Structure, Not Size: The Sample and Parameter Complexity of State Tracking in Selective State-Space Models

Anonymous Author(s)

Editor:

Abstract

A selective state-space model (SSM) with input-dependent transitions can represent any finite-state automaton, unlike a diagonal SSM whose commuting transitions are confined to abelian dynamics. We ask what it costs to *learn* such automata, separating information-theoretic cost from what gradient descent pays. First, under uniform passive sampling the sample complexity of empirical risk minimization for tracking a K-state permutation automaton is $\Theta(K^2 \log K)$ transitions, with order-matching bounds, independent of sequence length and blind to algebraic structure. Second, a parameter-complexity separation (optimizer-free, for the general linear-transition class): a faithful $O(d_{\min})$-dimensional tracker exists for a group, whereas a random K-state automaton ($K \geq 6$) requires hidden width at least $K - 1$, which width K always attains. Third, a commuting (diagonal) SSM cannot track any non-abelian automaton at any width or sample size. Empirically, gradient descent is not structure-blind: it generalizes from sub-coverage data on group automata, below the empirical-risk baseline, while failing on random automata despite a perfectly fit training set. This holds for both a selective SSM and a non-selective gated recurrent unit (GRU), so it reflects gradient-trained recurrence rather than selective gating, consistent with the parameter-complexity separation. A proved lower bound for the gradient-descent rate on structured targets remains the open problem.

Keywords: sample complexity, parameter complexity, state tracking, state-space models, gradient descent

1. Introduction

Selective state-space models such as Mamba (Gu and Dao, 2023) make the recurrent transition depend on the input, and this single change lifts them out of the expressive class of diagonal SSMs. A diagonal SSM applies a fixed, input-independent diagonal transition, so its transitions commute and it can only realize abelian dynamics; a recent line of work makes this limitation precise in terms of circuit complexity (Merrill et al., 2024). An input-selected transition, by contrast, can apply a different permutation of the state at every step, which is exactly what a finite-state automaton does. The expressive gap is well understood. What is far less understood is the *learning* gap: given samples from an automaton, how many does a selective SSM need to recover it, how wide must the model be, and when does gradient descent actually succeed?

These questions matter because the appeal of selective recurrences is practical efficiency, and efficiency is a statement about learning, not only representation. A model that can represent an automaton but needs exponentially many samples or parameters to learn it, or that gradient descent cannot reach, is not efficient in any useful sense. We therefore study

the canonical hard case for state tracking, the word problem of a finite group, where the target transition function $\delta(s, g) = s \cdot g$ is a permutation of the state set for each input symbol g, and we contrast it with random permutation automata that have the same state count but no algebraic structure.

Our central finding is a dissociation between two notions of complexity. The information-theoretic sample complexity of the problem, what any consistent learner needs, is $\Theta(K^2 \log K)$ and is completely blind to whether the target is a structured group or a structureless random automaton. Gradient descent, however, is not blind to that structure: it beats the information-theoretic rate on groups and fails entirely on random automata. This holds not only for a selective SSM but for a non-selective gated recurrent unit (GRU) as well, which indicates it is a property of gradient-trained recurrence rather than of selective gating, at least across the two architectures we test. We trace it to a parameter-complexity fact about the hypothesis class, namely that a low-dimensional faithful solution exists for a group and does not exist for a random automaton, and we show that gradient descent finds the low-dimensional solution when one exists.

Contributions.

- **An order-optimal sample-complexity theorem (Section 3).** Under uniform passive sampling, empirical risk minimization tracks a K-state permutation automaton exactly with $\Theta(K^2 \log K)$ transitions, order-matching upper and lower bounds, independent of sequence length. The rate is structure-blind, which we illustrate with a tabular learner (identical $K^2 \log K$ scaling for cyclic and random targets, log-log slope 2.13 to 2.14).

- **A gradient-descent dichotomy (Section 4).** A gradient-trained recurrent model, both a selective SSM and a non-selective GRU control, generalizes from sub-coverage data on group automata, at an empirical rate below the empirical-risk baseline, and fails to generalize on random automata at any sample size we can train (memorization despite super-coverage data and weight decay). The dichotomy is thus a property of gradient-trained recurrence; the selective model carries a quantitative data-efficiency edge that we report as optimizer-inclusive, since the two architectures are tuned separately.

- **A parameter-complexity account (Section 5).** For the linear-transition class we prove a faithful $O(d_{\min})$-dimensional solution exists for a group while a random K-automaton ($K \geq 6$) needs width at least $K - 1$; empirically gradient descent realizes the group solution at sublinear width (the width-to-K ratio falls from 0.75 at $K = 8$ to 0.33 at $K = 24$), and random automata are unlearnable at $H \leq 2K$ for $K \geq 12$. This accounts for the dichotomy at the level of the hypothesis class, a capacity correlate consistent with the gradient-descent behavior rather than a proof about the optimizer's trajectory; the PD-SSM inherits the necessity lower bound verbatim but not the $O(d_{\min})$ existence bound (Remark 5), though a defining-representation construction tracks any automaton at width K, pinning the PD-SSM tracking width for a random target to $K - 1 \leq H \leq K$ (Corollary 7). The GRU matches the pattern empirically.

- **A finite-versus-infinite separation (Section 6).** A commuting (diagonal) SSM cannot track any non-commuting automaton at any width or sample size; on non-solvable S_5 the selective model reaches 1.000 and the diagonal model 0.125.

We are explicit about scope. The sample-complexity theorem and the parameter-complexity theorem are proved for empirical risk minimization and for the hypothesis class, respectively. The sub-coverage gradient-descent rate on groups is an empirical observation; its matching lower bound requires controlling the gradient-descent trajectory, an optimization-to-statistics question we do not resolve and state as the central open problem (Section 8).

2. Setup

The learning problem. Fix an input alphabet Σ with $|\Sigma| = K$ and a state set Q with $|Q| = K$. The target is a deterministic automaton with transition function $\delta : Q \times \Sigma \to Q$ such that for each symbol g the map $\delta(\cdot, g)$ is a permutation of Q. Two families instantiate this. A *group* automaton uses a finite group of order K with $\delta(s, g) = s \cdot g$; we use cyclic $\mathbb{Z}_K$, dihedral D_n (order $2n$, so $D_{K/2}$ has order K as used in Table 2), and symmetric S_m (order $m!$, including the non-solvable S_5). A *random* automaton draws the K permutations independently and uniformly from the symmetric group on Q, so it has no algebraic relations. In both cases the realizable hypothesis class is the set of K-tuples of permutations, $\mathcal{H} = \mathrm{Sym}(Q)^K$, with $|\mathcal{H}| = (K!)^K$ and $\log |\mathcal{H}| = K \log(K!) = \Theta(K^2 \log K)$ nats. Throughout, log denotes the natural logarithm, and all reported normalized-cost constants $m^\star/(K^2 \log K)$ use this convention.

The model. A selective SSM with hidden width H keeps a state $h \in \mathbb{R}^H$ and updates it by $h_t = M(x_t)\, h_{t-1}$ from a learned initial state $h_0 \in \mathbb{R}^H$, where the input symbol g selects the transition $M(g) = P(g)\,\mathrm{diag}(d(g))$. The transition is built from per-symbol learned parameters indexed by $g \in \Sigma$: $P(g)$ is the row-softmax of a learned logit matrix $L(g) \in \mathbb{R}^{H\times H}$, so $P(g) \in \mathbb{R}^{H\times H}$ is row-stochastic (a soft permutation), and $d(g) = \tanh(a(g)) \in (-1, 1)^H$ for a learned vector $a(g) \in \mathbb{R}^H$ is a per-channel signed scale. The initial state h_0 is learned. A linear readout $h \mapsto \arg\max(Wh + b)$ with $W \in \mathbb{R}^{K\times H}$, $b \in \mathbb{R}^K$ decodes the state to one of the K labels. In the hard limit each row of $P(g)$ is one-hot, so $M(g)$ is a signed permutation and the recurrence is an exact finite-state machine; we use a straight-through estimator for this hard variant. This permutation-diagonal (PD-SSM) parametrization follows Terzić et al. (2025), but every property we use is established in this paper, so our results do not depend on any external claim about the class. The hidden width H is a model knob and need not equal K. The diagonal baseline uses input-selected complex-diagonal transitions $M(g) = \mathrm{diag}(d(g))$ (a per-symbol unitary phase per channel); because diagonal matrices commute, the transitions commute for every choice of input. Both produce a prediction at every position; writing $\hat{s}_t$ for the model prediction at step t on input $x_{1:T}$, we score exact running-product (sequence) accuracy, the all-positions indicator

$$\mathrm{Acc}_T \;=\; \mathbb{E}_{x_{1:T}}\big[\, \mathbf{1}\big[\, \hat{s}_t = s_t \text{ for all } t \in \{1, \ldots, T\} \,\big]\,\big]\,,$$

which requires tracking the full automaton because one wrong transition corrupts the entire suffix.

Data. Each training sequence is $x_1 \dots x_T$ with x_t drawn uniformly and independently from Σ, labeled by its running states $s_t = \delta(s_{t-1}, x_t)$ with s_0 fixed to the identity (index 0). One length-T sequence supplies T labeled transition observations $(s_{t-1}, x_t) \mapsto s_t$. We count sample complexity in *transitions* m (equivalently m/T sequences), the length-independent quantity. Because the alphabet is the full symbol set, the visited-state marginal is exactly uniform for $t \geq 1$ in any finite group, and equally for a random permutation automaton since each symbol acts as a fixed bijection of the state set (one-step mixing, Appendix A; over 200,000 sampled length-16 sequences the empirical marginal deviates from uniform by at most 3.6×10^{-4} for $\mathbb{Z}_{16}$ and 1.5×10^{-4} for S_5, at the Monte-Carlo noise floor), so the (s, g) cells are hit uniformly over the K^2 universe. The empirical thresholds below use a fresh-test sequence-accuracy bar of 0.95 as a finite-sample proxy for the exact recovery that the theory addresses; at $T = 16$ this already forces a per-transition accuracy above 0.99.

Standing assumptions. The theorems of Sections 3 and 5 hold under three assumptions, made explicit here so the statements are self-contained.

(A1) *Permutation automaton.* For each symbol g the map $\delta(\cdot, g)$ is a permutation of Q (the group and random families above), and the target lies in $\mathcal{H} = \mathrm{Sym}(Q)^K$ (the realizable setting).

(A2) *Uniform passive sampling.* Symbols x_t are drawn independently and uniformly from the full alphabet Σ, with the start state fixed to the identity; the learner observes the resulting labeled transitions and does not query or adaptively choose them.

(A3) *Linear readout.* The state is decoded by a fixed linear map followed by an arg max, so two automaton states that share a hidden vector cannot be assigned different labels.

3. The sample complexity of empirical risk minimization is $\Theta(K^2 \log K)$

Theorem 1 (Upper bound) *Under assumptions (A1) and (A2), after $m = O(K^2 \log K)$ uniform transition observations any empirical risk minimizer over $\mathcal{H}$ (any hypothesis consistent with the data) recovers δ exactly with high probability, hence tracks the state for all T. Quantitatively, for any failure target $\eta \in (0, 1)$ it suffices to take $m \geq C\, K^2 \log(K/\eta)$ with an absolute constant C (depending on T only through the factor $1/(1 - 1/T)$), after which the recovery failure probability is at most η.*

Proof sketch (full proof in Appendix B). Each observation reveals one of the K^2 cells $\delta(s, g)$, and the cells are hit uniformly, so observing them is a coupon-collector process over $N = K^2$ coupons. With $m = c\, K^2 \log K$, the probability that some cell is unseen tends to 0; the rigorous argument conditions on the state walk and a Chernoff bound on the per-cell visit counts, giving the threshold constant $c > 4/(1 - 1/T)$ (about 4.27 at $T = 16$), which accounts for the first-step bias and the Chernoff splitting (Appendix B). Once every cell is seen, the only K-tuple of permutations consistent with the data is δ itself, so any empirical risk minimizer over $\mathcal{H}$ recovers it exactly. Transitions within one sequence are statistically dependent because the predecessor state depends on the history; Appendix B handles this by conditioning on the state walk and treating the symbol draws as independent, which is the only randomness coverage needs. Permutation closure (the last cell of a column is forced

once the other $K-1$ are seen) improves the constant but not the order. The threshold is on the transition count and is therefore independent of T, because permutation transitions have operator norm 1 and the recurrence accumulates no length-dependent blow-up. ■

Theorem 2 (Lower bound) *Under uniform passive sampling (A2), any learner that recovers δ exactly (equivalently, attains zero test error at every sequence length, the guarantee of Theorem 1) needs $\Omega(K^2 \log K)$ transitions.*

Proof sketch (full proof in Appendix C). We give two lower bounds; the coverage argument is the binding one. *Coverage necessity.* Working per column, at $m \leq (1-\epsilon)K^2 \log K$ draws the expected number of unseen cells in a fixed column, summed over the $K-1$ non-start predecessor states, is $(K-1)\cdot \Pr[\text{cell unseen}] = (K-1)\cdot K^{-(1-\epsilon)-o(1)} = K^{\epsilon-o(1)} = \omega(1)$, and this count concentrates because the cells of a column use disjoint predecessor states (mutually independent conditional on the state walk). So with high probability the column holds at least two unseen cells beyond the one that permutation closure can force (Appendix C). Swapping the targets of two undetermined cells in that column yields a second hypothesis in $\mathcal{H}$ consistent with all data, so no learner predicts those cells better than chance and cannot recover δ exactly; any test transition landing on a disagreeing cell (probability $\Theta(1/K^2) > 0$) is wrong with probability at least $\frac{1}{2}$, so the test error is bounded away from zero. Hence exact recovery requires $m = \Omega(K^2 \log K)$, which matches Theorem 1 in order. *Fano cross-check.* Identifying the target among $|\mathcal{H}| = (K!)^K$ hypotheses needs $\log|\mathcal{H}| = \Theta(K^2 \log K)$ nats and each transition reveals at most $\log K$ nats, which gives the weaker $\Omega(K^2)$; the additional $\log K$ factor in the binding bound is the coupon-coverage term, not the counting term. ■

Theorems 1 and 2 match in order, so the empirical-risk sample complexity is $\Theta(K^2 \log K)$; the threshold constants are not tight (the upper bound covers all cells for $c > 4/(1-1/T)$, about 4.27 at $T = 16$, while the lower bound exhibits an undetermined column for $c < 1$), so the matching is at the level of the order, not the constant. The classical coupon-collector heuristic for K^2 uniform cells, with expected cover time $K^2 H_{K^2} \approx 2K^2 \log K$, places the true threshold constant near 2, inside our [1, 4.27] window; pinning the sharp constant is an open arithmetic question we do not pursue, as our claims are at the order level. The construction never touches gradient spectra or trajectory curvature; it is combinatorics and information theory over the transition table.

Corollary 3 (Adaptive sampling removes the logarithm) *Under adaptive access, where the learner chooses which $(state, symbol)$ cell to observe next, exact recovery of a K-state* random *automaton needs $\Theta(K^2)$ transitions, against the passive $\Theta(K^2 \log K)$ of Theorems 1 and 2. The $O(K^2)$ upper bound holds for any automaton.*

Proof. The upper bound queries each of the K^2 cells once. For the lower bound, the K symbol permutations of a random automaton are drawn independently, so no observed transition constrains an unobserved cell across symbols; the only inference is within a column, where one symbol acts as a permutation, so closure fixes at most one cell per symbol (its last entry), saving only K of the K^2. Hence $\Theta(K^2)$ cells must be seen. The gap to the passive rate is exactly the coupon-collector factor $\log K$ incurred when the same cells are covered by random draws; the adaptive regime is the membership-query model of Angluin

Table 1: Tabular (empirical-risk) sample complexity $m^\star$ in transitions, median over eight seeds. The cyclic group and the random automaton scale identically; log-log slope 2.14 and 2.13, $m^\star/(K^2 \log K)$ coefficient of variation 0.139 and 0.136. This is the structure-blind $\Theta(K^2 \log K)$ baseline.

K	8	12	16	20	24	32	48	64
group $m^\star$	288	880	1680	2336	3504	6576	15840	27024
random $m^\star$	320	816	1552	2784	3504	6576	15840	27024

(1987). A group, by contrast, admits adaptive recovery in $O(K)$ by composing the columns of a generating set, a further structure-exploitation we do not pursue. ■

The rate is structure-blind. A learner that does not exploit structure must cover all cells whether the target is a group or random, so the rate is identical for both. The theorem already settles this; we add a tabular learner to exhibit the constant and to make the structure-blindness visible side by side. The learner observes m transitions, fills the cells it has seen, and is tested on whether a fresh length-16 sequence has all of its cells observed. In the realizable setting this lookup is itself an empirical risk minimizer over $\mathcal{H}$: any hypothesis consistent with the seen cells agrees with δ once all cells are covered, so the tabular threshold is the ERM threshold, not a separate quantity. We therefore read it as an illustration of Theorems 1 and 2, not as independent evidence. Across $K \in \{8, \ldots, 64\}$ (eight seeds) the threshold $m^\star$ for 0.95 fresh-test sequence accuracy scales identically for cyclic and random targets (Table 1): the log-log slope is 2.14 (group, $r^2 = 0.997$) and 2.13 (random, $r^2 = 0.998$), and the $m^\star/(K^2 \log K)$ constant is flat (coefficient of variation 0.139 and 0.136). The values coincide exactly at $K \geq 24$ because both targets fall in the same cell of the geometric measurement grid at those sizes; the point is that structure does not move the rate.

4. A general recurrence dichotomy in gradient descent

Gradient descent is not structure-blind. We measure its statistical sample complexity directly: we fix a training set of m transitions, train the soft selective SSM to convergence, and read the smallest m at which it generalizes to fresh in-distribution sequences (median sequence accuracy at least 0.95, three seeds, $T = 16$). Each choice in that definition is deliberate. We take the median over seeds because the per-seed outcome is close to bimodal (a run either reaches the tracking solution or stalls well short of it, so the mean would sit in an unoccupied middle). We use the 0.95 sequence-accuracy bar as a finite-sample proxy for the exact recovery the theory addresses (at $T = 16$ it already forces per-transition accuracy above 0.99). We fix $T = 16$ so that this proxy gap is held constant as K varies (the T-dependence of the proxy is quantified in Section 8). Using fresh-test generalization rather than training fit certifies that the transition function was learned, not memorized. We then show, with a non-selective control, that the group-versus-random split this reveals is a property of gradient-trained recurrence, not of selective gating alone, and that the next section's capacity argument explains why.

Groups: gradient descent beats empirical risk minimization, abelian and non-abelian. Table 2 reports the gradient-descent threshold $m^\star$ for the cyclic groups $\mathbb{Z}_K$ (extended to $K = 48$ over five seeds) and the non-abelian dihedral groups D_n (order $2n = K$, three seeds).[1] Both sit below the empirical-risk rate. The cleanest summary is the normalized ratio $m^\star/(K^2 \log K)$, which the structure-blind tabular learner holds constant (Table 1) but which for the gradient-trained cyclic model trends downward from 26 at $K = 8$ to 8 at $K = 48$ (non-monotone at $K = 20$): gradient descent buys a discount over the covering rate that grows with K. So gradient descent generalizes from sub-coverage data on both an abelian and a non-abelian family, exploiting the algebraic consistency of a generator's action across states; the effect is not specific to cyclic groups. We are deliberately conservative about the exponent. The $m^\star$ values are mildly non-monotone (the cyclic $K = 24$ falls below $K = 20$). The largest single deviation, the $K = 20$ point, is a measurement-grid artifact rather than a genuine outlier: a post-hoc sensitivity check at $K = 20$ on a finer n_{seq} grid with eight seeds places the 0.95 crossing near 13,600 transitions rather than 25,584, which would restore monotonicity (8,896 at $K = 16$, 13,600 at $K = 20$, 19,312 at $K = 24$). The coarse-grid value arises because the original geometric grid jumps from 13,648 to 25,584 transitions with no point between, so a true crossing in that gap is rounded upward to the upper grid point, the 25,584 reported in Table 2. Table 2 retains the uniform five-seed, coarse-grid protocol across all K so the reported numbers are produced by one measurement procedure; the finer check confirms the downward trend rather than replacing the entry. The pre-registered discriminator that asks which of $K \log K$ or $K^2 \log K$ has a flat normalized constant returns inconclusive for the gradient-descent curve: over the extended cyclic range neither normalized constant is flat (coefficient of variation 0.25 for $K \log K$ and 0.46 for $K^2 \log K$; an unconstrained log-log fit returns an exponent near 1.6 with $r^2 = 0.96$, which we report but do not read as a clean rate; the dihedral family is likewise non-monotone, $m^\star$ at $K = 32$ falling below $K = 24$, and similarly inconclusive, with a fit r^2 of 0.89 and a $K^2 \log K$ normalized-constant coefficient of variation of 0.28). The dihedral arm (three seeds) is mildly non-monotone, so we read it as confirming that the sub-baseline behavior extends to a non-abelian family rather than as independent quantitative evidence; the cyclic family (five seeds, extended to $K = 48$) carries the quantitative sub-baseline claim. We therefore claim only that the gradient-descent rate on groups is empirically below the $\Theta(K^2 \log K)$ baseline, not a specific clean exponent.

Random automata: gradient descent fails to generalize. On a random $K = 16$ automaton, with $n = 3000$ training sequences (48,000 transitions, thirty-one times the tabular threshold of 1552), hidden width swept over $2K$, $4K$, and $8K$, forty thousand steps, and weight decay 0 to 0.1, the model fits its training set (to 1.00 at width $4K$ and above) but its fresh-test sequence accuracy never exceeds 0.54 and is only 0.40 at width $8K$, far short of the 0.95 bar. Crucially the fresh-test accuracy does not rise with width: it peaks at 0.54 at width $4K$ and falls to 0.40 at width $8K$ (eight times K, far above the necessity threshold $K - 1$), so the failure is not a capacity shortfall. A cyclic group at the same budget reaches 1.00. Weight decay, which prior work has found can accelerate delayed generalization (Power et al., 2022; Nanda et al., 2023), does not resolve it here.

1. Per-seed values are recorded in the project ledger (Appendix G); the median is reported due to the bimodal per-seed outcome distribution.

Table 2: Gradient-descent sample complexity $m^\star$ in transitions (median fresh-test threshold) for the selective SSM on a cyclic family (five seeds, extended to $K = 48$) and a non-abelian dihedral family (three seeds, to $K = 32$), and for a non-selective GRU control on the cyclic family (three seeds). All three lie below the structure-blind tabular rate (slope 2.13): the selective normalized ratio $m^\star/(K^2 \log K)$ falls from 26 at $K = 8$ to 8 at $K = 48$, whereas the tabular learner holds it constant. The GRU generalizes on groups too but needs more data: at $K = 32$, the largest size both models reach, its ratio is about 19 against the selective model's about 9 (which falls further to 8 at $K = 48$, beyond the GRU sweep). This quantitative ratio is optimizer-inclusive, since the two models run at different tuned learning rates (Section 4). Values are mildly non-monotone and the rate discriminator is inconclusive, so we read these as sub-baseline, not as a clean exponent. The GRU is censored above the grid at $K \leq 12$ and was not swept past $K = 32$.

K	8	12	16	20	24	32	40	48
selective, cyclic $\mathbb{Z}_K$	3472	8528	8896	25584	19312	33200	50336	72496
selective, dihedral $D_{K/2}$	1472	5072	8896	13648	19312	16048		
GRU, cyclic $\mathbb{Z}_K$	>grid	>grid	28384	47936	37600	68640		

This failure is a property of the random class rather than of one unlucky draw: across three independently drawn random automata at each of $K = 12, 16, 20$, all nine fail to generalize at any width up to $2K$ (Section 5). So on random targets gradient descent is strictly worse than empirical risk minimization: it memorizes rather than learning δ, even though the target is information-theoretically recoverable from far less data.

The dichotomy is general to recurrence, not selectivity. To test whether the input-selected transition is what creates this split, we run the same protocol on a standard GRU. The GRU's recurrent weight matrices are fixed across steps: the input enters through the gates and a learned additive embedding but never selects the transition matrix, so the GRU has no input-selected transition in the sense the selective SSM does. At the larger sample sizes used for the $m^\star$ measurements ($K \geq 16$) the GRU, trained with gradient clipping and a tuned learning rate, fits its training set perfectly, so the generalization gap at those points reflects inductive bias and not an optimization failure. The GRU reproduces the dichotomy. It generalizes on cyclic groups (Table 2, with $m^\star$ above the grid at $K \leq 12$ and finite for $K \geq 16$) and fails on random automata at every K (best fresh-test sequence accuracy 0.11 at $K = 8$, then 0.00 for $K \geq 12$), exactly as the selective model does. A second non-selective control, an LSTM, behaves the same way: it generalizes on the cyclic groups from $K = 16$ to 32 (threshold $m^\star$ of 28,384, 47,936, 37,600, and 68,640 transitions, the same as the GRU to grid resolution and sub-baseline with log-log slope 1.1) and fails on every random automaton (best fresh-test accuracy 0.00 for $K \geq 12$, forming no state homomorphism, Table 4), so the split holds across two distinct non-selective recurrences. A vanilla Elman RNN was also tested but failed to fit the larger training sets at this budget (a familiar long-range optimization failure), so we exclude it as a non-optimizing control rather than read its failure as evidence. The group-versus-random split is therefore a property of

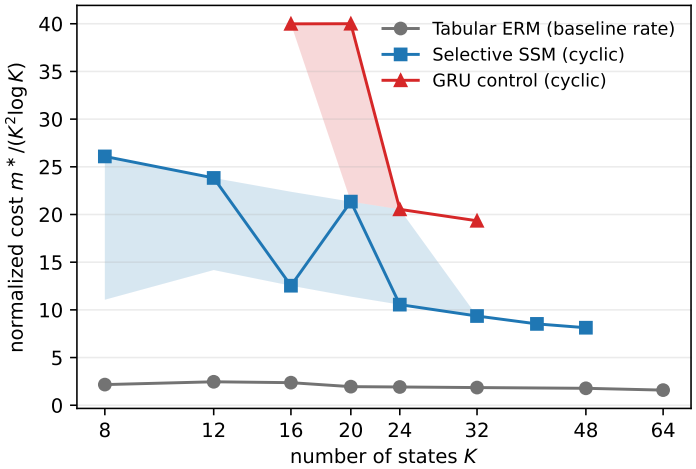

Figure 1: Normalized sample complexity $m^\star/(K^2 \log K)$ against K for the cyclic family. The tabular empirical-risk learner holds the ratio roughly flat (the $\Theta(K^2 \log K)$ baseline rate), while the gradient-trained selective SSM and the non-selective GRU both have a ratio that trends downward with K (mildly non-monotone for the GRU): their $m^\star$ grows more slowly than the baseline. The selective curve sits below the GRU curve. We report this gap as optimizer-inclusive rather than as a clean selectivity effect, because the two architectures are tuned separately (the GRU diverges at the selective model's learning rate), so the separation confounds the inductive bias with the optimizer. Absolute $m^\star$ for the neural learners exceeds the lookup-table threshold at small K; the dissociation is in the scaling, not the constant. Shaded bands are seed-bootstrap 95% intervals (resampling seeds with replacement); they are wide at small K, where seed noise is largest, and collapse onto the geometric measurement grid at large K, where the threshold crossing is stable. The downward trend and the selective-versus-GRU gap at $K = 32$ (about 9 against 19) survive the resampling.

gradient-trained recurrence, not of selective gating. Selectivity still helps quantitatively: at $K = 32$, the largest size both models reach, the selective normalized cost $m^\star/(K^2 \log K)$ is about 9 against the GRU's about 19, so the selective model learns groups from roughly half the data. We report this data-efficiency edge rather than a slope comparison, since both rate fits fall below our pre-registered precision gate. One confound bears on the size of this gap: the two models run at different optimizer settings (the selective model at learning rate 0.01 without clipping, the GRU at 0.003 with gradient clipping, because the GRU diverges at 0.01 and fails to fit its training set). The quantitative factor of two should therefore be read as optimizer-inclusive, not as a pure architectural constant. The qualitative dichotomy itself is unaffected: under each model's own tuned optimizer, both generalize sub-baseline on groups and both fail on random automata, so the split is a property of the hypothesis class rather than of any single shared learning rate. A matched-optimizer control isolates how much of the gap is architectural: rerun at the GRU's own learning rate 0.003 with the

same clipping, the selective model still reaches the threshold at fewer transitions on groups, at $K = 16$ (15,888 against the GRU's 28,384), $K = 20$ (13,648 against 47,936), and $K = 32$ (33,200 against 68,640), and ties at $K = 24$, while still failing on every random automaton. A substantial part of the data-efficiency edge therefore survives matched optimization and is not an artifact of the learning rate.

The empirical-risk curve (slope 2.13), the sub-baseline gradient-descent-on-groups curves for both the selective model and the GRU, and the random-automaton failure of both together establish the dissociation: the cost gradient descent pays is set by structure it can exploit, not by the state count.

5. A parameter-complexity account

Why can gradient descent beat the empirical-risk rate on groups but not on random automata? Because a low-dimensional faithful solution exists for a group and does not exist for a random automaton. This is a statement about the hypothesis class and needs no assumption about the optimizer. The existence is proved for the general linear-transition class (Theorem 4); the PD-SSM used in the experiments is more restrictive and reaches a faithful solution at a sublinear but larger width (Remark 5), so the mechanism accounts for the empirical split at the level of the hypothesis class while the exact attainable PD-SSM width stays an open parametrization question.

Theorem 4 (Existence for groups) *Let G be a finite group of order K with minimal faithful real representation dimension $d_{\min}$. Then in the class of selective recurrences with general linear transitions ($h_t = M(x_t)h_{t-1}$, $M(g) \in \mathbb{R}^{H\times H}$) and a linear readout, there is an exact tracker of hidden width $H = O(d_{\min})$. Here $d_{\min} = 2$ for cyclic $\mathbb{Z}_K$ and dihedral D_n, and, for the symmetric group S_m (so $K = m!$), $d_{\min} = m - 1 = \Theta(\log K/\log\log K)$ by Stirling.*

Proof sketch (full proof in Appendix D). Take a faithful real representation $\rho : G \to GL_{d_{\min}}(\mathbb{R})$ and set $M(g) = \rho(g)$, $h_0 = \rho(e)$ (the identity), so $h_t = \rho(s_t)$ tracks the state inside $\mathbb{R}^{d_{\min}}$. The K matrices $\rho(s)$ are distinct (faithfulness), so a linear readout on the flattened state, or on $h_t v$ for a generic vector v, separates the K labels. The width $O(d_{\min})$ is far below K for every group family. ■

Remark 5 (The PD-SSM parametrization is more restrictive) *Theorem 4 is about the general linear-transition class. The PD-SSM transition $M(g) = P(g)\,\mathrm{diag}(d(g))$, with $P(g)$ row-stochastic (non-negative entries) and $d(g) \in (-1, 1)^H$ a per-channel signed scale, cannot realize a pure two-dimensional rotation: a rotation matrix has the sign pattern $\left(\begin{smallmatrix} + & - \\ + & + \end{smallmatrix}\right)$, which a non-negative $P(g)$ rescaled by a single signed diagonal cannot reproduce. The same obstruction applies to dihedral groups: the standard two-dimensional representation of D_n contains these rotations together with reflections, whose matrices likewise carry negative off-diagonal entries (a reflection flips orientation), so a non-negative $P(g)$ with one signed diagonal cannot realize it either. Hence the PD-SSM does not attain $H = d_{\min} = 2$ for cyclic or dihedral groups. Empirically gradient descent reaches a faithful PD-SSM solution at a width that grows sublinearly in K, well below K (Table 3). Closing the gap between $d_{\min}$ and the PD-SSM-achievable width is a parametrization question we leave open.*

Theorem 6 (Necessity for random automata) *Let $K \geq 6$. With probability at least $1 - C/K$ over the draw of a random K-state automaton (with C an absolute constant), any exact tracker in the selective linear-transition class with a linear readout requires hidden width $H \geq K - 1$.*

Proof sketch (full proof in Appendix E). For $K \geq 6$ the K independent uniform permutations generate, with probability at least $1 - C/K$, a group Γ that is A_K or S_K (Dixon, 1969). Because A_K is simple for $K \geq 5$, every nontrivial real representation is faithful; and for $K \geq 6$ its smallest nontrivial real representation is the $(K-1)$-dimensional standard representation (Appendix E), so every faithful representation of Γ has dimension at least $K - 1$ (James and Kerber, 1981). (The single exception is $K = 5$, where A_5, the icosahedral rotation group, has a faithful 3-dimensional real representation; our experiments use $K \geq 8$.) An exact tracker with a linear readout must apply transitions $M(g)$ that act faithfully on the span of the reachable states (otherwise two states collapse and the readout cannot separate them); that action is a faithful representation of Γ and hence needs $H \geq K - 1$. There is no low-dimensional solution to find. ■

The necessity bound is matched from above inside the PD-SSM class itself.

Corollary 7 (Existence in the PD-SSM class at width K) *Every K-state permutation automaton, group or random, admits an exact tracker in the PD-SSM class with hidden width $H = K$. Take the defining permutation representation on the states: let $P(g)$ be the $K \times K$ matrix with $P(g)_{i,j} = 1$ when $\delta(j, g) = i$ and 0 otherwise, and let $d(g) = d_0 \mathbf{1}$ for any fixed $d_0 \in (0, 1)$, so $M(g) = P(g)\,\mathrm{diag}(d(g)) = d_0\, P(g)$. Each $P(g)$ has exactly one nonzero per row and per column, hence is non-negative and row-stochastic, and $d(g) \in (-1, 1)^K$, so the construction lies in the PD-SSM class. With $h_0 = e_{s_0}$, and since $P(g)\, e_s = e_{\delta(s,g)}$, the state is carried as a scaled one-hot vector, $h_t = M(x_t)\, h_{t-1} = d_0^t\, e_{s_t}$; the readout argmax is invariant to the positive scale d_0^t and recovers s_t exactly.*

Two consequences follow. For a random automaton, Theorem 6 gives $H \geq K - 1$ with high probability and the corollary gives $H \leq K$, so the PD-SSM tracking width is pinned to $K - 1 \leq H \leq K$, tight up to an additive constant. For a group, width K suffices but is far from optimal: Remark 5 rules out the $d_{\min}$-dimensional faithful representation inside the PD-SSM class, gradient descent nonetheless reaches a faithful solution at width sublinear in K (Table 3), and whether any $o(K)$ width is attainable in the PD-SSM class is the open parametrization question above.

Gradient descent realizes the gap. We measure the minimum width at which gradient descent generalizes, with the data made non-limiting (Table 3).[2] For cyclic groups the minimum width is $6, 8, 8, 10, 8$ for $K = 8, 12, 16, 20, 24$: it stays well below K, with the ratio to K falling from 0.75 to 0.33. The values are too few and too quantized by the width grid to support a precise exponent (a log-log fit gives slope 0.33 but only $r^2 = 0.60$, and the sequence is non-monotone at $K = 24$); the stable conclusion is that group width is consistently sublinear in K, not any specific power. Figure 1 shows the normalized-cost dissociation that underlies the rate claim: flat for the tabular baseline, declining for

2. Per-seed values are recorded in the project ledger (Appendix G); the median is reported due to the bimodal per-seed outcome distribution.

Table 3: Minimum hidden width H for gradient-descent generalization (median sequence accuracy at least 0.95, three seeds), with coverage made non-limiting. Group width is sublinear in K; random width is at least K and unlearnable at $H \leq 2K$ for $K \geq 12$. The random unlearnability is checked across three independently drawn automata at $K = 12, 16, 20$ (all fail); the $K = 24$ entry is a single draw from the original sweep, and additional draws at $K = 24$ are left to future work.

K	8	12	16	20	24
group min H	6	8	8	10	8
random min H	16	$> 2K$	$> 2K$	$> 2K$	$> 2K$

both gradient-trained models. For random automata the picture is sharp and stable across draws: the minimum width is $2K$ at $K = 8$, and for each of $K = 12, 16, 20$ all three independently drawn random automata fail to generalize at any width up to $2K$ (nine of nine cells unlearnable in our budget). The capacity separation is therefore sublinear versus at least linear, and it widens with K.

The achievable width, sublinear in K yet above the representation-theoretic $d_{\min} = 2$, is larger for the reason given in Remark 5: the PD-SSM parametrization cannot encode a pure rotation, so gradient descent settles at a larger but still sublinear width.

What the mechanism proves, and for which model. The argument has three levels of rigor, and we are explicit about each. (i) For the general linear-transition class, the separation is fully proved: a faithful $O(d_{\min})$-dimensional exact tracker exists for a group (Theorem 4) while a random automaton needs width at least $K-1$ (Theorem 6). (ii) The PD-SSM used in the experiments is a strict subclass (in the hard limit each $M(g) = P(g)\,\mathrm{diag}(d(g))$ is a signed permutation, hence invertible, and the necessity argument applies to its action on V unchanged), so it inherits the necessity lower bound verbatim; on the existence side it cannot reach $d_{\min}$ (Remark 5), and gradient descent realizes a sublinear-width group solution empirically (Table 3), with the exact achievable width an open parametrization question. (iii) The non-selective GRU lies outside the linear-transition class entirely, so the capacity theorems do not formally cover it; its matching group-versus-random dichotomy (Section 4) is empirical, and the capacity account is the natural conjectured explanation rather than a proved one. We did not run the minimum-width sweep for the GRU, so its capacity profile is left open. The proved content of this section is thus a class-level existence-versus-necessity separation; the experiments show both the selective SSM and the GRU behave consistently with it.

A measured representational signature. The capacity account predicts a concrete signature in the trained representation: a model should generalize on an automaton exactly when its hidden state collapses to a function of the running-product state, that is, when it discovers the state homomorphism. We measure this directly. For each trained model we group the hidden states by the true running-product state on fresh held-out prefixes and compute the alignment $A_{\text{fresh}} = \mathrm{SS}_{\text{between}}/\mathrm{SS}_{\text{total}}$, the fraction of representational variance explained by the state ($A_{\text{fresh}} \to 1$ means the representation is organized purely by group

Table 4: Homomorphism alignment versus generalization, mean over state counts $K \in \{16, 20, 24\}$ and three seeds. A_{fresh} is the fraction of fresh-prefix representational variance explained by the running-product state; cosine separation is the within-state minus between-state mean cosine. Group columns average the cyclic and non-abelian dihedral families. Every generalizing cell has high alignment; every failing cell sits near the floor. Test accuracy is the arithmetic mean over individual (K, seed) cells, so values below the per-seed median reflect the bimodal outcome distribution in which some seeds fail outright (computed from `results/exp_mechanism.json`).

	group (cyclic, dihedral)			random		
architecture	test acc	A_{fresh}	cos. sep.	test acc	A_{fresh}	cos. sep.
selective SSM	1.00	0.71	0.98	0.01	0.17	0.34
GRU	0.93	0.54	0.63	0.00	0.03	0.05
LSTM	0.93	0.49	0.50	0.00	0.07	0.04
vanilla RNN	0.00	0.06	0.06	0.00	0.04	0.02

element, $A_{\text{fresh}} \to 0$ means it is not). Across 108 cells spanning four architectures (selective SSM, GRU, LSTM, vanilla RNN), three families (cyclic, non-abelian dihedral, random), and three state counts, A_{fresh} tracks fresh-test accuracy with correlation 0.95 (Table 4); this coefficient is computed across all 108 individual cells and is driven in part by the large bimodal gap between generalizing and failing configurations rather than by within-condition variation, so we read it as corroborating the alignment-generalization link rather than as a precise effect size. Every architecture that generalizes on groups builds a clear homomorphism, while the same architectures on random automata, and the vanilla RNN everywhere, sit near the no-structure floor. This is a measurement and not a proof, but it is model-agnostic: it holds for the GRU and LSTM, which the capacity theorems do not formally cover, and it accounts for the vanilla RNN, which fails even on groups because it never forms the homomorphism. The train-fresh alignment gap is small in every cell, so the failure on random automata is the absence of any state-organized representation rather than memorization of specific prefixes.

6. Separation from diagonal state-space models

Proposition 8 *An SSM whose transitions $M(g)$ are all diagonal (for every input symbol g, including the input-selected diagonal baseline used here) generates a commutative matrix monoid and so realizes only abelian composition; it cannot represent any non-abelian automaton at any width. On any non-abelian group the task is therefore not realizable within the diagonal hypothesis class, so its sequence error cannot be driven to zero at any sample size: the limitation is representational, not statistical.*

The proof (Appendix F) is immediate from commutativity. Diagonal matrices commute irrespective of which one each input selects, so the monoid generated by $\{M(g)\}$ is com-

mutative and can realize only abelian compositions; no non-abelian group's word problem is in its range at any width, a fact corroborated from the circuit-complexity side by Merrill et al. (2024). We instantiate the finite-versus-infinite gap on the non-solvable group S_5 ($K = 120$): the selective model reaches median next-state accuracy 1.000 while the diagonal model reaches 0.125 (against a chance level of $1/120 \approx 0.008$; gap 0.875, six seeds, per-token accuracy under online training; the diagonal model cannot fit the data even on its own training distribution, the signature of a representational rather than statistical limit), and on S_3 ($K = 6$, chance $1/6$) the selective model reaches 1.000 versus diagonal 0.374. This separation is along a different axis from the dichotomy of Sections 4 and 5: the diagonal limit is abelian-only, whereas the selective model learns non-abelian groups (including S_5) without difficulty and fails only on structureless random automata. The two divisions, commuting versus selective and structured versus random, are distinct.

7. Related work

Expressivity of state-space models. The circuit-complexity limits of diagonal and fixed-transition SSMs are well established (Merrill et al., 2024; Gu and Dao, 2023), and selective or non-diagonal recurrences are known to recover finite-state expressivity (Terzić et al., 2025), and chain-of-thought lets transformers represent finite-state automata by externalizing intermediate state (Zhang et al., 2025). The correspondence between recurrent networks and automata classes is itself classical: saturated recurrent networks map onto specific automata families (Merrill, 2019), the formal languages that transformers and recurrences can recognize have been charted empirically (Bhattamishra et al., 2020), and programmatic accounts such as RASP make explicit the computation a transformer encoder can express (Weiss et al., 2021). These results concern representation; our contribution is the learning side, sample and parameter complexity, which they do not address.

Learning algebraic structure and grokking. Delayed generalization (grokking) on algorithmic and group-composition tasks was first reported empirically by Power et al. (2022); later work tied it, for two-layer networks, to the emergence of spectral or representational features (Nanda et al., 2023), including the preprint of He et al. (2026), with provable scaling laws for feature emergence on group-arithmetic targets (Tian, 2025), and showed mechanistically that networks learning a group operation implement it through the group's irreducible representations (Chughtai et al., 2023). A complementary optimization account addresses which functions gradient descent learns first: the leap (or staircase) complexity of a target governs the order and timing of SGD learning on neural networks (Abbe et al., 2023), so an algebraically structured target can be reached before an unstructured one. That line is feedforward and concerns which features emerge during training on a single group; our setting is recurrent state tracking, and our parameter-complexity theorem gives a representation-dimension account that holds at the level of the hypothesis class, independent of the optimizer. Two points distinguish it from the spectral grokking account (He et al., 2026): it covers the recurrent transition dynamics rather than a feedforward composition, and it supplies the matching random-automaton lower bound $H \geq K - 1$, which is what separates structured from unstructured targets and which a single-group spectral analysis does not address.

Learning finite automata. The learnability of finite automata is classical: exact identification from queries via Angluin's L^* algorithm (Angluin, 1987), identification in the limit (Gold, 1967), and the computational hardness of finding a minimum consistent deterministic finite automaton from passive examples (Pitt and Warmuth, 1993). Those results concern general DFAs and worst-case or query-based learning. Closer to the present setting, Giapitzakis et al. (2025) prove a statistical-query hardness result for learning semiautomata, derived from the internal state-transition structure via random walks on symmetric groups. That bound is, however, blind to algebraic structure in the sense that matters here: it does not distinguish the group automata that gradient descent learns from sub-coverage data, and its consequence for gradient-based learning runs through the statistical-query reduction of Abbe et al. (2023) rather than through plain gradient descent. The separation we observe, and the plain-gradient-descent lower bound that would explain it, lie outside that account. Our setting is the realizable, passively sampled, permutation (group and random) case, where the relevant quantity is cell coverage rather than VC dimension, and the contribution is the *tight two-sided* $\Theta(K^2 \log K)$ *rate* together with its dissociation from what gradient descent on a selective recurrence actually pays. Classical sample-complexity bounds for recurrent networks are VC-based and scale with the weight count and the input length (Koiran and Sontag, 1998); our coverage bound is instead combinatorial in the automaton and independent of length. A complementary line studies in-context rather than weight-learned acquisition of sequential structure, such as variable-order Markov chains by transformers (Zhou et al., 2024), which is the setting of the in-context extension we leave to future work. Coverage arguments of coupon-collector type for learning automata are themselves classical; what is new here, to our knowledge, is the combination of (i) the tight two-sided $\Theta(K^2 \log K)$ rate specialized to permutation automata with explicit constants, (ii) its dissociation from the cost that gradient descent on a selective recurrence actually pays, and (iii) the matching parameter-complexity separation, a faithful $O(d_{\min})$-dimensional solution for a group against $H \geq K - 1$ for a random automaton, that accounts for that dissociation.

8. Limitations and the open problem

We are explicit about what is proved and what is observed. Theorems 1 and 2 hold for empirical risk minimization (any consistent learner), and Theorems 4 and 6 hold for the hypothesis class. The sub-baseline rate of gradient descent on groups (Section 4) is an empirical observation, and we do not claim a clean exponent for it. Its matching lower bound, a proof that gradient descent needs at least this many samples, requires controlling the gradient-descent trajectory and is the central open problem; it is the same optimization-to-statistics obstruction that separates representation from reachability throughout this work. Concretely, a proof that gradient descent requires $\Omega(K^\alpha \log K)$ transitions for some $\alpha < 2$ on group automata would constitute this bound. What is known stops short of it on two counts. The statistical-query hardness of learning random semiautomata is established (Giapitzakis et al., 2025), but it is structure-blind, applying equally to the group automata that gradient descent learns sub-coverage, so it does not deliver the structured versus unstructured *separation*; and the bridge from a statistical-query bound to gradient descent currently passes through the large-batch, bounded-precision reduction of Abbe et al. (2023) rather

than plain gradient descent. Closing both, a plain-gradient-descent sample lower bound that distinguishes group from random targets, is the open problem. The adaptive-sampling rate for random targets is by contrast settled (Corollary 3: $\Theta(K^2)$, the coupon-collector logarithm removed). We leave open the extension to non-uniform passive distributions, to deep selective stacks, and the parametrization question of closing the gap between $d_{\min}$ and the gradient-descent-achievable width (Corollary 7 pins the PD-SSM width to $K-1 \leq H \leq K$ but the sublinear-width regime that gradient descent occupies is unproved).

Soft versus hard model. The statistical sample- and parameter-complexity sweeps use the soft selective model, which trains stably across the full range of K, while the theorems concern hard permutations and the diagonal separation uses the hard straight-through model. To check that the soft relaxation does not manufacture the sub-baseline group result, we re-measured $m^\star$ on cyclic groups with the hard straight-through model at $K = 12$ and $K = 20$. At $K = 12$ it reaches the bar at 1808 transitions, below the soft model and only about twice the empirical-risk coverage threshold; at $K = 20$ it reaches the bar at 25584 transitions, matching the soft model, though at this size the straight-through estimator is near its optimization budget and the agreement may be budget-limited. We do not report a hard-model exponent, because the straight-through estimator becomes optimization-fragile at larger K (the reason the main sweeps use the soft model); the qualitative phenomenon, sub-baseline generalization on groups, is present in both models.

Sequence length. The sample-complexity theorem is independent of T (Appendix B): the coverage threshold is a count of transitions and does not depend on sequence length. The *empirical* threshold $m^\star$ we measure is a distinct quantity: it is read off a fixed *sequence*-accuracy bar of 0.95, which tightens the per-transition requirement as T grows, so it upper-bounds the T-independent coverage threshold of the theorem rather than equaling it. Measured this way $m^\star$ rises with T (about 7.8k, 8.9k, and 16.9k transitions at $T = 8, 16, 32$ for $K = 16$, coefficient of variation 0.36 across the three lengths), roughly doubling from $T = 16$ to $T = 32$. This is a property of the proxy metric, not a contradiction of the theorem; a per-transition bar would remove it. The cross-K slope and coverage-ratio claims we report are all at the single fixed $T = 16$, where this proxy gap is held constant.

Pre-committed confounds. The visited-cell distribution is made uniform by the full-symbol alphabet (verified, Section 2); the start state is fixed to the identity; and the necessity result, being probabilistic over the random-automaton draw, is checked across three independent draws per K (Section 5).

9. Conclusion

Two notions of complexity come apart for gradient-trained recurrences. The information-theoretic sample complexity of tracking a K-state automaton is $\Theta(K^2 \log K)$ and depends only on K, blind to whether the target is a structured group or a structureless random automaton. The cost gradient descent actually pays does not: it falls below the baseline on groups, abelian and non-abelian alike, and diverges on random automata, because a faithful low-dimensional solution exists for a group and not for a random automaton. We prove this capacity separation for the selective linear-transition class; the same split appears empirically for a non-selective GRU, which the capacity account explains by analogy rather

than by proof, so the dichotomy reads as a property of gradient-trained recurrence across the two architectures we test. Selectivity contributes a quantitative data-efficiency advantage (normalized cost ≈ 9 for the selective model versus ≈ 19 for the GRU at $K = 32$, optimizer-inclusive) and is what makes the diagonal separation possible. Gradient-trained recurrences are structure-adaptive learners, and quantifying that adaptivity, including the open lower bound for the gradient-descent rate, is a precise open program.

Code and data availability. The model implementations, the procedural generators for the group and random automata, and the scripts that produce every table and figure in this paper are available from the author on request and will be released as a public repository upon publication. Each reported cell is reproducible from its configuration triple as described in Appendix G.

Appendix A. One-step uniform mixing

Lemma 9 (Exact uniformity) *For a group automaton of order K driven by symbols x_t drawn independently and uniformly from the full alphabet $\Sigma = G$, the visited state $s_t = s_0 \cdot x_1 \cdots x_t$ is exactly uniform on Q for every $t \geq 1$, and is independent of x_{t+1}.*

Proof. For $t \geq 1$, $s_t = s_{t-1} \cdot x_t$. If x_t is uniform on G and independent of s_{t-1}, then for any fixed $s_{t-1} = a$ the product $a \cdot x_t$ is uniform on G (left multiplication by a is a bijection), so s_t is uniform regardless of the law of s_{t-1}. Independence of s_t and x_{t+1} holds because x_{t+1} is drawn independently of the history. ■

Consequently a transition observed at step $t \geq 2$ lands on cell (s_{t-1}, x_t) with s_{t-1} uniform on Q and x_t uniform on Σ and independent of it, so it hits each of the K^2 cells with probability exactly $1/K^2$. The first step of each sequence ($t = 1$) is the only exception: its predecessor is the fixed start state $s_0 = e$, so it always lands in the identity row. These first steps are a $1/T$ fraction of all transitions, and they only add coverage to the identity row; we absorb their effect into the constants below (Appendix B), where it changes $\mathbb{E}[N_c]$ by a factor $1 - 1/T = 1 - o(1)$ and never reduces coverage. The same one-step argument applies verbatim to a random permutation automaton: each symbol g acts as a fixed bijection of Q, so if x_t is uniform and independent of s_{t-1} then s_t is uniform for $t \geq 1$, and the K^2 cells are again hit uniformly. Uniform cell coverage therefore holds for both the structured and the random targets.

Appendix B. Proof of Theorem 1 (upper bound)

We are in the realizable setting with hypothesis class $\mathcal{H} = \mathrm{Sym}(Q)^K$. Index the m observed transitions $t = 1, \ldots, m$ (concatenating sequences), and fix a cell $c = (s^*, g^*)$. Cell c is hit at step t iff $s_{t-1} = s^*$ and $x_t = g^*$. Let $N_c = |\{t : s_{t-1} = s^*\}|$ be the number of steps whose predecessor state is s^*. Since the symbols x_t are independent of the state walk and uniform, conditioned on the walk the events $\{x_t = g^*\}$ at the N_c relevant steps are independent Bernoulli$(1/K)$, so

$$\Pr[c \text{ unseen} \mid \text{walk}] = (1 - 1/K)^{N_c}, \qquad \Pr[c \text{ unseen}] = \mathbb{E}\big[(1 - 1/K)^{N_c}\big].$$

By Lemma 9 every step with $t \geq 2$ has $s_{t-1} = s^*$ with probability exactly $1/K$; the one first step per sequence has the fixed predecessor $s_0 = e$, so for $s^* \neq e$ it never contributes and

$\mathbb{E}[N_c] = (1 - 1/T)\, m/K = (1 - o(1))\, m/K$, while for $s^* = e$ it only adds visits. In either case $\mathbb{E}[N_c] \geq (1 - 1/T)\, m/K$, and the constant $1 - 1/T$ is absorbed into c below. Writing $\mu = \mathbb{E}[N_c] \geq (1 - 1/T)\, m/K$ and splitting on whether N_c is at least half its mean,

$$\Pr[c \text{ unseen}] \leq \Pr\left[N_c < \tfrac{\mu}{2}\right] + (1 - 1/K)^{\mu/2} \leq \Pr\left[N_c < \tfrac{\mu}{2}\right] + e^{-\mu/(2K)}.$$

The visit count N_c is a sum of per-sequence visit counts that are independent across sequences, each bounded in $[0, T-1]$ with mean at least $(T-1)/K$; rescaling each summand by $T-1$ gives a sum of independent $[0,1]$-bounded variables with total mean $\mu/(T-1)$, to which the multiplicative Chernoff lower tail (deviation $\frac{1}{2}$, Mitzenmacher and Upfal, 2017, Theorem 4.5) applies, yielding $\Pr[N_c < \mu/2] \leq e^{-\mu/(8(T-1))}$. We keep the two terms separate. By a union bound over the K^2 cells,

$$\mathbb{E}[\#\text{unseen cells}] \leq K^2 \left(e^{-\mu/(8(T-1))} + e^{-\mu/(2K)}\right).$$

For $m = c\, K^2 \log K$ and $\mu \geq (1 - 1/T)\, m/K$, the coverage term is $K^2 e^{-(1-1/T)\, m/(2K^2)} = K^{2-c(1-1/T)/2}$, which $\to 0$ once $c > 4/(1 - 1/T)$ (for $T = 16$, $c > 4.27$). The Chernoff term is $K^2 e^{-\mu/(8(T-1))} \leq K^2 e^{-c\, K \log K/(8T)}$ (using $(1 - 1/T)/(T-1) = 1/T$), which $\to 0$ super-polynomially in K for any fixed T and so is dominated by the coverage term. Both vanish, so with high probability every cell is observed. To make the rate explicit, Markov's inequality on the nonnegative integer count of unseen cells gives $\Pr[\text{some cell unseen}] \leq \mathbb{E}[\#\text{unseen cells}] \leq K^{2-c(1-1/T)/2} + K^2 e^{-cK \log K/(8T)}$; taking $c \geq (4 + 2\log(1/\eta)/\log K)/(1 - 1/T)$, equivalently $m \geq C\, K^2 \log(K/\eta)$ for an absolute constant C, drives this below any target $\eta \in (0,1)$, which is the quantitative form stated in Theorem 1. Once all cells are observed, the data determine $\delta(s,g)$ for every (s,g), so the only element of $\mathcal{H}$ consistent with the data is δ; every empirical risk minimizer returns it and tracks exactly for all T. The bound is in transitions m with no dependence on T, and because permutation transitions have operator norm 1 the recurrence carries no length-dependent error growth. ■

Appendix C. Proof of Theorem 2 (lower bound)

We instantiate the lower bound with a group automaton (for concreteness the cyclic group $\mathbb{Z}_K$), for which Lemma 9 gives that each cell is hit with probability exactly $1/K^2$ at every step. The lower bound is over the hypothesis class $\mathcal{H} = \mathrm{Sym}(Q)^K$, so a single hard instance suffices.

Coverage necessity. Fix a column g and take $m \leq (1 - \epsilon) K^2 \log K$ for a constant $\epsilon \in (0,1)$. Cell (s,g) is hit at step t iff $s_{t-1} = s$ and $x_t = g$; conditioned on the state walk, the per-cell miss events use the independent symbol draws, and as in Appendix B $\Pr[(s,g) \text{ unseen}] = \mathbb{E}[(1 - 1/K)^{N_{(s,g)}}]$. The map $x \mapsto (1 - 1/K)^x$ is convex, so Jensen's inequality gives $\mathbb{E}[(1 - 1/K)^{N_{(s,g)}}] \geq (1 - 1/K)^{\mathbb{E}[N_{(s,g)}]}$. For any predecessor state $s \neq s_0$ (the start state) the fixed-start-state bias of Appendix B only lowers the visit count, so $\mathbb{E}[N_{(s,g)}] \leq m/K$; the start state s_0 is the lone exception, where the extra first-step visit gives $\mathbb{E}[N_{(s_0,g)}] > m/K$, and we simply drop that single cell from the count below. Since the base is below 1, for $s \neq s_0$ we get $(1 - 1/K)^{\mathbb{E}[N_{(s,g)}]} \geq (1 - 1/K)^{m/K}$. With $\ln(1 - 1/K) \geq -1/(K-1)$ this yields $\Pr[(s,g) \text{ unseen}] \geq (1 - 1/K)^{m/K} \geq e^{-m/(K(K-1))} \geq K^{-(1-\epsilon)\, K/(K-1)} = K^{-(1-\epsilon)-o(1)}$. Summing over the $K - 1$ non-start cells of column g,

$$\mathbb{E}[\#\text{unseen cells in column } g] \geq (K-1) \cdot K^{-(1-\epsilon)-o(1)} = K^{\epsilon - o(1)} = \omega(1).$$

Within a column the K cells correspond to disjoint predecessor states. Condition on the state walk, that is on the visit counts N_s to each state s across the sample. The bound just derived on $\mathbb{E}[\#\text{unseen cells in column } g]$ averages over the walk, so before applying Chebyshev we pin the conditional expectation down on a high-probability set of walks. For each non-start state $s \neq s_0$ the count N_s is a sum of per-sequence visit counts, independent across the m/T sequences and bounded in $[0, T-1]$, with mean at most m/K; the multiplicative Chernoff upper tail gives $\Pr[N_s > (1+\delta)\, m/K] \leq e^{-\delta^2 m/(3KT)}$ for any $\delta \in (0,1)$, which with $m \leq (1-\epsilon)K^2 \log K$ is $e^{-\Omega(K \log K/T)}$, hence $K^{-\Omega(K)}$ for fixed T (the regime of the upper bound and of every experiment here) and $o(1/K)$ whenever $T = o(K)$; so by a union bound over the K states there is an event $\mathcal{W}$ of walks with $\Pr[\mathcal{W}] \geq 1 - o(1)$ on which every non-start state has $N_s \leq (1+\delta)\, m/K$. On $\mathcal{W}$ the per-cell miss probability obeys $(1-1/K)^{N_s} \geq (1-1/K)^{(1+\delta)m/K} = K^{-(1-\epsilon)(1+\delta)-o(1)}$, so the conditional expectation $\nu = \mathbb{E}[U \mid \text{walk}]$ of the count U of unseen non-start cells in the column satisfies $\nu \geq (K-1)\, K^{-(1-\epsilon)(1+\delta)-o(1)} = K^{\epsilon-(1-\epsilon)\delta-o(1)}$, which is $\omega(1)$ once $\delta < \epsilon/(1-\epsilon)$. Given the walk, whether cell (s, g) is seen depends only on the symbols drawn at the N_s steps whose predecessor is s, and these symbol sets are disjoint across distinct s and drawn independently; the K column-g miss-indicators are therefore mutually independent conditional on the walk. The conditional variance of U, a sum of independent indicators, satisfies $\mathrm{Var}[U \mid \text{walk}] \leq \nu$, so on $\mathcal{W}$ Chebyshev gives $\Pr[U < 2 \mid \text{walk}] \leq \nu/(\nu-2)^2 \to 0$ as $K \to \infty$. Since $\Pr[\mathcal{W}] \to 1$, taking expectations over the walk gives that with high probability column g has at least two unseen cells. Permutation closure determines a cell only when exactly one cell of the column remains unseen (its target is then the unique missing element); with at least two unseen cells, none is forced, so both are undetermined. Let (s_1, g) and (s_2, g) be two such undetermined cells. Define δ' equal to any data-consistent δ except with the images of s_1 and s_2 under column g swapped. Then $\delta' \in \mathcal{H}$ (column g is still a permutation), δ' agrees with all observed data, and $\delta' \neq \delta$ on these cells. No learner distinguishes δ from δ', so under the symmetric prior over the two it errs on each disagreeing cell with probability $\frac{1}{2}$. The learner therefore cannot recover δ exactly: a test transition that lands on one of these cells occurs with probability $\Theta(1/K^2) > 0$ and is wrong with probability at least $\frac{1}{2}$, so the expected test error is bounded strictly away from zero (and a sufficiently long test sequence is wrong with probability tending to 1). Exact recovery, the zero-error guarantee Theorem 1 delivers, thus requires every cell to be determined, which forces $m = \Omega(K^2 \log K)$, matching Theorem 1 in order. (The constants differ: the upper bound covers all cells for $c > 4/(1-1/T)$ and the lower bound exhibits an undetermined column for $c < 1$; the order $\Theta(K^2 \log K)$ is what we claim.)

Fano cross-check. Identifying the target among $|\mathcal{H}| = (K!)^K$ equiprobable hypotheses requires $\log|\mathcal{H}| = K \log(K!) = \Theta(K^2 \log K)$ nats of information. Each transition observation is one of K outcomes, carrying at most $\log K$ nats, so at least $\Theta(K^2 \log K)/\log K = \Omega(K^2)$ transitions are required. This is weaker than the coverage bound by a $\log K$ factor; the additional factor in the binding bound is the coupon-coverage term, not the information-counting term. ■

Appendix D. Proof of Theorem 4 (existence for groups)

Let $\rho : G \to GL_d(\mathbb{R})$ be a faithful real representation of dimension $d = d_{\min}$. Set the hidden state to the flattened matrix $h_t = \mathrm{vec}(\rho(s_t)) \in \mathbb{R}^{d^2}$ (or, more economically, $h_t = \rho(s_t)v \in \mathbb{R}^d$ for a vector v in general position), with transition $M(g)$ acting as left multiplication by $\rho(g)$ and $h_0 = \mathrm{vec}(\rho(e))$. Then h_t encodes $\rho(s_t)$, and since ρ is faithful the K matrices $\{\rho(s) : s \in G\}$ are distinct; a generic vector v separates them, so the K states map to K distinct points and a linear readout recovers the label. The width is $O(d_{\min})$.

The representation dimensions are standard. Cyclic $\mathbb{Z}_K$ has a faithful real representation of dimension 2 (rotation by $2\pi/K$); 1 is impossible because the only faithful real 1-dimensional representations are by ± 1. Dihedral D_n has the faithful 2-dimensional representation by the symmetries of the regular n-gon, so $d_{\min} = 2$. For the symmetric group S_m the minimal faithful representation is the $(m-1)$-dimensional standard representation, so $d_{\min} = m - 1$; writing the order $K = m!$ and applying Stirling, $\log K = \log m! = m \log m\,(1 + o(1))$, hence $m = \Theta(\log K/\log\log K)$ and $d_{\min} = m - 1 = \Theta(\log K/\log\log K)$. In all cases $d_{\min} = o(K)$. ■

Appendix E. Proof of Theorem 6 (necessity for random automata)

Fix $K \geq 6$. Let the random automaton be generated by K permutations drawn independently and uniformly from $\mathrm{Sym}(Q)$. By Dixon (1969), two uniform random permutations already generate A_K or S_K with probability at least $1 - C/K$ for an absolute constant C (Dixon's bound, which tends to 1 as $K \to \infty$). Adding more independent uniform generators only enlarges the generated subgroup, so with $K \geq 2$ symbol permutations the generated group Γ contains A_K with at least the Dixon probability; hence $\Gamma \supseteq A_K$ with probability at least $1 - C/K$. Since $\Gamma \subseteq \mathrm{Sym}(Q) = S_K$ and A_K has index 2 in S_K, the only subgroups of S_K containing A_K are A_K and S_K themselves, so on this event $\Gamma \in \{A_K, S_K\}$.

We need the minimal faithful real dimension of Γ. We first establish the complex minimal degree and then pass to real dimensions, since the model is real-valued. The $(K-1)$-dimensional standard representation of S_K (the sum-zero subspace of the permutation action on $\mathbb{R}^K$) is irreducible (James and Kerber, 1981). For $K \geq 5$ the group A_K acts 2-transitively on $\{1, \ldots, K\}$, and a 2-transitive permutation representation decomposes as the trivial representation plus a single irreducible complement; hence the standard representation stays irreducible on restriction to A_K, giving a $(K-1)$-dimensional irreducible of A_K. For $K \geq 5$ the alternating group A_K is simple, so any representation has trivial or full kernel; that is, every nontrivial representation is faithful.

For the matching lower bound on the degree we use the classical minimal-degree classification. The minimal degree of a nontrivial irreducible character of S_K is $K-1$ (the standard representation) by Rasala (Rasala, 1977). For A_K the corresponding minimal nontrivial degree is also $K - 1$ for every $K \geq 6$. This follows from Rasala's symmetric-group degrees through Clifford theory for the index-two subgroup $A_K \triangleleft S_K$: every irreducible of A_K is a constituent of the restriction of an irreducible of S_K, an irreducible of S_K that does not split on restriction contributes a constituent of the same degree (at least $K-1$ unless trivial or sign), and the self-conjugate partitions whose S_K-irreducibles do split contribute half-degree constituents whose degree still exceeds $K-1$ for $K \geq 6$ (the smallest, from $(3, 2, 1)$ at

$K = 6$, has degree $8 > 5$). For the moderate $K \leq 24$ used in our experiments the value $K-1$ is in any case immediate from the published character tables. The lone exception is $K = 5$, where A_5, the icosahedral rotation group, has two faithful 3-dimensional representations, so its minimal faithful dimension is 3 rather than $K - 1 = 4$; our experiments use $K \geq 8$. For S_K the only irreducibles below dimension $K - 1$ are the trivial and sign representations, both unfaithful (their kernels contain A_K). Hence for $K \geq 6$ every nontrivial irreducible representation of A_K, and every faithful representation of S_K, has complex degree at least $K - 1$.

We now pass to real dimensions. The standard representation is rational, hence realizable over $\mathbb{R}$ with Frobenius-Schur indicator $+1$, so its real dimension equals its complex degree $K-1$ and the existence half above is real. Conversely, let W be any faithful real representation of Γ. Its restriction to $A_K \subseteq \Gamma$ is again faithful, hence (by simplicity) nontrivial, so the complexification $W \otimes \mathbb{C}$ contains a nontrivial irreducible constituent of A_K, of complex degree at least $K - 1$ by the previous paragraph. Since $\dim_{\mathbb{R}} W = \dim_{\mathbb{C}}(W \otimes \mathbb{C}) \geq K - 1$, every faithful real representation of Γ has real dimension at least $K - 1$. Since Γ is A_K or S_K, its minimal faithful real dimension is exactly $K - 1$ for $K \geq 6$.

Now suppose a selective recurrence with hidden width H and a linear readout tracks the automaton exactly. Exact tracking means the map $s \mapsto h(s)$ from automaton state to reachable hidden state is injective (otherwise two states share a hidden vector and the linear readout cannot separate their labels), and it intertwines the automaton action with the linear operators $M(g)$: $h(\delta(s, g)) = M(g)\, h(s)$. Let $V \subseteq \mathbb{R}^H$ be the span of the reachable states $\{h(s) : s \in Q\}$ (for a permutation automaton every state is reachable). Each $M(g)$ maps V into V, since $M(g)\, h(s) = h(\delta(s, g))$ and $\delta(s, g) \in Q$, so V is $M(g)$-invariant. Moreover, since $\delta(\cdot, g)$ is a permutation of Q and $s \mapsto h(s)$ is injective, the map $h(s) \mapsto h(\delta(s, g))$ sends the spanning set $\{h(s) : s \in Q\}$ of V bijectively onto itself, so $M(g)|_V$ is a surjective (hence, V being finite-dimensional, bijective) linear operator and lies in $GL(V)$. The induced map $\phi : \Gamma \to GL(V)$ sends a group element γ to the product of the $M(g)|_V$ along any word representing γ; this is well defined because, by the intertwining relation, any such product acts on the spanning set as $h(s) \mapsto h(\gamma \cdot s)$, which depends only on γ and not on the chosen word. Thus $\phi(\gamma)\, h(s) = h(\gamma \cdot s)$, hence $\phi(\gamma_1\gamma_2) = \phi(\gamma_1)\phi(\gamma_2)$: ϕ is a group homomorphism, a representation of Γ on V. It is faithful: if $\phi(\gamma)$ acts as the identity on V then $h(\gamma \cdot s) = h(s)$ for all s, and injectivity of $s \mapsto h(s)$ forces $\gamma \cdot s = s$ for all s, so γ is the identity of Γ. For $K \geq 6$ a faithful representation of Γ has dimension at least $K - 1$, so $\dim V \geq K - 1$ and $H \geq K - 1$. By Dixon's theorem the draw yields $\Gamma \in \{A_K, S_K\}$ with probability at least $1 - C/K$, so with at least this probability there is no exact tracker of width below $K - 1$. ■

Appendix F. Proof of Proposition 8 (diagonal models cannot track non-abelian automata)

Let the transitions $M(g) \in \mathbb{R}^{H \times H}$ be diagonal for every symbol g. Any product $M(g_{i_1}) \cdots M(g_{i_n})$ is again diagonal, and diagonal matrices commute, so the monoid $\mathcal{M} = \langle M(g) : g \in \Sigma \rangle$ generated by the transitions is commutative.

Suppose, for contradiction, that such a model tracks an automaton with transition group Γ exactly, with an injective state embedding $s \mapsto h(s)$ and a linear readout. Exact tracking

requires $h(\delta(s,g)) = M(g)\,h(s)$, so on the span V of the reachable states the operators $\{M(g)|_V\}$ realize the action of Γ (as in Appendix E, the restriction is invertible and the induced map $\phi : \Gamma \to GL(V)$ is a homomorphism). Because the $M(g)$ commute, the operators $\phi(g)$ commute, so $\phi(\Gamma)$ is an abelian subgroup of $GL(V)$. The map ϕ is faithful (injectivity of $s \mapsto h(s)$, again as in Appendix E), so $\Gamma \cong \phi(\Gamma)$ is abelian. Contradiction whenever Γ is non-abelian. Hence no diagonal model of any width H tracks a non-abelian automaton exactly; the obstruction is the commutativity of the realized monoid and is independent of H and of the sample size. ■

Appendix G. Experimental details

All experiments run in MLX (version 0.31.2) on a single Apple M5 Max machine. The soft PD-SSM uses row-softmax $P(g)$ and $d(g) = \tanh(\cdot) \in (-1, 1)$; the hard variant uses a straight-through estimator with the inverse temperature annealed from 1 to 8 over training. Optimizer Adam, learning rate 0.01, batch size 64. Seeds are contiguous integers $(0, 1, 2, \ldots$ up to the stated seed count per sweep), and each run is reproducible from its (K, n, seed) triple: the training-data generator is seeded at $1000 + \text{seed}$, the held-out test generator at $7{,}000{,}000 + \text{seed}$, and the model initialization uses the seed directly. For the statistical sample-complexity sweeps, for each K we build a fixed training set of n sequences of length $T = 16$, train to convergence, and evaluate fresh-test sequence accuracy on 512 held-out sequences; $m^\star$ is the smallest nT whose median accuracy (over seeds) reaches 0.95. Each gradient-descent run trains for 8000 steps (the soft model converges well within this budget; train accuracy is monitored to confirm convergence). The size grid is geometric and spans from below $K \log K$ to forty times $K^2 \log K$ transitions, so it brackets both candidate rates. The minimum-width sweeps fix $n = 3000$ sequences (coverage made non-limiting), sweep H over a grid up to $2K$, and train for 12000 steps per run; the "GD-unlearnable at $H \leq 2K$" verdict for random automata is therefore relative to this fixed optimization budget. The tabular learner is the optimization-free coverage learner described in Section 3, run for K up to 64 over eight seeds.

The diagonal-versus-selective separation (Section 6) trains both arms online (fresh batches each step, so in-distribution accuracy equals training accuracy) at matched state width $2K$, batch size 64, learning rate 0.01, over six seeds; the selective arm trains for 8000 steps and the diagonal arm for 15000 (more budget, to rule out under-training as the cause of its failure). The sequence length is $T = 16$ for S_5 ($K = 120$) and $T = 48$ for S_3 ($K = 6$). The reported figures are median per-token next-state accuracy on freshly sampled evaluation batches; the diagonal arm's inability to reach high accuracy even on its own training distribution is the evidence that the limitation is representational, not a matter of sample size.

Four control studies support the disclosures in Section 8. (i) The gradient-descent $m^\star$ sweep is run for the non-abelian dihedral family $D_{K/2}$ as well as the cyclic family, with the family-to-order map chosen so the two share the same K axis (Table 2). (ii) The random-automaton minimum-width sweep is repeated for three independent draws of the K symbol permutations at each of $K = 12, 16, 20$; all nine cells are unlearnable at $H \leq 2K$. (iii) The cyclic $m^\star$ sweep is repeated with the hard straight-through model at $K = 12, 20$, with the inverse temperature annealed from 1 to 8. (iv) The sequence-length control fixes $K = 16$

and measures $m^\star$ at $T = 8, 16, 32$. Per-run configuration, seeds, and the resulting tables are recorded in the project ledger.

Non-selective GRU control. The control in Section 4 is a standard one-layer GRU of hidden width $H = 2K$ in which the input token enters only through a learned additive embedding, so the recurrent transition is identical at every step (no input-selected transition matrix). It follows the same fixed-training-set protocol as the selective sweep (geometric n grid, $T = 16$, three seeds, fresh-test bar 0.95), on the cyclic family and on random automata for $K = 8$ to 32. At learning rate 0.01 the GRU diverges for $K \geq 16$ and fails to fit its own training set; we therefore train it at learning rate 0.003 with global-norm gradient clipping at 1.0, under which it fits every training set to perfect accuracy, so its fresh-test failures reflect inductive bias and not an optimization failure. The cyclic extended sweep for the selective model in Table 2 uses five seeds and reaches $K = 48$.

References

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623. PMLR, 2023.

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116. Association for Computational Linguistics, 2020.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 6243–6267. PMLR, 2023.

John D. Dixon. The probability of generating the symmetric group. *Mathematische Zeitschrift*, 110:199–205, 1969.

George Giapitzakis, Kimon Fountoulakis, Eshaan Nichani, and Jason D. Lee. On the statistical query complexity of learning semiautomata: a random walk approach. *arXiv preprint arXiv:2510.04115*, 2025. (preprint).

E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Jianliang He, Leda Wang, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Neural networks provably learn spectral representations for group composition. *arXiv preprint arXiv:2606.02993*, 2026. (preprint).

Gordon James and Adalbert Kerber. *The Representation Theory of the Symmetric Group*. Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

Pascal Koiran and Eduardo D. Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

William Merrill. Sequential neural networks as automata. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 1–13. Association for Computational Linguistics, 2019.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, 2024.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.

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

Leonard Pitt and Manfred K. Warmuth. The minimum consistent DFA problem cannot be approximated within any polynomial. *Journal of the ACM*, 40(1):95–142, 1993.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR Workshop on Mathematical Reasoning in General Artificial Intelligence*, 2022.

Richard Rasala. On the minimal degrees of characters of S_n. *Journal of Algebra*, 45(1): 132–181, 1977.

Aleksandar Terzić, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Structured sparse transition matrices to enable state tracking in state-space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. *arXiv preprint arXiv:2509.21519*, 2025. (preprint).

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 11080–11090. PMLR, 2021.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Ruida Zhou, Chao Tian, and Suhas Diggavi. An information-theoretic approach to understanding transformers' in-context learning of variable-order markov chains. *arXiv preprint arXiv:2410.05493*, 2024.