological condensation of the network during early-stage SGD optimization.

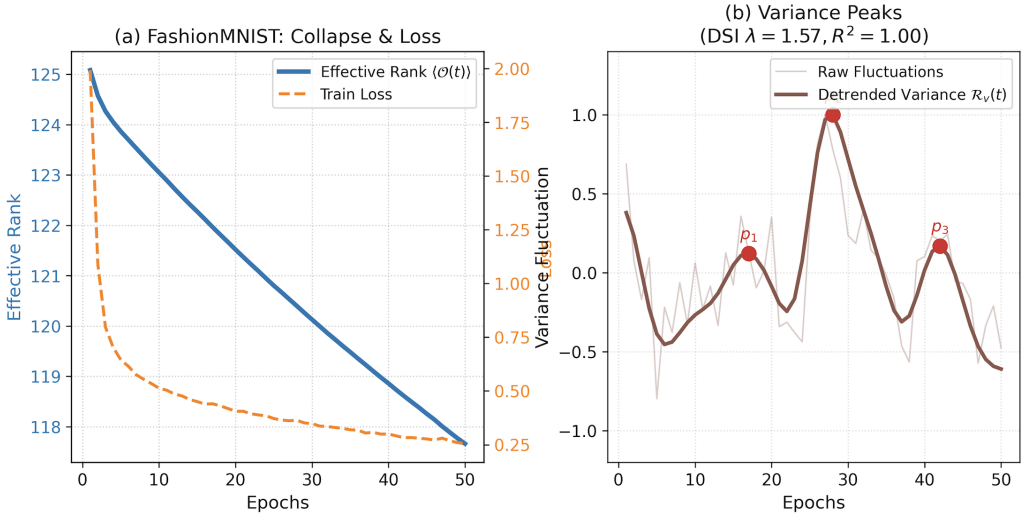


Figure 11: **FashionMNIST (Vision):** Detection of consecutive variance peaks highlighting the discretized nature of capacity reduction ($\lambda_t = 1.57, \text{FPR} = 10.7\%$).

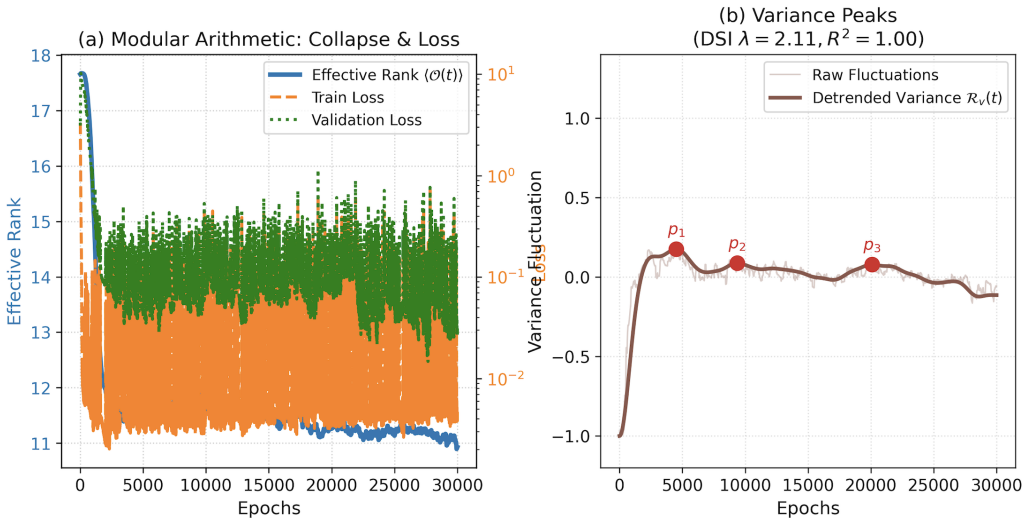


Figure 12: **Modular Arithmetic (Grokking):** A shallow Transformer trained with AdamW experiences a delayed topological collapse. The macroscopic decay of the token-embedding table's effective rank is composed of a log-linear 3-peak cascade ($\lambda_t = 2.11, \text{FPR} = 0.1\%$).