FROM MEMORIZATION TO LENGTH GENERALIZATION: GROKKING IN RECURRENT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent models can in principle process arbitrarily long sequences by reusing the same state transition, but fitting short sequences does not guarantee that the learned transition will generalize to longer ones. We study this gap in an over-parameterized linear RNN trained on sequential modular addition. We show that zero training loss alone is insufficient for length generalization because decoder-invisible hidden directions can remain unconstrained. In contrast, sufficient hidden-state coverage forces every zero-loss solution to implement the correct reusable recurrence. Combined with a small-initialization convergence result, this gives a tractable setting where exact arbitrary-length generalization can be characterized.We then explain why length generalization may emerge much later than short-horizon fitting. Near a generalizing solution, the delay is controlled by slowly decaying training modes and by their visibility at different rollout horizons. For decoder-invisible directions, we derive the local curvature and link the slowest learning rate to the conditioning of hidden-state coverage. This provides a quantitative mechanism for *length grokking*: recurrent errors can become negligible on short rollouts while remaining significant on longer ones. Our results connect hidden-state coverage, optimization dynamics, and the emergence of reusable recurrent computation from finite-horizon data.

1 INTRODUCTION

Recurrent neural networks maintain a hidden state and repeatedly update it using a shared transition rule as new inputs arrive (Elman, 1990; Panigrahi & Goyal, 2021). This structure makes recurrent models natural candidates for solving problems that require state tracking or repeated algorithmic computation.(Delétang et al., 2023; Buitrago & Gu, 2025; Van Rossem & Saxe, 2025). Unlike architectures whose computation is tied to a fixed input length, a recurrent model can in principle continue applying the same learned transition for arbitrarily many steps. This makes recurrence an appealing architectural mechanism for *length generalization*: once a transition rule has been learned, the same rule can in principle be reused for arbitrarily many steps.

The ability to *run* a recurrence for more steps, however, should not be confused with the ability to *learn the correct recurrence* from finite-horizon data. A recurrent architecture may be expressive enough to represent the desired transition rule, while optimization from short trajectories identifies only its behavior on the hidden states encountered during training. Two recurrent systems can therefore agree on all observed short trajectories yet differ on hidden states reached after further composition. Consequently, perfect short-horizon fitting may not guarantee that the learned transition can be safely reused beyond the training horizon.

This distinction is visible empirically. Recurrent and state-space models can perform well within their training context while degrading substantially on longer sequences (Anil et al., 2022; Delétang et al., 2023; Fan et al., 2024; Ben-Kish et al., 2025; Azizi et al., 2025; Buitrago & Gu, 2025; Lu et al., 2025). The issue is particularly relevant for state tracking, algorithmic sequence processing, and recurrent long-context modeling, where inaccuracies in the learned state transition are repeatedly composed over long rollouts (Delétang et al., 2023; Buitrago & Gu, 2025; Azizi et al., 2025; Lu et al., 2025). Thus the central question is not simply whether a recurrent architecture *can* represent a length-generalizing computation, but under what training conditions finite-horizon observations are sufficient to *identify and learn* a transition rule that remains correct under repeated composition.

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.

Prior work has studied length generalization from several directions. Theory on recurrent models has characterized their ability to represent and learn sequential computations, but largely at bounded sequence lengths (Panigrahi & Goyal, 2021). Work on algorithmic and formal-language tasks has shown that strong training performance need not extrapolate to longer inputs, and that the recurrent architecture itself can strongly affect length generalization (Delétang et al., 2023; Fan et al., 2024). Similar failures have recently been observed in modern recurrent and state-space models (Ben-Kish et al., 2025; Azizi et al., 2025; Lu et al., 2025).

Two recent lines of work are especially relevant to our question. Buitrago & Gu (2025) show that poor coverage of recurrent states can cause length-generalization failure, and improve it by exposing the model to a broader set of hidden states. Separately, work on grokking shows that fitting and generalization can occur on very different optimization timescales (Power et al., 2022; Gromov, 2023; Nanda et al., 2023). For recurrent models, Van Rossem & Saxe (2025) observe a related transition when learning streaming parity: fitting observed trajectories precedes the emergence of the state structure needed for infinite generalization.

Taken together, these results leave a central question open: when does finite-horizon training recover a recurrent rule that remains correct at arbitrary lengths, and why can long-horizon generalization emerge much later than short-horizon fitting? We study both questions in a setting where the endpoint of learning and the optimization dynamics can be analyzed explicitly. More specifically, we connect these two questions through the geometry of hidden-state coverage. Weakly covered hidden directions are learned slowly, and the resulting errors can be nearly invisible on short rollouts while remaining important at longer horizons.

Our contributions. Our main contribution is a tractable framework for studying length generalization in recurrent models as a learning-dynamics problem.

1. **A solvable framework for recurrent length generalization.** We introduce an overparameterized linear RNN that preserves the central difficulty of length generalization while remaining analytically tractable. The model makes precise the distinction between fitting finite trajectories and learning a reusable recurrent rule: ordinary training can reach zero loss without length generalization, whereas sufficient hidden-state coverage forces every zero-loss solution to implement the correct decoded recurrence. Together with our small-initialization convergence result, this gives an end-to-end setting in which we can characterize when finite-horizon training leads to exact arbitrary-length generalization.
2. **A theory of length grokking.** We show that short-horizon fitting and length generalization can occur on widely separated optimization timescales even when the correct recurrent rule is ultimately learned. We derive a spectral mechanism for this delay: weakly trained recurrent modes decay slowly and can be nearly invisible on short rollouts while remaining important at long horizons. In our model, these modes arise from weakly covered hidden directions, giving a quantitative link between hidden-state geometry and the timescale of length generalization.

2 Setup of Framework

In this section, we introduce the modular addition task and the overparameterized linear RNN studied throughout the paper. In Section 2.1, we formulate modular addition as a recurrent state-transition problem. In Section 2.2, we introduce the linear RNN architecture and specify the training setup for its recurrent dynamics.

2.1 Modular Addition

Modular arithmetic has become a standard setting for studying the transition from memorization to generalization because it generates a finite, exactly specified dataset while still exhibiting nontrivial learning dynamics, including grokking (Power et al., 2022; Gromov, 2023; Nanda et al., 2023). For recurrent models considered in this paper, modular arithmetic has an additional advantage: the computation can be expressed as repeated updates of a finite task state. Longer instances therefore require no new operation; they only require the same transition rule to be composed more times.

This allows us to distinguish a model that fits the transitions observed during training from one that has learned a transition rule that remains correct under repeated composition.

We use modular addition over $\mathbb{Z}_d = \{0, 1, \ldots, d-1\}$ as a sequential state-tracking task, where $d \geq 2$ is the number of possible task states. For $x, a \in \mathbb{Z}_d$, define $x \oplus a := (x + a) \bmod d$. Given an initial state x_0 and actions $a_1, \ldots, a_L$, the state evolves as $x_t = x_{t-1} \oplus a_t$, so that $x_L = (x_0 + \sum_{t=1}^{L} a_t) \bmod d$.

We represent each state $x \in \mathbb{Z}_d$ by its one-hot vector $e_x \in \mathbb{R}^d$. For a fixed action a, the map $x \mapsto x \oplus a$ simply permutes the d possible states. It can therefore be represented by a permutation matrix $P_a \in \mathbb{R}^{d \times d}$, i.e., a matrix that permutes the standard basis vectors:

$$P_a e_x = e_{x \oplus a}, \qquad x, a \in \mathbb{Z}_d. \tag{1}$$

Writing $y_t = e_{x_t}$, the task dynamics become $y_t = P_{a_t} y_{t-1}$, and hence $y_L = P_{a_L} \cdots P_{a_1} e_{x_0}$. These transition matrices inherit the algebraic structure of modular addition: $P_b P_a = P_{a \oplus b}$.

2.2 Overparameterized Linear RNN

We study a linear recurrent neural network for the modular addition task. This provides us a theoretically tractable model that focuses on the crucial recurrence, hidden-state overparameterization dynamics and repeated composition. Despite the simple form, it preserves the nontrivial temporal behavior, and long-term dynamics (Stolzenburg et al., 2018).

Formally, let h denote the hidden dimension, with $h > d$ being the overparameterized setting. Given an initial task state $x_0 \in \mathbb{Z}_d$ and an action sequence $(a_1, \ldots, a_L)$, the model maintains hidden states $z_0, z_1, \ldots, z_L \in \mathbb{R}^h$ and produces decoded task states $\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_L \in \mathbb{R}^d$. The model is parameterized by an encoder $E \in \mathbb{R}^{h \times d}$, a decoder $O \in \mathbb{R}^{d \times h}$, and a collection of recurrent matrices $\{R_a \in \mathbb{R}^{h \times h} : a \in \mathbb{Z}_d\}$. The encoder maps the one-hot representation of the initial task state into the hidden space. At each step, the current action selects the corresponding recurrent matrix, which updates the hidden state linearly:

$$z_0 = E\mathbf{e}_{x_0}, \qquad z_t = R_{a_t} z_{t-1}, \qquad \hat{y}_t = O z_t, \quad t = 1, \ldots, L. \tag{2}$$

Training setup. Our goal is to learn recurrent transitions that remain correct under composition beyond the training horizon. For an initial task state $x_0 \in \mathbb{Z}_d$ and an action sequence $a_{1:T}$, exact length generalization requires

$$O R_{a_T} \cdots R_{a_1} E e_{x_0} = P_{a_T} \cdots P_{a_1} e_{x_0}, \qquad \forall T \geq 1. \tag{3}$$

We separate the encoder–decoder representation from the learning of the recurrent dynamics. We fix an encoder $E \in \mathbb{R}^{h \times d}$ and decoder $O \in \mathbb{R}^{d \times h}$ satisfying $OE = I_d$. The encoder and decoder are then frozen, and only the recurrent matrices $\{R_a\}_{a \in \mathbb{Z}_d}$ are optimized.

Recurrent initialization. The recurrent matrices are initialized independently as

$$(R_a(0))_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\varepsilon^2}{h}\right), \qquad a \in \mathbb{Z}_d, \tag{4}$$

where $\varepsilon > 0$ is the recurrent initialization gain. Our convergence theorem will characterize the sufficiently small-ε regime.

Training objective. For a training horizon L, we first consider ordinary training from the reset states Ee_x. We use every-step supervision and define the *reset-state training loss*

$$\mathcal{L}_L^{\text{reset}}(R) = \frac{1}{2Ld^{L+1}} \sum_{x \in \mathbb{Z}_d} \sum_{a_{1:L} \in \mathbb{Z}_d^L} \sum_{t=1}^{L} \| O R_{a_t} \cdots R_{a_1} E e_x - P_{a_t} \cdots P_{a_1} e_x \|_2^2 . \tag{5}$$

3 From Zero-Loss to Length Generalization

We now ask whether minimizing the reset-state training objective in Eq. (5) is sufficient to learn a recurrent rule that remains correct beyond the training horizon. This raises two distinct questions.

First, does optimization converge to a zero-loss solution? Second, does every zero-loss solution implement a reusable recurrence that generalizes to longer sequences?

We address these questions separately. We first show that, under sufficiently small initialization, gradient flow converges to zero training loss. However, zero loss alone does not guarantee length generalization, because training from the reset states constrains the recurrent dynamics only on the hidden states encountered during training. This will motivate a hidden-state augmentation that makes the recurrent rule identifiable.

Small initialization reaches zero loss. We first consider ordinary training from the reset states Ee_x. Throughout this subsection, for the convergence results we assume that the encoder has orthonormal columns and $O = E^\top$. Equivalently, after an orthogonal change of coordinates we may take $E = (I_d, 0)^\top$ and $O = (I_d, 0)$.

Theorem 3.1 (Small-initialization convergence)**.** *For every fixed finite d, h, and $L \geq 1$, there exists $\delta_{\mathrm{reset}} > 0$ such that any initialization satisfying $\left(\sum_{a\in\mathbb{Z}_d} \|R_a(0)\|_F^2\right)^{1/2} < \delta_{\mathrm{reset}}$ converges under gradient flow on $\mathcal{L}_L^{\mathrm{reset}}$ to a finite limit R_∞ with $\mathcal{L}_L^{\mathrm{reset}}(R_\infty) = 0$.*

The proof is in Appendix A.2. Theorem 2.1 resolves the optimization problem near zero initialization, but zero loss alone does not imply length generalization. Ordinary training starts from the reset states Ee_x, which span only $\mathrm{Im}(E)$. When $h > d$, the hidden space decomposes as $\mathbb{R}^h = \mathrm{Im}(E) \oplus \ker(O)$. , which means the directions in $\ker(O)$ are invisible to the current decoder, but they can still affect future outputs after being propagated by the recurrent dynamics.

Decoder-consistent source augmentation. We therefore augment the reset states along $\ker(O)$. Let $k = h-d$, and let $K \in \mathbb{R}^{h\times k}$ have orthonormal columns spanning $\ker(O)$. Since $O(Ee_x+u) = e_x$ for every $u \in \ker(O)$, such perturbations change the hidden state without changing its task-level label.

We use the same Gaussian decoder-invisible directions for every task state. Let $g_1, \ldots, g_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_k)$. For every $x \in \mathbb{Z}_d$ and $j = 1, \ldots, n$, define

$$h_{x,j}^{\pm} = Ee_x \pm \rho K g_j, \tag{6}$$

where $\rho > 0$ controls the augmentation scale. Together with the reset states, **the balanced Gaussian source pool** is

$$\mathcal{S} = \{(Ee_x, e_x) : x \in \mathbb{Z}_d\} \cup \{(Ee_x + \rho K g_j, e_x), (Ee_x - \rho K g_j, e_x)\}_{x,j}. \tag{7}$$

The total number of sources is $N = d(1 + 2n)$. Writing $H = [h_1, \ldots, h_N]$ and $Y = [y_1, \ldots, y_N]$, we have $Y = OH$.

The symmetric pairs provide hidden-space coverage while preserving the symmetry of the reset problem. In the orthogonal coordinates $[E, K]$, write each source as $h_i = (y_i, u_i)^\top$. Then

$$\frac{1}{N}\sum_{i=1}^N y_i y_i^\top = \frac{1}{d} I_d, \qquad \frac{1}{N}\sum_{i=1}^N y_i u_i^\top = 0.$$

Moreover, $h_{x,j}^+ - h_{x,j}^- = 2\rho K g_j$, so the added sources expose the Gaussian directions in $\ker(O)$.

Proposition 3.1 (Full hidden-state coverage)**.** *If $n \geq h - d$, then the source matrix H in Eq. (7) satisfies $\mathrm{rank}(H) = h$ almost surely.*

Proof. See Appendix A.3. □

We train the recurrent matrices from the source pool with the same length-L, every-step supervision:

$$\mathcal{L}_{L,\mathcal{S}}(R) = \frac{1}{2LNd^L}\sum_{i=1}^N \sum_{a_{1:L}\in\mathbb{Z}_d^L} \sum_{t=1}^L \|OR_{a_t}\cdots R_{a_1}h_i - P_{a_t}\cdots P_{a_1}y_i\|_2^2 . \tag{8}$$

Only the recurrent matrices are optimized: $\dot{R}_a = -\nabla_{R_a}\mathcal{L}_{L,\mathcal{S}}$, which is summarized as follows.

Theorem 3.2 (Convergence from sufficiently small initialization)**.** *Assume $E^\top E = I_d$ and $O = E^\top$, and consider the balanced decoder-consistent source pool in Eq.* (7)*. For every fixed finite d, h, and $L \geq 1$, there exists $\delta_{\mathcal{S}} > 0$ such that, if $\left(\sum_{a\in\mathbb{Z}_d} \|R_a(0)\|_F^2\right)^{1/2} < \delta_{\mathcal{S}}$, then gradient flow on $\mathcal{L}_{L,\mathcal{S}}$ converges to a finite limit R_∞ satisfying $\mathcal{L}_{L,\mathcal{S}}(R_\infty) = 0$.*

Proof. See Appendix A.4. □

Theorem 3.2 is an optimization result only, and no rank condition on H is required for convergence. Full rank is what converts zero-loss convergence into length generalization (proof in Appendix A.5).

Theorem 3.3 (Full source coverage forces length generalization)**.** *Consider the decoder-consistent source pool $\mathcal{S}$ defined in Eq.* (7)*, with source matrix H and label matrix Y. If $\mathrm{rank}(H) = h$, then, for every finite $L \geq 1$,*

$$\mathcal{L}_{L,\mathcal{S}}(R) = 0 \quad \Longleftrightarrow \quad OR_a = P_a O, \qquad \forall a \in \mathbb{Z}_d. \tag{9}$$

Consequently, every zero-loss solution generalizes exactly to arbitrary sequence lengths.

Together, the preceding results give an end-to-end guarantee: the source augmentation preserves convergence to zero loss, while its full-rank coverage ensures that every zero-loss limit is length-generalizing (proof in Appendix A.6).

Corollary 3.1 (Small initialization yields length generalization)**.** *Consider the balanced decoder-consistent source pool in Eq.* (7)*, and assume $n \geq h - d$. Then, almost surely over the Gaussian source directions, there exists $\delta > 0$ such that every initialization satisfying $\left(\sum_{a\in\mathbb{Z}_d} \|R_a(0)\|_F^2\right)^{1/2} < \delta$ converges under gradient flow to a finite limit R_∞ satisfying $OR_{a,\infty} = P_a O$ for every $a \in \mathbb{Z}_d$.*

In particular, if $(R_a(0))_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \varepsilon^2/h)$, then $\Pr(R_\infty$ is length-generalizing$) \longrightarrow 1$ as $\varepsilon \to 0$.

In the canonical coordinates, writing $R_a = \left(\begin{smallmatrix} A_a & B_a \\ C_a & D_a \end{smallmatrix}\right)$, Eq. (9) gives $A_{a,\infty} = P_a$ and $B_{a,\infty} = 0$. Hence every limiting solution in Corollary 3.1 has the form

$$R_{a,\infty} = \begin{pmatrix} P_a & 0 \\ C_{a,\infty} & D_{a,\infty} \end{pmatrix}.$$

The blocks $C_{a,\infty}$ and $D_{a,\infty}$ need not vanish; they describe decoder-invisible dynamics that cannot feed back into the output. Numerical experiments in Appendix D.2 verify the predicted coverage threshold and the small-initialization regime of Corollary 3.1.

4 Memorization and Length-Generalization Times

The previous result characterizes the endpoint of training. With a full-rank balanced source pool and sufficiently small initialization, gradient flow converges to a zero-loss solution that generalizes to every finite horizon. But *when and how does length generalization occur?*

To separate short-horizon fitting from long-horizon generalization, we evaluate the trained recurrent dynamics only on the original task distribution, starting from the reset states Ee_x. For any evaluation horizon $K \geq 1$, define a quantity that measures the population error over all prefixes up to K:

$$\mathcal{E}^{(K)}(R) = \frac{1}{2Kd^{K+1}} \sum_{x\in\mathbb{Z}_d} \sum_{a_{1:K}\in\mathbb{Z}_d^K} \sum_{t=1}^{K} \|OR_{a_t}\cdots R_{a_1}Ee_x - P_{a_t}\cdots P_{a_1}e_x\|_2^2 . \tag{10}$$

For the training horizon L, we define the memorization error as $\mathcal{E}^{(L)}_{\mathrm{mem}}(R) := \mathcal{E}^{(L)}(R)$. It measures whether the model has learned the task at the same horizon used during training. For a longer horizon $T > L$, define the length-generalization error as $\mathcal{E}^{(T)}_{\mathrm{gen}}(R) := \mathcal{E}^{(T)}(R)$. For an error threshold $\tau > 0$, define the persistent memorization and generalization time, respectively, by

$$\begin{aligned} t^{(L)}_{\mathrm{mem}}(\tau) &= \inf\left\{t_0 : \mathcal{E}^{(L)}_{\mathrm{mem}}(R(t)) \leq \tau \text{ for all } t \geq t_0\right\}, \\ t^{(T)}_{\mathrm{gen}}(\tau; L) &= \inf\left\{t_0 : \mathcal{E}^{(T)}_{\mathrm{gen}}(R(t)) \leq \tau \text{ for all } t \geq t_0\right\}. \end{aligned} \tag{11}$$

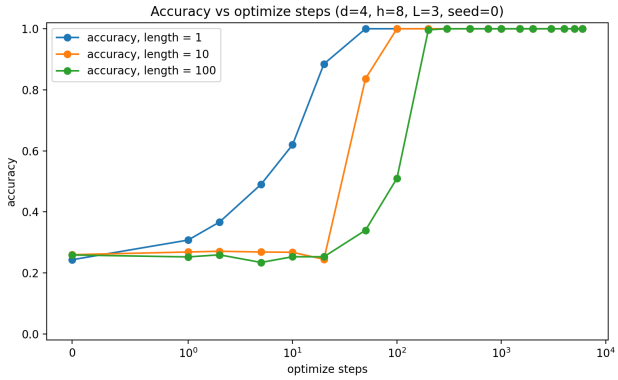


Figure 1: **Delayed length generalization under balanced source-pool training.** For an overparameterized linear RNN with $d = 4$, $h = 8$, and training horizon $L = 3$, we train the recurrent transitions using the balanced decoder-consistent source pool in Eq. (7), with $\rho = 1$ and recurrent initialization scale $\sigma_R = 0.05$. Accuracy is evaluated throughout optimization at rollout lengths 1, 10, and 100. Short rollouts become accurate first, while longer rollouts remain inaccurate for additional optimization steps before eventually reaching perfect accuracy.

We define their difference as $\Delta t_{L,T}(\tau) = t^{(T)}_{\mathrm{gen}}(\tau; L) - t^{(L)}_{\mathrm{mem}}(\tau)$. A positive and substantial $\Delta t_{L,T}$ means that *the model has already learned the task at the training horizon while still making errors on much longer rollouts.* This is the recurrent form of **grokking**, the delayed generalization, that we aim to explain, and thus we call $\Delta t_{L,T}$ the grokking delay in such case.

Figure 1 gives an illustration of this separation. We evaluate the same model at sequence lengths 1, 10, and 100 throughout training. Short rollouts become accurate first, while substantially longer rollouts remain near chance for many additional optimization steps before rapidly improving. This delayed emergence of long-horizon accuracy is the phenomenon that the local spectral analysis below aims to explain. By Corollary 3.1, both errors eventually vanish. The problem is therefore dynamical: *what makes the long-horizon error decay much later than the short-horizon error?*

4.1 A Local Spectral Mechanism for the Delay

For small τ, the two hitting times occur late in training, after the trajectory has entered a neighborhood of its limiting length-generalizing solution $R_\star$. We can therefore ask two local questions: which residual parameter errors decay slowly near $R_\star$, and how strongly are those errors seen at different rollout horizons? The first question is controlled by the local curvature of the training objective.

Definition 4.1 (Local training mode)**.** *Let* $H^{(L)}_{\mathrm{tr}} := \nabla^2 \mathcal{L}_{L,\mathcal{S}}(R_\star)$. *A* local training mode *is a positive-curvature eigenvector* $u_j = \{U_{j,a}\}_{a\in\mathbb{Z}_d}$ *in the normal space of the zero-loss manifold such that* $H^{(L)}_{\mathrm{tr}} u_j = \kappa_j(L) u_j, \qquad \kappa_j(L) > 0.$

Thus u_j is a coordinated perturbation of the recurrent matrices, and $\kappa_j(L)$ is its local decay rate under gradient flow. Small $\kappa_j(L)$ corresponds to a slowly removed residual parameter error.

We assume the late-time trajectory lies in the regular local regime

$$R(t) - R_\star = \sum_j c_j e^{-\kappa_j(L)t} u_j + o(e^{-\kappa_{\min} t}), \tag{12}$$

where $\kappa_{\min}$ is the smallest positive rate among the active modes. The coefficient c_j is the amplitude of mode u_j along the actual trajectory; we call the mode active when $c_j \neq 0$.

Slowness alone, however, does not produce delayed length generalization. A slow residual mode creates a gap between the two evaluation errors only if short and long rollouts respond to it differently. This motivates a second quantity.

Definition 4.2 (Horizon observability)**.** *Let* $H_K := \nabla^2 \mathcal{E}^{(K)}(R_\star)$*. The* horizon-$K$ observability *of mode* u_j *is* $q_j(K) := \langle u_j, H_K u_j \rangle$*. For the two horizons of interest, write* $m_j(L) := q_j(L)$ *and* $g_j(T) := q_j(T)$*.*

Thus $m_j(L)$ measures how strongly a residual perturbation along u_j affects the memorization error, while $g_j(T)$ measures its effect on the long-horizon generalization error. If $g_j(T) \gg m_j(L)$, the same parameter error can be almost harmless on training-length rollouts while remaining important on longer ones.

Not every slow mode controls a hitting time. It must be present along the trajectory and visible to the corresponding evaluation error. Accordingly, define

$$\kappa_{\mathrm{mem}} = \min\{\kappa_j(L) : c_j \neq 0,\ m_j(L) > 0\}, \qquad \kappa_{\mathrm{gen}} = \min\{\kappa_j(L) : c_j \neq 0,\ g_j(T) > 0\}.$$

Both are rates of the same training dynamics; they can differ only because the two evaluation horizons may observe different active modes. Specifically, we can show that (proof in Appendix B.1)

Theorem 4.1 (Local spectral-delay law)**.** *Suppose the late-time expansion Eq.* (12) *holds, and suppose the slowest active-visible modes defining* κ_{mem} *and* κ_{gen} *are unique, with indices* j_m *and* j_g*. Then, as* $\tau \downarrow 0$*,*

$$\Delta t_{L,T}(\tau) = \frac{1}{2}\left(\frac{1}{\kappa_{\mathrm{gen}}} - \frac{1}{\kappa_{\mathrm{mem}}}\right)\log\frac{1}{\tau} + \frac{\log A_{\mathrm{gen}}}{2\kappa_{\mathrm{gen}}} - \frac{\log A_{\mathrm{mem}}}{2\kappa_{\mathrm{mem}}} + o(1), \tag{13}$$

where $A_{\mathrm{mem}} = \frac{1}{2}c_{j_m}^2 m_{j_m}(L)$ *and* $A_{\mathrm{gen}} = \frac{1}{2}c_{j_g}^2 g_{j_g}(T)$*.*

The result follows by inserting Eq. (12) into the quadratic expansions of the two evaluation errors: each late-time error is dominated by its slowest active-visible mode, and solving the resulting exponential decay for the threshold-crossing time gives Eq. (13). The mechanism is especially transparent when the same mode controls both errors (proof in Appendix B.2).

Corollary 4.1 (Same-mode delay)**.** *If the same active mode* $u_\star$ *controls both memorization and generalization, then* $\kappa_{\mathrm{mem}} = \kappa_{\mathrm{gen}} = \kappa_\star$*, and* $\Delta t_{L,T}(\tau) = \frac{1}{2\kappa_\star}\log\frac{g_\star(T)}{m_\star(L)} + o(1),\ \tau \downarrow 0.$

Corollary 4.1 isolates the two ingredients of the delay. A small $\kappa_\star$ makes the residual parameter error persist, while a large observability ratio $g_\star(T)/m_\star(L)$ makes the same error much more consequential on long rollouts than on training-length rollouts. The local theory therefore tells us what to look for: a recurrent parameter direction that is both weakly trained and strongly horizon-dependent. We now identify a concrete candidate in the decoder-invisible dynamics.

5 Visible and Invisible Recurrent Dynamics

Recall the canonical block form $R_a = \left(\begin{smallmatrix} A_a & B_a \\ C_a & D_a \end{smallmatrix}\right)$. By Eq. (9), every length-generalizing endpoint satisfies $A_{a,\star} = P_a$ and $B_{a,\star} = 0$. Residual error in A_a directly perturbs the task transition, whereas B_a maps decoder-invisible hidden coordinates back into the visible coordinates. A residual B_a therefore represents leakage of hidden memory into the task prediction.

The balanced source pool in Eq. (7) determines how strongly these invisible directions are constrained. Writing $G = [g_1, \ldots, g_n] \in \mathbb{R}^{k\times n}$, with $k = h - d$, define the invisible source covariance

$$\Sigma_u := \frac{1}{N}\sum_{i=1}^{N} u_i u_i^\top = \frac{2\rho^2}{1+2n} G G^\top. \tag{14}$$

A small eigenvalue of Σ_u means that the source pool contains little variation along the corresponding decoder-invisible direction. We next show that this weak coverage translates directly into weak training curvature for the corresponding B-direction.

5.1 Exact B-Sector Curvature

Consider the canonical length-generalizing zero-loss point $R_a^0 = \begin{pmatrix} P_a & 0 \\ 0 & 0 \end{pmatrix}$. We use R^0 only as an analytically clean reference; the actual endpoint reached by gradient flow may have nonzero $C_{a,\star}$

and $D_{a,\star}$. Restrict attention to B-sector perturbations $\Delta R_a = \begin{pmatrix} 0 & Y_a \\ 0 & 0 \end{pmatrix}$, $Y_a \in \mathbb{R}^{d\times k}$. These perturbations measure how strongly training penalizes a small amount of decoder-invisible information feeding back into the visible state, as shown below (proof in Appendix C.1).

Theorem 5.1 (Exact B-sector curvature)**.** *Let $H_B^{(L)}$ denote the restriction of the training Hessian at R^0 to the B-blocks. Then*

$$\left\langle Y, H_B^{(L)} Y \right\rangle = \frac{1}{d} \sum_{a\in\mathbb{Z}_d} \operatorname{tr}\left(Y_a \Sigma_u Y_a^\top\right). \tag{15}$$

In particular, the B-sector curvature at R^0 is exactly independent of the training horizon L.

The reason is simple. At R^0, the decoder-invisible component can enter the visible output to first order only through the first-step B-perturbation. Subsequent multiplication by permutation matrices preserves its norm, so every supervised depth contributes the same quadratic term and the average over depths cancels the factor $1/L$.

Theorem 5.1 directly connects source geometry to training speed. If $\Sigma_u v_q = \mu_q v_q$, then a B-mode reading from the invisible direction v_q has local rate $\kappa_{B,q} = \mu_q/d$. Hence

$$\kappa_{B,\min} = \frac{\lambda_{\min}(\Sigma_u)}{d}. \tag{16}$$

Its characteristic decay time is therefore $d/\lambda_{\min}(\Sigma_u)$ (verified in Appendix D.1). This provides a concrete candidate for the slow mode in Corollary 4.1.

5.2 Slow Modes Near the Coverage Threshold

Proposition 3.1 shows that $n \geq k = h - d$ is enough for the Gaussian source directions to span the entire decoder-invisible subspace almost surely. Full rank guarantees identifiability, but it does not guarantee good conditioning. Let $\gamma := k/n = (h-d)/n$. In the high-dimensional regime $k, n \to \infty$ with $k/n \to \gamma < 1$, the Marchenko–Pastur lower edge, together with Eq. (14), gives

$$\lambda_{\min}(\Sigma_u) \longrightarrow \rho^2(1-\sqrt{\gamma})^2, \qquad \kappa_{B,\min} \sim \frac{\rho^2}{d}(1-\sqrt{\gamma})^2. \tag{17}$$

Thus, as $\gamma \uparrow 1$, the source pool can already have full rank while its weakest invisible direction remains extremely poorly conditioned. Since $(1-\sqrt{\gamma})^2 \sim (1-\gamma)^2/4$, the associated local timescale grows as $d/[\rho^2(1-\gamma)^2]$, up to a constant factor. At finite dimension there is no rank contradiction at $n = k$: a square Gaussian matrix is full rank almost surely. The slowdown above is a high-dimensional hard-edge conditioning effect. The finite-dimensional simulations in Figure 2 show the corresponding approach to the asymptotic prediction: the empirical smallest eigenvalue $\lambda_{\min}(\Sigma_u)/\rho^2$ for Gaussian source pools is compared with the asymptotic prediction $(1-\sqrt{\gamma})^2$, where $\gamma = (h-d)/n$. As the invisible dimension increases, the empirical smallest eigenvalue approaches the Marchenko–Pastur lower edge, while the largest finite-size deviations occur near the hard edge $\gamma = 1$.

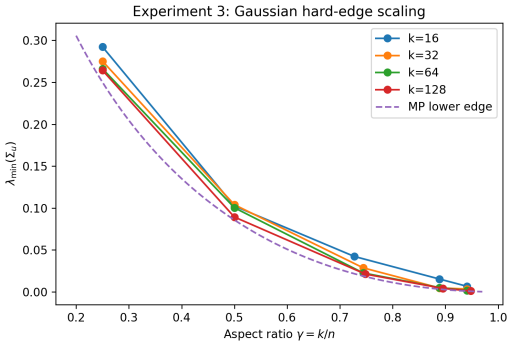


Figure 2: Gaussian hard-edge scaling.

5.3 From the Slow Mode to the Grokking Delay

The canonical calculation above supplies the first ingredient of Corollary 4.1: a parameter direction whose training rate can become very small. To turn this into a prediction for the actual training trajectory, two further properties are required. The weak B-direction must persist as a slow active

mode at the endpoint reached by gradient flow, and that same mode must be substantially more observable at horizon T than at horizon L.

The first point is not automatic, because the actual endpoint may contain nonzero $C_{a,\star}$ and $D_{a,\star}$, which can modify the local Hessian through longer recurrent paths. We therefore make the dependence explicit.

Assumption 5.2 (Weak-mode persistence)**.** *In the weak-source regime of interest, the actual endpoint $R_\star$ has an active local mode $u_\star$ associated with the weakest canonical B-direction and satisfying $\kappa_\star = \frac{\lambda_{\min}(\Sigma_u)}{d}\big(1 + o(1)\big)$.*

If this mode also controls both evaluation errors, then Corollary 4.1 yields

$$\Delta t_{L,T} = \frac{d}{2\lambda_{\min}(\Sigma_u)} \log \frac{g_\star(T)}{m_\star(L)} + o\left(\frac{1}{\lambda_{\min}(\Sigma_u)}\right). \tag{18}$$

Combining this with Eq. (17) gives the high-dimensional Gaussian scaling

$$\Delta t_{L,T} \sim \frac{d}{2\rho^2(1-\sqrt{\gamma})^2} \log \frac{g_\star(T)}{m_\star(L)}. \tag{19}$$

The two factors have distinct roles. The source-conditioning term $1/\lambda_{\min}(\Sigma_u)$ determines how slowly the residual mode is removed, while $g_\star(T)/m_\star(L)$ determines how differently that same mode appears at short and long horizons.

Eqs. (18)–(19) are conditional on Assumption 5.2 and on the same-mode mechanism of Corollary 4.1. The exact result established in this section is the canonical B-sector curvature in Theorem 5.1.

We finally test whether this conditional mechanism is realized along actual training trajectories. Figure 5 in Appendix D.3 provides a trajectory-level test of each step in the proposed mechanism. First, the training-horizon error decays before the long-horizon error, producing a clear memorization–generalization delay. Second, the late-time decay rate of the dominant slow mode closely matches the corresponding eigenvalue of the endpoint Hessian. Third, this mode remains strongly concentrated in the decoder-invisible B-sector, with a decay rate on the same scale as the canonical prediction $\lambda_{\min}(\Sigma_u)/d$. Finally, using the measured decay rate and horizon observabilities in Corollary 4.1 predicts the observed late-time delay at the correct scale. While Assumption 5.2 is not proved in full generality, these experiments show that the proposed mechanism is realized along the training trajectories considered here.

This also clarifies the role of the training horizon. There is no universal law such as $\Delta t \propto L$ or $\Delta t \propto 1/L$. At the canonical solution, the weakest B-sector training rate is exactly independent of L; the training horizon instead enters the observed delay through the short-horizon observability $m_\star(L)$, and more generally through the actual endpoint and its local Hessian. The resulting picture is therefore simple: source coverage determines whether the length-generalizing recurrence is identifiable, while the conditioning of that coverage determines how slowly it can emerge.

6 Conclusion

We developed a tractable theory for the transition from short-horizon fitting to length generalization in recurrent models. In our overparameterized linear RNN, fitting finite training sequences does not guarantees identifying a reusable recurrence. We showed that sufficient hidden-state coverage resolves this identifiability problem, while sufficiently small initialization ensures convergence to a length-generalizing solution.We then studied why length generalization can nevertheless emerge much later than short-horizon fitting. The key distinction is between whether a hidden direction is covered and how strongly it is covered. Weakly covered decoder-invisible directions induce slowly learned recurrent modes, and a large grokking delay can arise when such a mode is much more consequential for long rollouts than for training-length rollouts.

AI USE STATEMENT

In this work, we used generative AI tools as research and writing assistants. They were used to help refine the theoretical framework, formulate and check mathematical derivations, assist in drafting proofs, provide feedback on experimental design, generate and edit research code, and assist with the interpretation and presentation of intermediate results. Generative AI tools were also used to help organize the manuscript, improve clarity and readability, and assist with literature search and manuscript editing.

We did not use generative AI tools to fabricate or alter experimental results, generate benchmark datasets, perform qualitative or thematic data analysis, or replace human judgment in evaluating the validity of the research. Translation and human-subject-related tasks were not applicable to this work.

All AI-assisted mathematical arguments, proofs, and theoretical claims were reviewed and checked by the authors. The authors also reviewed the literature and citations used in the paper and independently assessed the interpretation of the experimental and theoretical results. Generative AI outputs were treated as suggestions rather than as authoritative sources.

We take full responsibility for the final content of this work, including all text, mathematical claims, proofs, code, experimental results, figures, and other artifacts produced with the aid of generative AI.

REFERENCES

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Seyedarmin Azizi, Souvik Kundu, Mohammad Sadeghi, and Massoud Pedram. MambaExtend: A training-free approach to improve long context extension of Mamba. In *The Thirteenth International Conference on Learning Representations*, 2025.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. DeciMamba: Exploring the length extrapolation potential of Mamba. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ricardo Buitrago and Albert Gu. Understanding and improving length generalization in recurrent models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5856–5874. PMLR, 2025.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *The Eleventh International Conference on Learning Representations*, 2023.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Ting-Han Fan, Ta-Chung Chi, and Alexander Rudnicky. Advancing regular language reasoning in linear recurrent neural networks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 45–53. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.4.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

Peng Lu, Jerry Huang, Qiuhao Zeng, Xinyu Wang, Boxing Chen, Philippe Langlais, and Yufei Cui. Mamba modulation: On the length generalization of Mamba models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0653.

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.

Abhishek Panigrahi and Navin Goyal. Learning and generalization in RNNs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Frieder Stolzenburg, Sandra Litz, Olivia Michael, and Oliver Obst. The power of linear recurrent neural networks. *arXiv preprint arXiv:1802.03308*, 2018.

Loek Van Rossem and Andrew M. Saxe. Algorithm development in neural networks: Insights from the streaming parity task. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60764–60791. PMLR, 2025.

A Proofs of the Main Theoretical Results

Throughout this appendix, for the collection of recurrent matrices $R = \{R_a\}_{a\in\mathbb{Z}_d}$, we use the product Frobenius norm $\|R\|_{\mathcal{R}}^2 := \sum_{a\in\mathbb{Z}_d} \|R_a\|_F^2$. All gradient flows are taken with respect to this Euclidean structure.

A.1 A local convergence lemma for analytic losses

We first record a standard local argument that will be used in the proofs of Theorems 3.1 and 3.2.

Lemma A.1 (Local trapping near a zero-loss point)**.** *Let $F : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ be real analytic, and let $\theta_\star \in \mathbb{R}^p$ satisfy $F(\theta_\star) = 0$. Then there exists an open neighborhood $\mathcal{U}_\star$ of $\theta_\star$ such that every gradient-flow trajectory $\dot\theta = -\nabla F(\theta)$ initialized in $\mathcal{U}_\star$ exists for all $t \geq 0$, has finite total path length, and converges to a finite limit θ_∞ satisfying $F(\theta_\infty) = 0$.*

Proof. Since F is real analytic and $F(\theta_\star) = 0$, the Łojasiewicz gradient inequality gives a neighborhood $\mathcal{U}$ of $\theta_\star$, constants $c > 0$, and $\vartheta \in [1/2, 1)$ such that

$$\|\nabla F(\theta)\| \geq c\, F(\theta)^\vartheta, \qquad \theta \in \mathcal{U}. \tag{20}$$

Choose $r > 0$ such that the closed ball $\overline{B_{2r}(\theta_\star)}$ is contained in $\mathcal{U}$. By continuity of F and $F(\theta_\star) = 0$, there exists $0 < r_0 < r$ such that

$$\frac{F(\theta_0)^{1-\vartheta}}{c(1-\vartheta)} < r$$

whenever $\theta_0 \in B_{r_0}(\theta_\star)$.

Consider a trajectory initialized at such a point and suppose, temporarily, that it remains in $\mathcal{U}$. Along gradient flow, $\dot F = -\|\nabla F\|^2$. Hence, whenever $F > 0$,

$$-\frac{d}{dt}F^{1-\vartheta} = (1-\vartheta)F^{-\vartheta}\|\nabla F\|^2 \geq c(1-\vartheta)\|\nabla F\|,$$

where we used Eq. (20). Integrating from 0 to T gives

$$\int_0^T \|\dot\theta(t)\|\, dt \leq \frac{F(\theta_0)^{1-\vartheta}}{c(1-\vartheta)} < r.$$

The trajectory therefore cannot leave $B_{2r}(\theta_\star)$: starting inside $B_{r_0}(\theta_\star) \subset B_r(\theta_\star)$, reaching the boundary of $B_{2r}(\theta_\star)$ would require path length at least r. Thus the preceding estimate is valid for all time.

The trajectory remains in a compact subset on which the gradient vector field is bounded and locally Lipschitz, so the solution exists globally. Moreover,

$$\int_0^\infty \|\dot\theta(t)\|\, dt < \infty.$$

Hence $\theta(t)$ is Cauchy as $t \to \infty$ and converges to a finite limit θ_∞.

Finally, Eq. (20) gives $\dot{F} \le -c^2 F^{2\vartheta}$ while the trajectory remains in the neighborhood. Therefore $F(\theta(t)) \to 0$. By continuity, $F(\theta_\infty) = 0$. Taking $\mathcal{U}_\star = B_{r_0}(\theta_\star)$ completes the proof. □

A.2 Proof of Theorem 3.1

Proof of Theorem 3.1. Because $E^\top E = I_d$ and $O = E^\top$, we may extend the columns of E to an orthogonal basis of $\mathbb{R}^h$. An orthogonal change of hidden coordinates preserves the Frobenius metric, the loss, and gradient flow. We may therefore work without loss of generality in the canonical coordinates

$$E = \begin{pmatrix} I_d \\ 0 \end{pmatrix}, \qquad O = \begin{pmatrix} I_d & 0 \end{pmatrix}.$$

It is useful first to rewrite the reset loss using prefixes. Every length-t prefix occurs exactly d^{L-t} times among the length-L action sequences. Therefore Eq. (5) is equivalently

$$\mathcal{L}_L^{\mathrm{reset}}(R) = \frac{1}{2Ld} \sum_{t=1}^{L} \mathbb{E}_{a_{1:t}} \left\| O R_{a_t} \cdots R_{a_1} E - P_{a_t} \cdots P_{a_1} \right\|_F^2, \tag{21}$$

where the expectation is uniform over $(a_1, \ldots, a_t) \in \mathbb{Z}_d^t$.

We now construct an explicit reference trajectory. For $\alpha \in \mathbb{R}$, define

$$R_a(\alpha) = \begin{pmatrix} \alpha P_a & 0 \\ 0 & 0 \end{pmatrix}, \qquad a \in \mathbb{Z}_d.$$

For every prefix $\alpha_{1:t} = (a_1, \ldots, a_t)$,

$$O R_{a_t}(\alpha) \cdots R_{a_1}(\alpha) E = \alpha^t P_{a_t} \cdots P_{a_1}.$$

Hence the loss restricted to this one-dimensional family is

$$f(\alpha) = \frac{1}{2L} \sum_{t=1}^{L} (\alpha^t - 1)^2.$$

We next verify that this family is invariant under the full matrix gradient flow. Write a perturbation of R_b in block form as

$$\Delta R_b = \begin{pmatrix} \Delta A_b & \Delta B_b \\ \Delta C_b & \Delta D_b \end{pmatrix}.$$

Along the reference family, every hidden state generated from a reset state lies entirely in the visible subspace. Consequently, a perturbation in ΔB_b or ΔD_b receives zero invisible input. A perturbation in ΔC_b creates an invisible component, but this component is annihilated by the readout if the perturbation occurs at the final step, and by the next reference transition if another recurrent step follows. Thus the directional derivatives in the B, C, and D blocks vanish.

It remains to compute the A_b derivative. Fix a prefix of length t and an occurrence s with $a_s = b$. At the reference point, the first-order change in the output matrix caused by ΔA_b is

$$\alpha^{t-1} P_{a_t} \cdots P_{a_{s+1}} \Delta A_b P_{a_{s-1}} \cdots P_{a_1}.$$

The corresponding residual is $(\alpha^t - 1) P_{a_t} \cdots P_{a_1}$. Since permutation matrices are orthogonal, their Frobenius inner product reduces to

$$(\alpha^t - 1) \alpha^{t-1} \langle \Delta A_b, P_b \rangle_F.$$

Under the uniform action distribution, the expected number of occurrences of action b in a length-t prefix is t/d. Substituting into Eq. (21) yields

$$\nabla_{A_b} \mathcal{L}_L^{\mathrm{reset}} = \frac{1}{Ld^2} \left[\sum_{t=1}^{L} t \alpha^{t-1} (\alpha^t - 1) \right] P_b.$$

Therefore the full gradient-flow vector field is tangent to the reference family, and the scalar parameter obeys

$$\dot{\alpha} = \frac{1}{Ld^2} \sum_{t=1}^{L} t\alpha^{t-1}(1-\alpha^t). \tag{22}$$

Starting from the zero recurrent initialization gives $\alpha(0) = 0$. For every $\alpha \in [0, 1)$, the right-hand side of Eq. (22) is strictly positive. The point $\alpha = 1$ is stationary. By uniqueness of solutions, the interval $[0, 1]$ is forward invariant, so $\alpha(t)$ is increasing and bounded. Let $\bar{\alpha}$ be its limit. If $\bar{\alpha} < 1$, continuity of the vector field would give $\dot{\alpha} \geq c_0 > 0$ for all sufficiently large times, contradicting convergence of $\alpha(t)$. Hence $\bar{\alpha} = 1$.

Thus the reference trajectory starting from $R = 0$ converges to

$$R_a^\star = \begin{pmatrix} P_a & 0 \\ 0 & 0 \end{pmatrix}, \qquad a \in \mathbb{Z}_d,$$

and $\mathcal{L}_L^{\text{reset}}(R^\star) = 0$.

The loss $\mathcal{L}_L^{\text{reset}}$ is a nonnegative polynomial in the entries of the recurrent matrices and is therefore real analytic. Lemma A.1 gives an open neighborhood $\mathcal{U}_\star$ of $R^\star$ such that every gradient-flow trajectory entering $\mathcal{U}_\star$ converges to a finite zero-loss limit.

Since the reference trajectory converges to $R^\star$, there exists a finite time T such that $R^{\text{ref}}(T) \in \mathcal{U}_\star$. Choose an open set $\mathcal{V}$ satisfying $R^{\text{ref}}(T) \in \mathcal{V}$ and $\overline{\mathcal{V}} \subset \mathcal{U}_\star$.

The gradient vector field is polynomial and hence locally Lipschitz. Solutions therefore depend continuously on their initial conditions on every finite interval on which the reference solution remains bounded. Consequently, there exists $\delta_{\text{reset}} > 0$ such that every initial condition satisfying $\|R(0)\|_{\mathcal{R}} < \delta_{\text{reset}}$ has a solution on $[0, T]$ and satisfies $R(T) \in \mathcal{V} \subset \mathcal{U}_\star$. Lemma A.1 then implies global existence and convergence to a finite zero-loss limit.

This proves Theorem 3.1. □

A.3 Proof of Proposition 3.1

Proof of Proposition 3.1. Let $k = h - d$. Since $OE = I_d$, the matrix E has rank d, so the reset states $\{Ee_x : x \in \mathbb{Z}_d\}$ span the d-dimensional subspace $\mathrm{Im}(E)$.

For every x and j, the balanced source pool contains $h_{x,j}^+ = Ee_x + \rho K g_j$ and $h_{x,j}^- = Ee_x - \rho K g_j$. Their difference is

$$h_{x,j}^+ - h_{x,j}^- = 2\rho K g_j.$$

Hence the column span of H contains $\mathrm{span}\{Kg_1, \ldots, Kg_n\}$.

Let $G = (g_1, \ldots, g_n) \in \mathbb{R}^{k \times n}$. When $n \geq k$ and the g_j are independent standard Gaussian vectors, G has row rank k almost surely. Indeed, the event $\mathrm{rank}(G) < k$ is the zero set of all $k \times k$ minors of G; at least one such minor is a nonzero polynomial in the Gaussian entries, and a nonzero polynomial vanishes on a Lebesgue-null set. Thus $\mathrm{span}\{g_1, \ldots, g_n\} = \mathbb{R}^k$ almost surely. Since the columns of K form a basis of $\ker(O)$,

$$\mathrm{span}\{Kg_1, \ldots, Kg_n\} = \ker(O)$$

almost surely.

It remains to show that the visible and invisible subspaces form a direct sum. If $v \in \mathrm{Im}(E) \cap \ker(O)$, then $v = Ec$ for some $c \in \mathbb{R}^d$ and $0 = Ov = OEc = c$. Hence $v = 0$. Therefore $\mathrm{Im}(E) \cap \ker(O) = \{0\}$. Their dimensions are d and $h - d$, respectively, so

$$\mathbb{R}^h = \mathrm{Im}(E) \oplus \ker(O).$$

The columns of H contain vectors spanning both summands. Consequently $\mathrm{rank}(H) = h$ almost surely. □

A.4 Proof of Theorem 3.2

Proof of Theorem 3.2. As in the proof of Theorem 3.1, the assumptions $E^\top E = I_d$ and $O = E^\top$ allow us to work in the canonical coordinates

$$E = \begin{pmatrix} I_d \\ 0 \end{pmatrix}, \qquad O = \begin{pmatrix} I_d & 0 \end{pmatrix}.$$

In these coordinates, each source state can be written as $h_i = (y_i, u_i)^\top$, where $y_i \in \{e_0, \dots, e_{d-1}\}$ and $u_i \in \mathbb{R}^{h-d}$.

The balanced construction implies two identities. First, each task label appears exactly $1+2n$ times, so

$$\frac{1}{N}\sum_{i=1}^{N} y_i y_i^\top = \frac{1}{d} I_d. \tag{23}$$

Second, every invisible perturbation $+\rho g_j$ is paired with $-\rho g_j$ carrying the same label, so

$$\frac{1}{N}\sum_{i=1}^{N} y_i u_i^\top = 0. \tag{24}$$

Using prefix counting as before, the source-pool objective can be written as

$$\mathcal{L}_{L,\mathcal{S}}(R) = \frac{1}{2L}\sum_{t=1}^{L} \mathbb{E}_{a_{1:t}} \frac{1}{N}\sum_{i=1}^{N} \|O R_{a_t} \cdots R_{a_1} h_i - P_{a_t} \cdots P_{a_1} y_i\|_2^2 .$$

Consider again the reference family

$$R_a(\alpha) = \begin{pmatrix} \alpha P_a & 0 \\ 0 & 0 \end{pmatrix}.$$

For any source $h_i = (y_i, u_i)^\top$, the first recurrent step gives

$$R_{a_1}(\alpha) h_i = \begin{pmatrix} \alpha P_{a_1} y_i \\ 0 \end{pmatrix}.$$

Thus the invisible source component is removed after the first reference transition. Inductively,

$$O R_{a_t}(\alpha) \cdots R_{a_1}(\alpha) h_i = \alpha^t P_{a_t} \cdots P_{a_1} y_i.$$

Since $\|y_i\|_2 = 1$, the restriction of the full augmented loss to the reference family is again

$$f(\alpha) = \frac{1}{2L}\sum_{t=1}^{L} (\alpha^t - 1)^2.$$

We now verify that the full augmented gradient remains tangent to this family. Write $\Delta R_b = \left(\begin{smallmatrix} \Delta A_b & \Delta B_b \\ \Delta C_b & \Delta D_b \end{smallmatrix}\right)$.

A C_b perturbation maps a visible reference state into the invisible subspace. If it occurs at the final step, it is killed by O; if another step follows, the next reference matrix annihilates it. A D_b perturbation has the same property. Therefore the C_b and D_b directional derivatives vanish.

For a B_b perturbation, the invisible input is zero at every recurrent step except possibly the first one. If $a_1 = b$, the first-order output perturbation at depth t is

$$\alpha^{t-1} P_{a_t} \cdots P_{a_2} \Delta B_b u_i.$$

Its inner product with the reference residual is proportional to $\langle \Delta B_b u_i, P_b y_i \rangle$. Averaging over the source pool gives

$$\frac{1}{N}\sum_i \langle \Delta B_b u_i, P_b y_i \rangle = \left\langle \Delta B_b,\ P_b \left(\frac{1}{N}\sum_i y_i u_i^\top \right) \right\rangle_F = 0$$

by Eq. (24). Hence $\nabla_{B_b}\mathcal{L}_{L,\mathcal{S}} = 0$ on the reference family.

Finally, consider an A_b perturbation at position s of a length-t prefix. Averaging its inner product with the residual over the source pool gives

$$\frac{1}{N}\sum_i \left\langle \Delta A_b P_{a_{s-1}} \cdots P_{a_1} y_i,\, P_b P_{a_{s-1}} \cdots P_{a_1} y_i \right\rangle = \frac{1}{d}\langle \Delta A_b, P_b\rangle_F,$$

where Eq. (23) and orthogonality of the permutation matrices were used. Averaging over action sequences contributes the expected number t/d of occurrences of action b. Therefore

$$\nabla_{A_b}\mathcal{L}_{L,\mathcal{S}} = \frac{1}{Ld^2}\left[\sum_{t=1}^{L} t\alpha^{t-1}(\alpha^t - 1)\right] P_b.$$

Thus the augmented gradient flow restricted to the reference family is exactly the scalar flow in Eq. (22). Starting from $R = 0$, it follows that $\alpha(t) \uparrow 1$ and

$$R_a(t) \longrightarrow R_a^\star = \begin{pmatrix} P_a & 0 \\ 0 & 0 \end{pmatrix}.$$

For this limit every source trajectory is correct at every supervised prefix, so $\mathcal{L}_{L,\mathcal{S}}(R^\star) = 0$.

The loss $\mathcal{L}_{L,\mathcal{S}}$ is a finite sum of squared polynomial functions of the recurrent entries and is therefore a nonnegative real-analytic function. By Lemma A.1, there exists an open neighborhood $\mathcal{U}_\star$ of $R^\star$ such that every trajectory entering $\mathcal{U}_\star$ converges to a finite zero-loss limit.

The reference trajectory from $R = 0$ enters this neighborhood after some finite time T. By finite-time continuous dependence of the polynomial gradient-flow ODE on its initial condition, there exists $\delta_{\mathcal{S}} > 0$ such that every initialization satisfying $\|R(0)\|_{\mathcal{R}} < \delta_{\mathcal{S}}$ is defined on $[0, T]$ and satisfies $R(T) \in \mathcal{U}_\star$. Applying Lemma A.1 from time T onward proves that every such trajectory converges to a finite limit R_∞ with $\mathcal{L}_{L,\mathcal{S}}(R_\infty) = 0$.

No rank assumption on H was used in this argument. The source-pool symmetry is sufficient for the convergence statement; full rank is needed only for the identifiability result proved next. □

A.5 Proof of Theorem 3.3

Proof of Theorem 3.3. Recall that the source construction is decoder-consistent: $y_i = Oh_i$ for every source pair (h_i, y_i).

Suppose first that $OR_a = P_a O$ for every action a. Repeated application of this identity gives, for every prefix $(a_1, \ldots, a_t)$,

$$OR_{a_t} \cdots R_{a_1} = P_{a_t} \cdots P_{a_1} O.$$

Therefore, for every source h_i,

$$OR_{a_t} \cdots R_{a_1} h_i = P_{a_t} \cdots P_{a_1} O h_i = P_{a_t} \cdots P_{a_1} y_i.$$

Every summand in $\mathcal{L}_{L,\mathcal{S}}$ is zero, and hence $\mathcal{L}_{L,\mathcal{S}}(R) = 0$.

Conversely, suppose $\mathcal{L}_{L,\mathcal{S}}(R) = 0$. The loss is a sum of nonnegative squared norms, so every supervised residual must vanish. In particular, its $t = 1$ terms imply

$$OR_a h_i = P_a y_i = P_a O h_i$$

for every action a and every source state h_i. Thus

$$(OR_a - P_a O)H = 0.$$

Since $\mathrm{rank}(H) = h$, the columns of H span $\mathbb{R}^h$. Therefore the linear map $OR_a - P_a O$ vanishes on all of $\mathbb{R}^h$, and hence

$$OR_a = P_a O, \qquad \forall a \in \mathbb{Z}_d.$$

This establishes the equivalence in Eq. (9).

Finally, applying the intertwining identity repeatedly and using $OE = I_d$ gives, for every $T \geq 1$ and every action sequence,

$$OR_{a_T} \cdots R_{a_1} E = P_{a_T} \cdots P_{a_1} OE = P_{a_T} \cdots P_{a_1}.$$

Hence every zero-loss solution generalizes exactly to arbitrary finite sequence lengths. □

A.6 Proof of Corollary 3.1

Proof of Corollary 3.1. Let $k = h - d$ and suppose $n \geq k$. By Proposition 3.1, the Gaussian source directions satisfy $\mathrm{rank}(H) = h$ with probability one.

Fix any source-pool realization in this probability-one event. Theorem 3.2 then provides a radius $\delta = \delta(\mathcal{S}, L) > 0$ such that every recurrent initialization with $\|R(0)\|_{\mathcal{R}} < \delta$ converges to a finite zero-loss limit R_∞. Since H is full rank, Theorem 3.3 implies

$$OR_{a,\infty} = P_a O$$

for every action a. The same theorem then implies exact generalization to every finite sequence length. This proves the deterministic small-norm statement, almost surely over the Gaussian source directions.

Now suppose, independently of the source-pool draw, that the recurrent entries satisfy $(R_a(0))_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \varepsilon^2/h)$. There are dh^2 recurrent entries in total, so conditional on the source pool,

$$\|R(0)\|_{\mathcal{R}}^2 \overset{d}{=} \frac{\varepsilon^2}{h}\, \chi^2_{dh^2}.$$

For every fixed full-rank source realization, its associated radius $\delta(\mathcal{S}, L)$ is strictly positive. Hence

$$\Pr(\|R(0)\|_{\mathcal{R}} < \delta(\mathcal{S}, L) \,|\, \mathcal{S}) = \Pr\left(\chi^2_{dh^2} < \frac{h\,\delta(\mathcal{S}, L)^2}{\varepsilon^2}\right) \longrightarrow 1$$

as $\varepsilon \to 0$.

On this event, the preceding deterministic argument guarantees that the limit is length-generalizing. Therefore, conditional on almost every source-pool realization,

$$\Pr(R_\infty \text{ is length-generalizing} \,|\, \mathcal{S}) \longrightarrow 1.$$

If probability is taken jointly over both the Gaussian source directions and the independent recurrent initialization, the same conclusion follows from dominated convergence, since the conditional probabilities are bounded by one and the source pool is full rank almost surely. Thus

$$\Pr(R_\infty \text{ is length-generalizing}) \longrightarrow 1 \qquad \text{as } \varepsilon \to 0.$$

This proves Corollary 3.1. □

B Proofs for the Local Spectral Delay Mechanism

We next prove the results in Section 4.1. The argument is local around the limiting length-generalizing solution $R_\star$.

For an evaluation horizon K, let $H_K = \nabla^2 \mathcal{E}^{(K)}(R_\star)$, as in Definition 4.2. Because $\mathcal{E}^{(K)}$ is a sum of squared residuals and $\mathcal{E}^{(K)}(R_\star) = 0$, its gradient vanishes at $R_\star$ and $H_K \succeq 0$.

The phrase "regular local regime" in Section 4.1 refers to the regime in which the quadratic evaluation term is asymptotically dominant along the convergent trajectory. More precisely, for $K \in \{L, T\}$, if κ_K denotes the smallest active rate with $q_j(K) > 0$, we use

$$\mathcal{E}^{(K)}(R(t)) = \frac{1}{2}\,\langle R(t) - R_\star,\, H_K(R(t) - R_\star)\rangle + o(e^{-2\kappa_K t}). \tag{25}$$

This is the local quadratic regime in which the Hessian observability $q_j(K)$ controls the corresponding hitting time.

B.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Let $\xi(t) := R(t) - R_\star$. By Eq. (12),

$$\xi(t) = \sum_j c_j e^{-\kappa_j(L)t} u_j + o(e^{-\kappa_{\min} t}).$$

We first consider the memorization error. Since $H_L \succeq 0$, the condition $m_j(L) = 0$ implies $H_L u_j = 0$. Indeed, $m_j(L) = \langle u_j, H_L u_j \rangle = \|H_L^{1/2} u_j\|^2$, so zero quadratic form implies $H_L^{1/2} u_j = 0$, and hence $H_L u_j = 0$. Therefore any L-invisible mode has zero quadratic cross term with every other mode.

Let j_m be the unique active mode achieving κ_{mem}. Every other active mode with positive L-observability has strictly larger training rate. Consequently,

$$\frac{1}{2}\langle \xi(t), H_L \xi(t) \rangle = \frac{1}{2} c_{j_m}^2 m_{j_m}(L) e^{-2\kappa_{\mathrm{mem}} t}\big(1 + o(1)\big).$$

Using Eq. (25),

$$\mathcal{E}_{\mathrm{mem}}^{(L)}(R(t)) = A_{\mathrm{mem}} e^{-2\kappa_{\mathrm{mem}} t}\big(1 + o(1)\big), \qquad A_{\mathrm{mem}} = \frac{1}{2} c_{j_m}^2 m_{j_m}(L). \tag{26}$$

The same argument at horizon T gives

$$\mathcal{E}_{\mathrm{gen}}^{(T)}(R(t)) = A_{\mathrm{gen}} e^{-2\kappa_{\mathrm{gen}} t}\big(1 + o(1)\big), \qquad A_{\mathrm{gen}} = \frac{1}{2} c_{j_g}^2 g_{j_g}(T). \tag{27}$$

It remains to convert the two error asymptotics into persistent hitting times. We use the following elementary fact. If $f(t) = Ae^{-2\kappa t}(1 + o(1))$, where $A, \kappa > 0$, and

$$t_f(\tau) = \inf\{t_0 : f(t) \le \tau \text{ for all } t \ge t_0\},$$

then

$$t_f(\tau) = \frac{1}{2\kappa} \log \frac{A}{\tau} + o(1) \qquad (\tau \downarrow 0).$$

To see this, put $t_\tau = (2\kappa)^{-1} \log(A/\tau)$. For every fixed $\epsilon > 0$,

$$Ae^{-2\kappa(t_\tau - \epsilon)} = \tau e^{2\kappa\epsilon} > \tau, \qquad Ae^{-2\kappa(t_\tau + \epsilon)} = \tau e^{-2\kappa\epsilon} < \tau.$$

Since the relative error in $f(t)/(Ae^{-2\kappa t})$ tends uniformly to zero on sufficiently late tails, for all sufficiently small τ the persistent crossing occurs between $t_\tau - \epsilon$ and $t_\tau + \epsilon$. Since ϵ is arbitrary, the displayed hitting-time expansion follows.

Applying this fact to Eqs. (26) and (27) yields

$$t_{\mathrm{mem}}^{(L)}(\tau) = \frac{1}{2\kappa_{\mathrm{mem}}} \log \frac{A_{\mathrm{mem}}}{\tau} + o(1),$$

and

$$t_{\mathrm{gen}}^{(T)}(\tau; L) = \frac{1}{2\kappa_{\mathrm{gen}}} \log \frac{A_{\mathrm{gen}}}{\tau} + o(1).$$

Subtracting the first expansion from the second gives Eq. (13):

$$\Delta t_{L,T}(\tau) = \frac{1}{2}\left(\frac{1}{\kappa_{\mathrm{gen}}} - \frac{1}{\kappa_{\mathrm{mem}}}\right) \log \frac{1}{\tau} + \frac{\log A_{\mathrm{gen}}}{2\kappa_{\mathrm{gen}}} - \frac{\log A_{\mathrm{mem}}}{2\kappa_{\mathrm{mem}}} + o(1).$$

□

B.2 Proof of Corollary 4.1

Proof of Corollary 4.1. If the same active mode $u_\star$ controls both errors, then $j_m = j_g = \star$ and $\kappa_{\mathrm{mem}} = \kappa_{\mathrm{gen}} = \kappa_\star$. Moreover,

$$A_{\mathrm{mem}} = \frac{1}{2} c_\star^2 m_\star(L), \qquad A_{\mathrm{gen}} = \frac{1}{2} c_\star^2 g_\star(T).$$

Substituting these identities into Eq. (13), the $\log(1/\tau)$ term cancels. The common factor $c_\star^2/2$ also cancels inside the logarithm. Therefore

$$\Delta t_{L,T}(\tau) = \frac{1}{2\kappa_\star} \log \frac{g_\star(T)}{m_\star(L)} + o(1),$$

which is the conclusion of Corollary 4.1. □

C Invisible-Sector Curvature and Source Conditioning

We now prove the exact B-sector result in Section 5.1 and derive the Gaussian hard-edge scaling used in Section 5.2.

C.1 Proof of Theorem 5.1

Proof of Theorem 5.1. At the canonical zero-loss point,

$$R_a^0 = \begin{pmatrix} P_a & 0 \\ 0 & 0 \end{pmatrix}.$$

Consider a perturbation supported only in the invisible-to-visible block,

$$\Delta R_a = \begin{pmatrix} 0 & Y_a \\ 0 & 0 \end{pmatrix}, \qquad Y_a \in \mathbb{R}^{d\times k}.$$

Because all training residuals vanish at R^0, the Hessian quadratic form is the average squared first-order output perturbation. More explicitly, if $F_{i,a_{1:t}}(R)$ denotes the output residual for source i and prefix $a_{1:t}$, then

$$\left\langle Y, H_B^{(L)} Y \right\rangle = \frac{1}{L}\sum_{t=1}^{L} \mathbb{E}_{a_{1:t}} \frac{1}{N}\sum_{i=1}^{N} \left\| DF_{i,a_{1:t}}(R^0)[Y] \right\|_2^2.$$

Write a source state as $h_i = (y_i, u_i)^\top$. Under the reference dynamics,

$$R_{a_1}^0 h_i = \begin{pmatrix} P_{a_1} y_i \\ 0 \end{pmatrix}.$$

Thus after the first recurrent step the invisible component is zero. Consequently, a B-perturbation can contribute to first order only at the first recurrent step. A perturbation inserted at a later step receives zero invisible input.

For a prefix of length t, the first-order output perturbation is

$$DF_{i,a_{1:t}}(R^0)[Y] = P_{a_t} \cdots P_{a_2} Y_{a_1} u_i.$$

Every P_a is orthogonal, so

$$\left\| DF_{i,a_{1:t}}(R^0)[Y] \right\|_2^2 = \|Y_{a_1} u_i\|_2^2.$$

The expression is independent of the supervised depth t.

The invisible covariance of the balanced source pool is Eq. (14). Indeed, for each task label x and each Gaussian direction g_j, the pair $\pm\rho g_j$ contributes $2\rho^2 g_j g_j^\top$. Since there are d task labels and $N = d(1+2n)$,

$$\frac{1}{N}\sum_{i=1}^{N} u_i u_i^\top = \frac{2\rho^2}{1+2n} G G^\top = \Sigma_u.$$

Therefore

$$\frac{1}{N}\sum_{i=1}^{N} \|Y_{a_1} u_i\|_2^2 = \operatorname{tr}\left(Y_{a_1} \Sigma_u Y_{a_1}^\top\right).$$

Finally, a_1 is uniform on $\mathbb{Z}_d$, so

$$\mathbb{E}_{a_1} \operatorname{tr}\left(Y_{a_1} \Sigma_u Y_{a_1}^\top\right) = \frac{1}{d}\sum_{a\in\mathbb{Z}_d} \operatorname{tr}\left(Y_a \Sigma_u Y_a^\top\right).$$

Since this expression is identical for every depth t, averaging over $t = 1, \ldots, L$ leaves it unchanged. Hence

$$\left\langle Y, H_B^{(L)} Y \right\rangle = \frac{1}{d}\sum_{a\in\mathbb{Z}_d} \operatorname{tr}\left(Y_a \Sigma_u Y_a^\top\right),$$

which is Eq. (15). In particular, the quadratic form contains no dependence on L, proving the claimed exact L-independence.

For completeness, suppose $\Sigma_u v_q = \mu_q v_q$ with $\|v_q\|_2 = 1$. For any action a and visible coordinate r, choose a perturbation whose only nonzero block is $Y_a = e_r v_q^\top$. Then

$$H_B^{(L)} Y = \frac{\mu_q}{d} Y.$$

Thus the B-sector eigenvalues are μ_q/d, and in particular

$$\kappa_{B,\min} = \frac{\lambda_{\min}(\Sigma_u)}{d},$$

which is Eq. (16). □

C.2 Derivation of the Gaussian Hard-Edge Scaling

We derive Eq. (17). Let $k = h - d$ and $G \in \mathbb{R}^{k \times n}$ have independent standard Gaussian entries. From Eq. (14),

$$\lambda_{\min}(\Sigma_u) = \frac{2\rho^2 n}{1 + 2n} \lambda_{\min}\left(\frac{1}{n} G G^\top\right).$$

Suppose $k, n \to \infty$ with $k/n \to \gamma \in (0, 1)$. The almost-sure lower-edge limit for a Gaussian sample covariance matrix gives

$$\lambda_{\min}\left(\frac{1}{n} G G^\top\right) \longrightarrow (1 - \sqrt{\gamma})^2.$$

Since $2n/(1 + 2n) \to 1$,

$$\lambda_{\min}(\Sigma_u) \longrightarrow \rho^2 (1 - \sqrt{\gamma})^2.$$

Combining this limit with Eq. (16) yields

$$\kappa_{B,\min} \sim \frac{\rho^2}{d} (1 - \sqrt{\gamma})^2,$$

which is Eq. (17).

Near the threshold $\gamma = 1$,

$$1 - \sqrt{\gamma} = \frac{1 - \gamma}{1 + \sqrt{\gamma}} \sim \frac{1 - \gamma}{2}.$$

Hence the reciprocal local rate satisfies

$$\kappa_{B,\min}^{-1} \sim \frac{4d}{\rho^2 (1 - \gamma)^2}.$$

This is the near-threshold slowdown stated in Section 5.2.

C.3 Derivation of the Conditional Delay Scaling

We finally justify Eqs. (18)–(19) under the assumptions stated in Section 5.3.

Assumption 5.2 gives

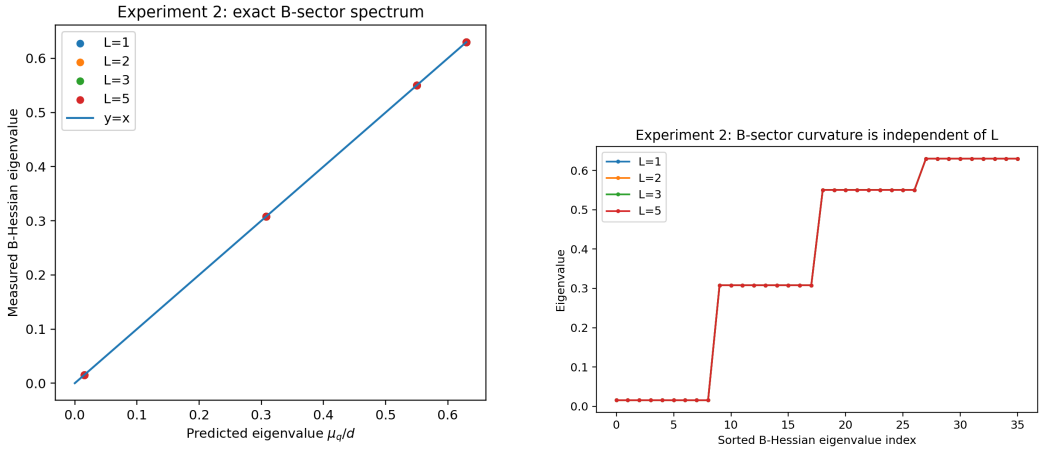
$$\kappa_\star = \frac{\lambda_{\min}(\Sigma_u)}{d} \big(1 + o(1)\big).$$

If the same mode $u_\star$ controls both evaluation errors, Corollary 4.1 gives

$$\Delta t_{L,T} = \frac{1}{2\kappa_\star} \log \frac{g_\star(T)}{m_\star(L)} + o(\kappa_\star^{-1})$$

in the corresponding joint late-time and weak-source regime. Substitution of the persistence relation gives

$$\Delta t_{L,T} = \frac{d}{2\lambda_{\min}(\Sigma_u)} \log \frac{g_\star(T)}{m_\star(L)} + o\left(\frac{1}{\lambda_{\min}(\Sigma_u)}\right),$$



(a) Predicted versus measured curvature.

(b) Independence from L.

Figure 3: **Direct numerical verification of Theorem 5.1.** (a) Numerically computed B-sector Hessian eigenvalues versus the exact prediction μ_q/d. The maximum absolute discrepancy is 2.8×10^{-16}. (b) The complete B-sector spectrum for several training horizons L; the curves coincide up to numerical precision, confirming the predicted L-independence.

which is Eq. (18).

Finally, using the high-dimensional Gaussian limit $\lambda_{\min}(\Sigma_u) \sim \rho^2(1-\sqrt{\gamma})^2$ gives

$$\Delta t_{L,T} \sim \frac{d}{2\rho^2(1-\sqrt{\gamma})^2} \log \frac{g_\star(T)}{m_\star(L)},$$

which is Eq. (19).

These final two relations are conditional consequences of Assumption 5.2 and the same-mode hypothesis. They do not strengthen the exact statement of Theorem 5.1.

D Additional Experiments

D.1 Verification of Theorem 5.1

We verify the exact curvature prediction directly in Figure 3. The B-sector Hessian is computed by automatic differentiation of the full sequence loss at R^0, independently of the closed-form expression in Theorem 5.1. Its eigenvalues agree with μ_q/d to numerical precision, and the entire spectrum is unchanged across the tested training horizons.

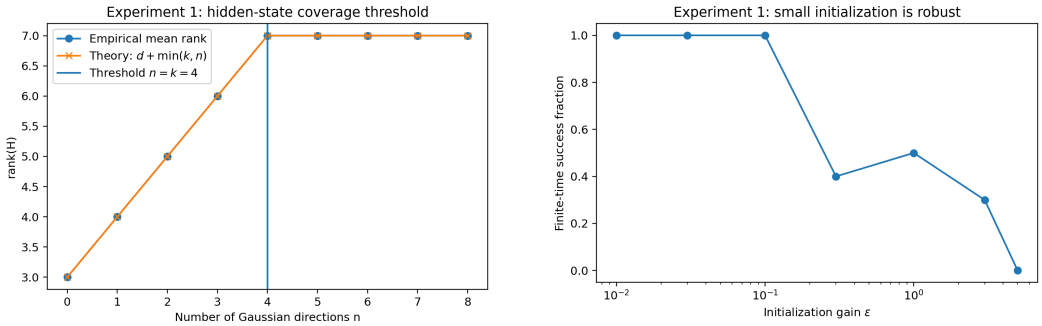
D.2 Numerical Verification of Coverage and Convergence

We numerically verify the two ingredients of the endpoint guarantee. For the coverage experiment, we use $d = 3$, $h = 7$, and hence $k = h - d = 4$, with $\rho = 1$. We vary the number of Gaussian source directions n and compute the rank of the resulting source matrix. Figure 4a shows that the empirical rank agrees with

$$\operatorname{rank}(H) = d + \min\{h - d, n\},$$

and becomes full exactly at the predicted threshold $n = h - d$. Below this threshold, zero training loss does not force the intertwining condition $OR_a = P_a O$; we construct solutions with essentially zero training loss but nonzero operator defect.

We also test the small-initialization prediction with full source coverage. Figure 4b reports the fraction of runs that reach negligible training loss, operator defect, and long-horizon error as the recurrent initialization scale is varied. The success rate is one throughout the sufficiently small-initialization regime and decreases for larger initialization, consistent with Corollary 3.1.



(a) Hidden-state coverage.

(b) Small-initialization convergence.

Figure 4: **Numerical verification of coverage and convergence.** (a) The empirical rank of the balanced source matrix agrees with $d + \min\{h - d, n\}$ and becomes full at $n = h - d$. (b) With full source coverage, sufficiently small recurrent initialization reliably converges to solutions with negligible training loss, operator defect, and long-horizon error.

D.3 Experiment 4: Actual Training Dynamics

We further test whether the local spectral mechanism developed in Sections 4.1 and 5 is realized along actual training trajectories.

We use the canonical encoder–decoder coordinates

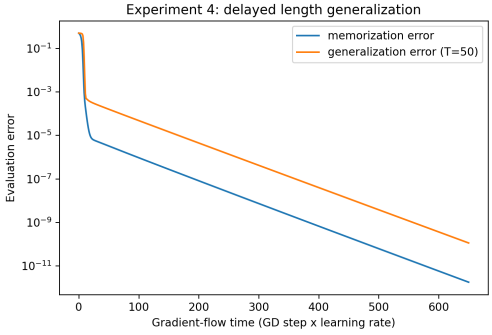
$$E = \begin{pmatrix} I_d \\ 0 \end{pmatrix}, \qquad O = (I_d \quad 0),$$

with task dimension $d = 2$, hidden dimension $h = 5$, and hence $k = h - d = 3$ decoder-invisible dimensions. The balanced decoder-consistent source pool in Eq. (7) is constructed with $n = 4$ Gaussian source directions and augmentation scale $\rho = 1$. We train at horizon $L = 3$ and evaluate long-horizon generalization at $T = 50$.

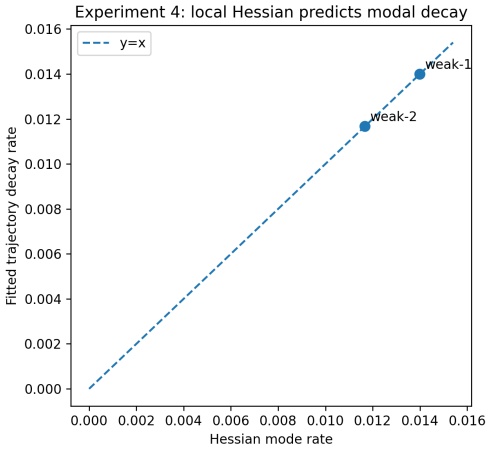
The recurrent matrices are initialized according to Eq. (4) with $\varepsilon = 0.1$. We use full-batch gradient descent with learning rate 0.1 as a numerical approximation to gradient flow. The population losses are evaluated exactly through second-moment recursion rather than Monte Carlo rollouts. We report two weakly conditioned Gaussian source-pool realizations, denoted `weak-1` and `weak-2`. The initialization is held fixed across the two runs, while the Gaussian source directions are varied.

For each trajectory, we compute the training-horizon error $\mathcal{E}^{(L)}$ and the long-horizon error $\mathcal{E}^{(T)}$ on the original reset-state distribution. At the resulting length-generalizing endpoint, we compute the Hessian of the source-pool training objective and identify the slow decoder-invisible B-dominated mode. We then compare its Hessian eigenvalue with the decay rate measured directly along the training trajectory. Finally, we evaluate the horizon observabilities of the same mode and use Corollary 4.1 to predict the late-time grokking delay.

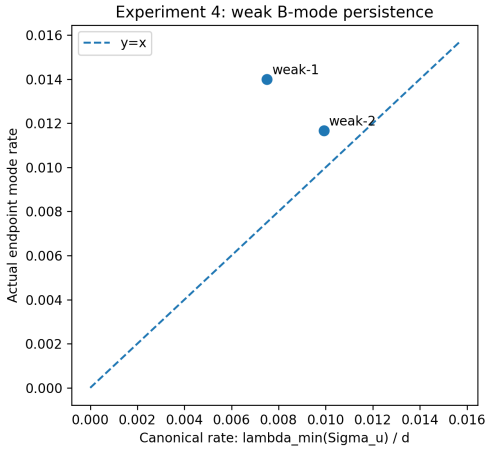
The results support the proposed mechanism at each stage. First, the training-horizon error becomes small before the long-horizon error, demonstrating the delayed emergence of length generalization. Second, the late-time decay of the selected mode is accurately described by the corresponding endpoint-Hessian eigenvalue. Third, the slow mode remains strongly concentrated in the decoder-invisible B-sector and has a rate of the same scale as the canonical prediction $\lambda_{\min}(\Sigma_u)/d$. Finally, the spectral-delay formula predicts the observed late-time delay at the correct scale.



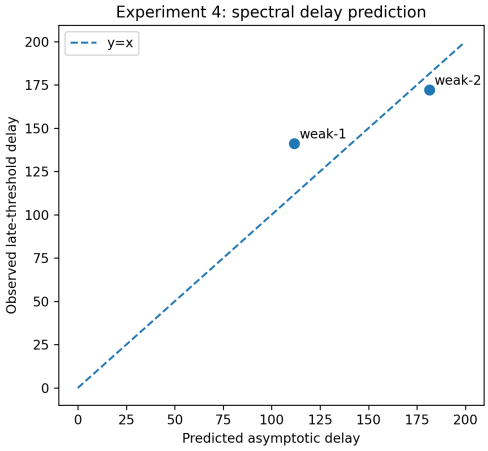
(a) Memorization and generalization errors.



(b) Local spectral decay.



(c) Weak B-mode persistence.



(d) Predicted versus observed delay.

Figure 5: **Validation of the local spectral mechanism for delayed length generalization.** We use $d = 2$, $h = 5$, $n = 4$, $\rho = 1$, training horizon $L = 3$, evaluation horizon $T = 50$, and recurrent initialization scale $\varepsilon = 0.1$. We report two weakly conditioned Gaussian source-pool realizations, `weak-1` and `weak-2`. (a) The training-horizon error decays before the long-horizon error, producing a clear memorization–generalization delay. (b) The decay rate fitted directly from the late-time trajectory closely matches the corresponding positive eigenvalue of the endpoint training Hessian, supporting the local spectral approximation in Theorem 4.1. (c) The actual endpoint slow-mode rate is compared with the canonical B-sector prediction $\lambda_{\min}(\Sigma_u)/d$. The slow modes remain strongly concentrated in the decoder-invisible B-sector, although nonzero $C_\star$ and $D_\star$ can renormalize their curvature away from the canonical value. (d) The delay predicted by $\frac{1}{2\kappa_\star} \log \frac{g_\star(T)}{m_\star(L)}$ is compared with the observed late-time hitting-time delay.