ACTIVATION-SPECTRUM REGULARIZATION FOR GROKKING-STYLE GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Overparameterized networks can memorize a small algorithmic training set long before they generalize. We test whether compressing the spectrum of hidden-state covariance favors representations that reuse directions across examples. Likelihood-gated activation-spectrum regularization (ASR) applies a trace-based Rényi-2 proxy, spectral entropy, or ridge Log-Det only as predictive confidence grows. The Log-Det gradient weights an eigenvalue by the inverse of its ridge-shifted magnitude, and Sylvester's identity permits an exact computation through the smaller of the batch Gram and feature covariance matrices. ASR lowered 15 of 18 task–objective means and won 66 of 87 matched comparisons across the fixed six-task suite. It reduced the required data fraction by 7.37 percentage points on average (95% stratified bootstrap interval $[5.71, 9.27]$), or 11.40% relative to the matched baseline ($[5.87, 17.03]\%$). Rényi-2 lowered the mean threshold on all six tasks, while spectral entropy gave the lowest thresholds on two heterogeneous arithmetic mixtures. On UCI Superconductivity, every ASR objective improved all six mean test metrics under validation-selected and fixed-budget evaluation, although no objective won every metric. Full-run ModAdd trajectories show recurrent spectral-tail contractions aligned with transient accuracy drops. ASR is therefore a measurable representation-level bias whose benefit depends on architecture, task, and spectral objective.

1 INTRODUCTION

Grokking is the delayed transition from fitting a finite algorithmic training set to generalizing beyond it (Power et al., 2022). It isolates a basic question: after a network memorizes the observed examples, what steers optimization toward a rule that also applies to unobserved examples?

One answer is simplicity. Solomonoff induction assigns more prior weight to shorter computable explanations, and minimum description length prefers models that describe observations concisely (Solomonoff, 1964; Rissanen, 1978). Neither principle yields a direct neural-network objective: Kolmogorov complexity is uncomputable, and parameter count poorly measures the information used by a particular computation. Weight decay supplies an indirect bias, but it contracts parameters rather than the representations that they produce.

Our hypothesis concerns representation geometry. Two networks can have the same training loss while distributing hidden states across different numbers of activation directions. A concentrated covariance spectrum indicates that examples reuse a smaller set of directions; a broad spectrum indicates more diffuse variation. Penalizing spectral spread is a measurable, differentiable simplicity bias. It neither estimates program length nor asserts that every useful algorithm has a low-rank representation.

We call this bias *activation-spectrum regularization* (ASR). We compare three objectives: a trace-based proxy for Rényi-2 entropy, spectral entropy, and a ridge Log-Det volume. The first two depend on the normalized eigenvalue distribution; Log-Det supplies stronger ridge-limited pressure on small directions. A likelihood gate activates each penalty as the model becomes confident, reducing interference with initial task learning.

The paper makes three contributions. First, it defines likelihood-gated ASR and relates its three objectives at the eigenvalue level. Second, it derives a fixed-trace Taylor connection between Rényi-2

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.

and Log-Det and an exact small-matrix Log-Det identity. Third, it evaluates sample efficiency, censoring, interactions with weight decay and dropout, wall-clock cost, heterogeneous arithmetic, and an untuned regression transfer. The results are conditional rather than universal: they are strong on the MLP and mixed-task suites, unresolved in aggregate on Transformers, and sensitive to the chosen spectral objective.

Relation to prior work. ASR complements feature-learning, implicit-bias, and compression accounts of grokking, including weight-spectrum entropy regularization (Power et al., 2022; Lyu et al., 2024; Mohamadi et al., 2024; DeMoss et al., 2025). Its distinct intervention is confidence-gated compression of centered minibatch activation covariance through three spectral objectives. ASR also differs from covariance-preserving self-supervision and coding-rate objectives (Zbontar et al., 2021; Bardes et al., 2022; Yu et al., 2020). Extended comparisons appear in Appendix A.

2 Activation-spectrum regularization

Let $H \in \mathbb{R}^{B\times d}$ denote a matrix of hidden activations for one layer and one minibatch. We center each feature over the batch,

$$\widetilde{H} = \left(I_B - \frac{1}{B}\mathbf{1}\mathbf{1}^\top\right) H, \qquad \mathbf{1}^\top \widetilde{H} = 0, \tag{1}$$

then form the empirical covariance

$$C = \frac{1}{B}\widetilde{H}^\top \widetilde{H} \succeq 0, \qquad \tau = \mathrm{Tr}(C), \qquad \rho = \frac{C}{\tau + \delta}. \tag{2}$$

When several hidden tensors are selected, we compute one penalty per tensor and average them. The normalized matrix ρ describes spectral shape: apart from the stabilizer $\delta > 0$, uniform rescaling does not change it. Centering also removes the batch-mean direction, so the penalty acts on variation across examples rather than on a shared offset.

Numerical conventions. The participation-ratio implementation uses a 10^{-12} denominator stabilizer and routes through the smaller Gram or feature-covariance matrix. Spectral entropy uses the $1/B$ covariance normalization, floors eigenvalues at 10^{-8}, and retains the feature-space null multiplicity when a smaller matrix is diagonalized. Log-Det uses the per-tensor RMS normalization in Equation 9, ridge $\epsilon = 10^{-5}$, and the smaller-matrix route in Equation 13.

2.1 A tractable prior over representational codebooks

The rank of ρ counts active directions only in an ideal noiseless limit; it is discontinuous and brittle in a trained network. Spectral entropy offers a soft substitute. In particular, the participation ratio

$$r_2(C) \coloneqq \frac{\mathrm{Tr}(C)^2}{\mathrm{Tr}(C^2)} = \frac{1}{\mathrm{Tr}(\rho^2)} = \exp(R_2(\rho)) \tag{3}$$

is the number of equally weighted eigen-directions that would have the same purity, where the equalities hold for a nonzero exactly trace-normalized covariance. It therefore acts as a continuous effective codebook size for the batchwise representation. The reported Rényi-2 implementation minimizes $r_2(C)$ (with a 10^{-12} denominator stabilizer), rather than the logarithm $R_2(\rho)$. This is a trace-based proxy: minimizing r_2 has the same ordering and the same optima as minimizing R_2, while avoiding a logarithm. For an ideal nonzero covariance, $r_2 \geq 1$, with equality at rank one.

This yields a concrete relaxation of an Occam-style preference:

$$\min_\theta \ \mathbb{E}_{(x,y)}[\ell(f_\theta(x), y)] + \beta \underbrace{\mathcal{C}(\rho_\theta)}_{\text{spectral codebook complexity}}. \tag{4}$$

The term $\mathcal{C}$ is not a code length in bits and is not invariant to all function-preserving transformations. Its value is operational: it rewards a particular class of internal solutions—those that reuse activation directions across a batch—that can be directly measured, optimized, and falsified on held-out examples.

2.2 A HIERARCHY OF SPECTRAL COMPRESSORS

Write the eigenvalues of ρ as $p_1, \ldots, p_d$. When $\delta = 0$ and $\tau > 0$, they are nonnegative and sum to one. We compare the following quantities:

$$R_2(\rho) = -\log \operatorname{Tr}(\rho^2) = -\log \sum_i p_i^2, \tag{5}$$

$$R_1(\rho) = -\operatorname{Tr}(\rho \log \rho) = -\sum_i p_i \log p_i, \tag{6}$$

$$V_\epsilon(C) = \log\det(C + \epsilon I) = \sum_i \log(\lambda_i(C) + \epsilon). \tag{7}$$

R_2 is the order-two Rényi entropy (and is monotonically related to purity $\operatorname{Tr}(\rho^2)$); R_1 is spectral entropy. Lower entropy means a more concentrated normalized spectrum. The implementation uses the monotone participation-ratio proxy $r_2 = \exp(R_2)$ from Equation 3, not the logarithmic quantity R_2 itself. V_ϵ is a ridge Log-Det volume whose eigenvalue derivative is

$$\frac{\partial V_\epsilon}{\partial \lambda_i} = \frac{1}{\lambda_i + \epsilon}. \tag{8}$$

Thus, in a minimization objective, the ridge Log-Det exerts greater pressure on small nonzero spectral directions than a polynomial penalty with derivative proportional to λ_i. In finite precision this pressure is bounded by $1/\epsilon$; it is not literally infinite.

The three objectives should not be presented as interchangeable. The participation ratio r_2 and R_1 measure spectral *shape*; the mathematical raw V_ϵ can also depend on scale. The reported Log-Det implementation removes this degree of freedom before evaluating the objective: it rescales each selected tensor as

$$\begin{aligned} \bar{H} &= \frac{\widetilde{H}}{\sqrt{\|\widetilde{H}\|_F^2/B + 10^{-12}}}, & \bar{C} &= \frac{1}{B}\bar{H}^\top \bar{H}, \\ V_{\min}(d,\epsilon) &= \log(1+\epsilon) + (d-1)\log\epsilon, & \epsilon &= 10^{-5}, \\ R_{\mathrm{LD}}(H) &= \max\left\{ \frac{V_\epsilon(\bar{C}) - V_{\min}(d,\epsilon)}{d}, 0 \right\}. \end{aligned} \tag{9}$$

Apart from the numerical stabilizer, $\operatorname{Tr}(\bar{C}) = 1$, and $V_{\min}$ is the fixed-trace ridge Log-Det value of a rank-one spectrum. Consequently, $R_{\mathrm{LD}} \geq 0$, and minimizing it reduces the raw Log-Det toward its rank-one fixed-trace minimum; the offset and clamp are essential conventions, not cosmetic presentation. Uniform activation rescaling therefore cannot lower the Log-Det loss used in the reported experiments. More generally, the improvements from the normalized Rényi-2 and spectral-entropy objectives— with spectral entropy strongest on the heterogeneous arithmetic tasks—also show that a purely scale-based account cannot explain the family-level ASR result.

Fixed-trace connection. At fixed trace, the second-order expansion of Log-Det is a purity objective: minimizing its quadratic term is equivalent to minimizing the participation ratio. Log-Det retains higher trace moments and, unlike the participation ratio, has inverse-ridge pressure on weak directions. Appendix B gives the expansion and eigenvalue-level comparison.

2.3 LIKELIHOOD-GATED PRESSURE

Fixed-strength spectral penalties can suppress task-relevant directions before the predictor is competent. The likelihood gate delays substantial pressure until the prediction loss is low and reduces that pressure when the loss rises.

Let $p_i = p_\theta(y_i \mid x_i)$ be the probability assigned to the correct label for batch example i. Cross-entropy gives the exact identity

$$\exp(-\ell_{\mathrm{CE}}) = \exp\left(\frac{1}{B}\sum_{i=1}^{B} \log p_i\right) = \left(\prod_{i=1}^{B} p_i\right)^{1/B}. \tag{10}$$

Thus $\exp(-\ell_{\mathrm{CE}})$ is the geometric-mean correct-label likelihood of the batch: it is close to zero while the task is poorly modeled and approaches one only after the network has become predictively certain. Let $R(H)$ denote the *implemented nonnegative penalty*: $r_2(C)$ for the Rényi-2 proxy, $R_1(\rho)$ for spectral entropy, or $R_{\mathrm{LD}}(H)$ for Log-Det. We use the detached adaptive multiplier

$$\mathcal{L} = \ell_{\mathrm{CE}} + \gamma R(H), \qquad \gamma = w \ \mathrm{stopgrad}[g(R(H)) \exp(-\ell_{\mathrm{CE}})], \tag{11}$$

where w is a tuned base weight. The implementation uses $g(R) = R - 1$ for the participation-ratio Rényi-2 proxy and $g(R) = R$ for spectral entropy and shifted Log-Det; there is no additional temperature or threshold. In the ideal nonzero case, $r_2 \geq 1$, $R_1 \geq 0$, and $R_{\mathrm{LD}} \geq 0$. Thus $g(R) \geq 0$ for every reported objective: the detached gate scales the regularization pressure but never reverses its direction. With the stop-gradient, the ASR contribution to the hidden-state update is proportional to

$$\nabla_H \mathcal{L}_{\mathrm{ASR}} = w \, g(R(H)) \exp(-\ell_{\mathrm{CE}}) \nabla_H R(H). \tag{12}$$

For Log-Det away from the clamp, $\nabla_H R_{\mathrm{LD}}$ is the positive $1/d$-rescaling of the gradient of the raw fixed-trace V_ϵ; gradient descent therefore retains the small-eigenvalue compression direction described in Equation 8. Equation 12 defines feedback rather than a fixed annealing schedule. Low cross-entropy increases the compression gradient; a subsequent loss increase reduces it and gives the task gradient more influence. The spectrum can therefore contract and recover rather than decrease monotonically. Figure 1 shows this pattern over complete 100k-step ModAdd runs. The ASR panels show recurrent tail contractions, and several sharp events align with brief train and test accuracy drops followed by recovery at the 100-step logging resolution. The weight-decay baseline has shallower, localized contractions, so ASR strengthens and organizes the pattern rather than creating every spectral disturbance.

We therefore call the scheme *likelihood-gated* regularization. The name refers only to its dependence on predictive confidence: $\exp(-\ell_{\mathrm{CE}})$ is not a curvature estimate.

2.4 An algebraic small-matrix Log-Det identity

Directly forming $C \in \mathbb{R}^{d \times d}$ is undesirable when d is much larger than batch size B. Sylvester's identity, $\det(I + AB) = \det(I + BA)$, gives the exact equality

$$\log\det\left(\epsilon I_d + \frac{1}{B}\widetilde{H}^\top \widetilde{H}\right) = \log\det\left(\epsilon I_B + \frac{1}{B}\widetilde{H}\widetilde{H}^\top\right) + (d - B)\log\epsilon, \quad B < d. \tag{13}$$

The final term has zero gradient with respect to network parameters and may be omitted when only optimizing. It must be restored when reporting the true feature-space value or comparing widths. We route the computation through the smaller of the $B \times B$ Gram and $d \times d$ feature matrices. This removes the decomposition's direct dependence on the larger dimension, but forming the Gram matrix and factoring a $\min(B, d)$ square matrix still have nonzero cost. The equality is algebraically exact; its finite-precision realization requires the numerical checks reported in Appendix J.

3 Experimental protocol

3.1 Task families and architectures

ModAdd. We exhaustively enumerate modular addition over integers $0, \ldots, 30$. Each example encodes two operands as concatenated 31-way one-hot vectors (62 input features); the label is $(a + b) \bmod 31$. The task has 31 classes and $31^2 = 961$ examples.

S_5 composition. We exhaustively enumerate composition in the symmetric group S_5. Inputs are two concatenated 120-way one-hot permutation identifiers (240 features), and the label is the identifier of their composed permutation. The task contains $120^2 = 14{,}400$ examples and 120 classes. Unlike ModAdd, it requires learning a non-commutative composition rule. It therefore tests ASR on a larger class space and a different algebraic structure from cyclic ModAdd, without assuming that the network learns a literal group representation.

Parity. The parity task contains 4,096 random binary vectors with 50 Bernoulli input bits. Its binary label is the XOR of bit positions 2, 14, and 42. It is a counterpoint to the group-composition tasks: it has a short rule but need not benefit from the same covariance geometry. It separates program simplicity from compatibility with a particular representation penalty.

Sequential variants. For Transformer experiments, we present the corresponding operand or bit representations as sequences (ModStream, S5Stream, and ParitySeq) and predict the same task labels. This tests whether ASR survives a change in the native representation and routing mechanism rather than only an MLP input format.

Implemented architectures. The MLPs use the architecture input $\to$ 128 $\to$ 128 $\to$ classes, with two post-SiLU hidden layers. Linear weights are Xavier initialized and linear biases are initialized to zero. The ModAdd, S_5, and parity models contain 28,575, 62,840, and 23,298 parameters, respectively. ASR is applied to both post-SiLU layers. The executable three-task ModAdd+ModMul+ModSquare model uses the same MLP, a 65-dimensional input (including a three-way task ID), 93 disjoint-range output classes, and 36,957 parameters.

The Transformer models use two causal decoder blocks with model width 128, four heads of dimension 32, SwiGLU inner width 256, fixed RoPE, pre-RMSNorm ($\epsilon = 10^{-6}$), and residual connections (237,570 parameters). Float parity inputs use a learned $50 \to 128$ linear embedding. For each attention head, we form $A_h = \mathrm{softmax}(Q_h K_h^\top / \sqrt{d_h})$ and context $Z_h = A_h V_h$. ASR is applied to Z_h before head concatenation and the output projection W_o; it is not applied to the projected residual write or to FFN activations. These are standard decoder-only Transformer components (Vaswani et al., 2017).

3.2 Three-stage evaluation

We compare ordinary weight decay (None) with Rényi-2, spectral-entropy, and ridge-Log-Det ASR, each combined with weight decay. All reported sweeps use five seeds $(0, 42, 67, 89, 123)$. Training uses full-batch AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3} and no learning-rate schedule; unspecified AdamW settings use the Optax defaults. Linear weights use Xavier initialization (Glorot & Bengio, 2010) and linear biases are initialized to zero. Unless stated otherwise, the primary threshold and timing comparisons fix $w = 0.01$ and weight decay 0.1 for every ASR objective; the baseline uses the same weight decay.

Stage I: full-data hyperparameter sweeps. We first train on the full task dataset while sweeping ASR weight $w \in \{0, 0.01, 0.05\}$ and weight decay $\lambda_{\mathrm{wd}} \in \{0.05, 0.1, 0.5\}$. We record the first step T at which test accuracy reaches 99%, with a 100k-step budget. A run that does not grok by the budget is a right-censored observation:

$$\widetilde{T} = \min(T, T_{\max}), \qquad \Delta = \mathbf{1}\{T \le T_{\max}\}, \qquad T_{\max} = 100{,}000. \tag{14}$$

The hyperparameter grids are therefore summarized by both the number of uncensored trajectories and time-to-grok among runs that finish; a timeout is never treated as an ordinary 100k-step measurement or silently discarded.

Regularizer-interaction controls. On full-data ModAdd, we additionally test dropout (Srivastava et al., 2014) with probability 0.1. With weight decay fixed at 0.1, this gives a 2×4 factorial control: dropout on or off crossed with no ASR, Rényi-2, spectral entropy, or Log-Det. We also test each ASR objective without weight decay. These controls use the same five matched seeds and the same 100k-step, 99% time-to-grok criterion as Stage I.

Stage II: data-threshold search. We then use a binary-search-style procedure to find the smallest exposed training-data fraction that reaches 99% test accuracy within 100k steps. This is the primary sample-efficiency metric. Lower is better.

Stage III: heterogeneous-task stress tests. Finally, we evaluate task-ID-conditioned mixtures. A maximally heterogeneous pilot combining zero-padded ModAdd, S_5, and parity examples did not grok for any compared method in the initial runs. We therefore use the more controlled ModAdd+ModMul and ModAdd+ModMul+ModSquare mixtures for the quantitative heterogeneity study. These retain a shared modular input domain while testing whether a global spectrum penalty can support multiple simultaneously learned rules. Because multi-task grokking is substantially slower, the heterogeneous data-threshold searches use a 250k-step cap rather than the 100k-step cap used for Stages I–II.

Table 1: Minimum exposed training-data fraction (%) needed to achieve 99% test accuracy by 100k steps. Entries are mean ± sample standard deviation over observed thresholds; lower is better. All cells contain five seeds unless marked †. Bold marks the lowest mean in each row.

Architecture	Task	None	Rényi-2	Spectral	Log-Det
MLP	ModAdd	66.5 ± 3.3	52.1 ± 0.7	$\mathbf{49.2 \pm 1.6}$	63.3 ± 6.5
MLP	S_5	55.9 ± 3.0	39.4 ± 1.0	37.9 ± 1.3	$\mathbf{33.6 \pm 1.8}$
MLP	Parity	24.6 ± 5.1	24.4 ± 1.2	$\mathbf{23.3 \pm 2.4}$	25.5 ± 3.0
Transformer	ModStream	25.6 ± 11.1	$\mathbf{23.3 \pm 3.5}$	26.0 ± 8.1	25.3 ± 11.8
Transformer	S5Stream	51.9 ± 16.4 $(4/5)^{\dagger}$	43.5 ± 6.3	42.5 ± 15.3	$\mathbf{28.5 \pm 4.3}$
Transformer	ParitySeq	2.68 ± 0.41	$\mathbf{2.27 \pm 0.34}$	2.63 ± 0.85	2.83 ± 0.53

The grid sweeps evaluate every update; the trajectory and threshold runs evaluate every 100 updates. Grid and trajectory runs use 80/20 train/test splits, while the generic and threshold runs use 90/10 splits. All summaries retain success counts for censored searches. Dedicated 100k-step ModAdd runs log full eigenspectra every 100 updates for the baseline and each ASR objective. Activation trace, effective rank, and gate values are available only for a subset of diagnostic runs.

Untuned real-data transfer. We additionally evaluate an MLP on the UCI Superconductivity regression benchmark (Hamidieh, 2018). All four objectives use the same data split (seed 464), 100k-step budget, and evaluation interval of 100 steps. We transfer the ASR coefficient $w = 0.01$ unchanged from the algorithmic experiments and use weight decay 10^{-4}; no coefficient is tuned on this dataset. Ten matched optimization seeds $(0, 42, 67, 89, 123, 321, 404, 420, 555, 600)$ are used for every objective. The headline metric reports test performance at the checkpoint selected only by best validation R^2. We separately report fixed-budget test performance at 100k steps; oracle selection by test performance is not used for either comparison.

4 Results

Full-data coefficient grids and the dropout/weight-decay interaction controls are reported in Appendices C and D. They contextualize the common $w = 0.01$, weight-decay-0.1 configuration used in the main comparisons.

4.1 ASR lowers thresholds across the fixed suite

Table 1 reports data thresholds. ASR lowered 15 of 18 task–objective means and won 66 of 87 observed matched comparisons. Rényi-2 lowered the mean threshold on all six tasks; spectral entropy did so on five, and Log-Det on four. Across the four arithmetic and composition tasks, 11 of 12 means and 48 of 57 matched comparisons favored ASR. Each objective separately had positive stratified-bootstrap intervals under both the absolute and baseline-normalized suite summaries. The mean absolute reduction was 7.37 percentage points (95% stratified bootstrap interval $[5.71, 9.27]$), and the mean baseline-normalized reduction was 11.40% ($[5.87, 17.03]\%$). Correlation-preserving sign-flip estimates were $p = 1.62 \times 10^{-4}$ and $p = 0.0095$, respectively. Thus the suite-wide improvement held under both absolute and relative summaries.

The MLP result was stable under both summaries. Averaged across its three tasks and objectives, ASR reduced the threshold by 10.26 percentage points ($[8.91, 11.73]$) or 16.57% relative ($[12.57, 20.52]\%$). Rényi-2 and spectral entropy both reduced ModAdd, and all three objectives reduced S_5; five of nine paired MLP intervals excluded zero. Log-Det gave the largest S_5 reduction, from $55.9 \pm 3.0\%$ to $33.6 \pm 1.8\%$. MLP parity remained unresolved.

The Transformer-family sign-flip test remained unresolved, but its task-level pattern improved. All three objectives reduced S5Stream in every observed baseline-matched seed, and all three paired intervals excluded zero; Log-Det lowered its mean threshold from 51.9% to 28.5%. Rényi-2 also lowered the ModStream and ParitySeq means, although neither paired interval excluded zero. Spectral entropy approximately matched the ParitySeq baseline, while Log-Det raised it modestly. Transformer effects therefore depended on both the task and spectral objective.

Table 2: Untuned transfer to UCI Superconductivity regression over ten matched optimization seeds and one fixed data split. All ASR methods use the synthetic-selected $w = 0.01$ and weight decay 10^{-4}, with no dataset-specific ASR tuning. Upper panel: test metrics at the checkpoint selected by validation R^2 only. Lower panel: test metrics after the fixed 100k-step budget. Entries are mean $\pm$ sample standard deviation; higher R^2 and lower errors are better.

Validation-selected checkpoint

Objective	Test R^2 (%)	Test MAE (K)	Test RMSE (K)
None	92.178 ± 0.241	5.395 ± 0.120	9.389 ± 0.145
Rényi-2	$\mathbf{92.520 \pm 0.330}$	5.281 ± 0.105	$\mathbf{9.180 \pm 0.203}$
Spectral	92.468 ± 0.242	$\mathbf{5.257 \pm 0.129}$	9.213 ± 0.148
Log-Det	92.506 ± 0.241	5.334 ± 0.086	9.190 ± 0.148

Fixed 100k-step budget

Objective	Test R^2 (%)	Test MAE (K)	Test RMSE (K)
None	92.063 ± 0.317	5.339 ± 0.066	9.457 ± 0.190
Rényi-2	92.394 ± 0.292	5.266 ± 0.066	9.257 ± 0.177
Spectral	92.323 ± 0.384	$\mathbf{5.259 \pm 0.123}$	9.300 ± 0.234
Log-Det	$\mathbf{92.459 \pm 0.248}$	5.279 ± 0.076	$\mathbf{9.218 \pm 0.151}$

4.2 Untuned ASR transfer improves real-data regression

The synthetic results establish that the useful spectrum objective depends on representation geometry. Table 2 shows that this dependence persisted on a non-algorithmic, real-data regression benchmark without retuning the ASR coefficient. Every ASR objective improved all three mean test metrics under both checkpoint conventions. No objective won every metric. At the validation-selected checkpoint, Rényi-2 gave the best mean R^2 and RMSE, while spectral entropy gave the best MAE. At 100k steps, Log-Det gave the best R^2 and RMSE, while spectral entropy again gave the best MAE. Depending on objective and metric, these mean gains occurred in six to eight of ten paired validation-selected runs. Paired bootstrap intervals for the primary test R^2 excluded zero for all three objectives (Appendix G).

The conclusion is not an early-convergence claim. At common validation thresholds, ASR typically reached the target slightly later: for example, the baseline reached 90% validation R^2 after $5{,}960 \pm 740$ steps, compared with $6{,}640 \pm 556$, $6{,}700 \pm 660$, and $6{,}880 \pm 658$ steps for Rényi-2, spectral entropy, and Log-Det, respectively. The result therefore concerns the quality after sustained training, not faster initial convergence.

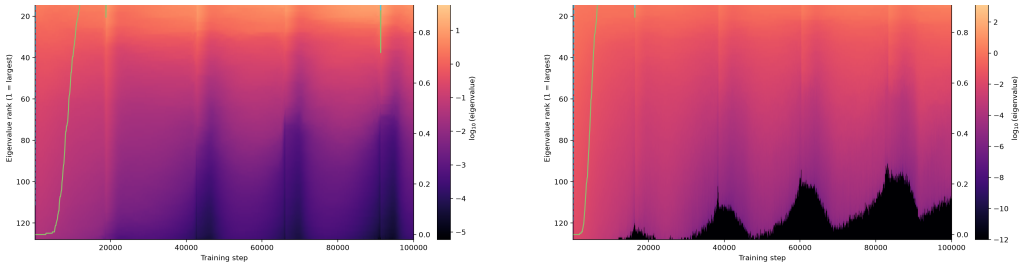
4.3 Spectral dynamics provide a mechanistic diagnostic

Figure 1 shows the logged mean activation spectra and aligned accuracy trajectories for full ModAdd MLP runs. Because the covariance uses centered activations, the plotted changes describe example-to-example variation rather than motion of the batch-mean activation vector. The weight-decay baseline remained comparatively stable, with several localized tail contractions. All three ASR runs showed deeper recurrent contraction and recovery throughout training. Several sharp ASR events coincided with short train and test accuracy drops followed by recovery. We refer to the lowest viable part of this operating range as the *compression boundary*.

The objectives produced distinct dynamics. Rényi-2 showed broad, smooth contractions, spectral entropy showed pronounced compression–recovery waves, and Log-Det repeatedly produced the deepest depleted tail. This ordering is consistent with the inverse-ridge pressure in Equation 8, but the figure does not identify a unique mechanism. The panels use different color limits, so color intensity must not be compared directly across objectives; the relevant visual evidence is the within-panel temporal structure and the rank extent of the dark tail.

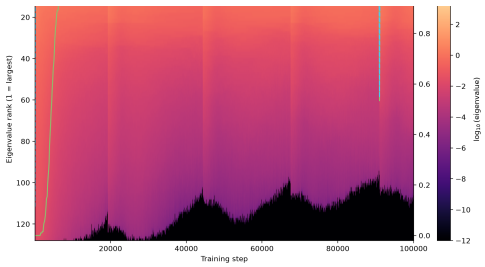
4.4 Spectral entropy extends ASR gains to heterogeneous arithmetic

We also construct task-ID-conditioned mixed datasets. Table 3 shows all five seeds with a 250k-step limit. On ModAdd+ModMul, spectral entropy lowered the threshold from $79.7 \pm 3.0\%$ to $56.7 \pm 3.2\%$; Rényi-2 also reached all five thresholds at $64.9 \pm 3.1\%$. On the three-task mixture, the

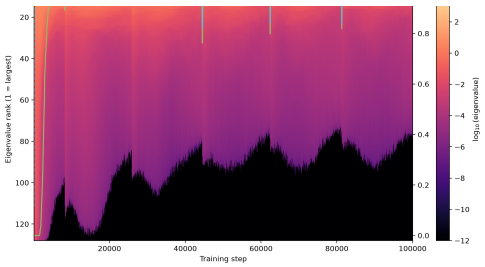


(a) Weight decay

(b) Rényi-2



(c) Spectral entropy



(d) Log-Det

Figure 1: Eigenvalues of centered activation covariance over complete 100k-step ModAdd MLP runs. Eigenvalues are sorted at each checkpoint and shown on a logarithmic scale. Dashed cyan and solid green curves show train and test accuracy. The baseline shows occasional localized contractions, whereas ASR produces deeper recurrent compression–recovery, especially under Log-Det. Several sharp events align with transient accuracy drops. Each panel uses its own color limits, so colors are comparable over time within a panel but not in magnitude across panels. The alignment is diagnostic, not causal.

Table 3: Heterogeneous MLP data-fraction results under a 250k-step limit. Entries are mean $\pm$ sample standard deviation over uncensored thresholds; the parenthesis reports the number reaching the criterion. Runs that do not reach the criterion within the prescribed data-fraction search and 250k-step budget are right-censored, denoted by †, and are *not* assigned a fabricated threshold. Lower is better.

Mixture	None	Rényi-2	Spectral	Log-Det
Add+Mul	79.7 ± 3.0 (5/5)	64.9 ± 3.1 (5/5)	$\mathbf{56.7 \pm 3.2}$ **(5/5)**	77.6 ± 6.9 (4/5)†
Add+Mul+Square	82.2 (1/5)†	68.9 ± 8.1 (5/5)	$\mathbf{65.3 \pm 4.4}$ **(5/5)**	78.5 ± 10.2 (4/5)†

baseline succeeded once, whereas Rényi-2 and spectral entropy succeeded in every seed. Spectral entropy again gave the lowest observed mean, $65.3 \pm 4.4\%$. Log-Det was less reliable and had a higher threshold on both mixtures. Thus the normalized spectral objectives transfer better than Log-Det when one batch contains several rules.

One explanation is cross-task interference. A global batch penalty can remove directions that distinguish rules even when each direction is weak in the pooled covariance. This interpretation does not imply that the network learns orthogonal task-specific subspaces. It does identify a testable design issue: the spectral objective must preserve the diversity required by a heterogeneous batch. The stronger Rényi-2 and spectral-entropy results show that this is an objective-selection problem, not a general failure of ASR.

4.5 Wall-clock benefit depends on objective and task

A fixed-horizon microbenchmark does not measure time to generalization. We therefore record wall-clock time to the first observed 99% test accuracy. Appendix H reports both steps and seconds.

The trace-only Rényi-2 proxy improved ModAdd from 4.16 to 2.17 seconds and S_5 from 12.30 to 3.02 seconds. Spectral entropy often reduced the step count, but its eigendecomposition dominates

these small models and made every MLP wall-clock result slower. Log-Det was also slower on the MLPs, and only three of five ModAdd timing runs reached the criterion.

All three ASR objectives reduced the success-only mean on Transformer ModStream: 52.46 seconds for the baseline, 11.16 for Rényi-2, 17.45 for spectral entropy, and 13.77 for the four successful Log-Det runs. On S5Stream, every ASR objective succeeded in all five seeds while the baseline succeeded in four, but the observed wall times did not show a clear speed gain. Every ASR objective was slower on ParitySeq. ASR can therefore reduce total time when it saves enough updates, but the result is neither task-independent nor guaranteed by a lower step count.

When the backbone is small, dense spectral linear algebra is a large fraction of each update. When $B < d$, the exact Sylvester identity in Equation 13 replaces the $d \times d$ Log-Det factorization with its smaller $B \times B$ Gram counterpart. None of the timing rows activates that route because $B \geq d$. The identity may therefore help in wider, small-batch settings, but the present timings do not test that regime. Spectral entropy still requires an eigendecomposition of the selected matrix.

5 Discussion

The results support a suite-level claim with architecture-dependent detail. ASR lowers 15 of 18 task–objective means, including 11 of 12 on the four arithmetic and composition tasks, and lowers both the absolute and baseline-normalized aggregate thresholds. The MLP effect is consistent, especially on ModAdd and S_5. The Transformer family remains unresolved as a separate aggregate, but all three objectives reduce S5Stream and Rényi-2 also lowers the ModStream and ParitySeq means.

Objective choice matters. Log-Det is strongest on MLP S_5 and in parts of the full-data grid, where its inverse-ridge gradient removes weak directions aggressively. Spectral entropy is strongest on the heterogeneous mixtures and records the most paired wins in the six-task suite, while Rényi-2 gives the largest mean relative reduction. On superconductivity, all objectives improve every mean metric, but the best objective changes with the metric and checkpoint convention. These results do not support a universal ordering of spectral penalties.

The trajectories add a dynamical observation. The weight-decay baseline shows occasional contractions, while ASR produces deeper and more recurrent compression–recovery cycles. Several cycles align with transient accuracy loss. The alignment is consistent with a compression boundary but does not identify a unique cause. More generally, a low-dimensional representation can memorize, and a reusable rule can require many directions. ASR tests one specific bias: whether reducing batchwise spectral spread helps a given model and task.

6 Limitations and reproducibility

Our evidence combines synthetic algorithmic tasks with one real-data tabular regression transfer, not a broad survey of modalities or scales. The superconductivity experiment uses ten optimization seeds but one fixed data split, so its variation quantifies optimization randomness rather than split-to-split variation. The data-fraction sweep is expensive and produces censored outcomes; success-only means must therefore be read with their success counts. The displayed coefficient study covers a 3×3 grid, not the larger space of plausible ASR and weight-decay values. The gate adds a base weight and a stop-gradient choice. The mathematical raw Log-Det is scale-sensitive, although the implementation RMS-normalizes each selected tensor before applying it. The trajectory figure uses one run per objective and panel-specific color limits, and temporal alignment is not causal identification. Finally, all reported runs use a fixed ridge, so sensitivity to ridge selection remains untested.

7 Conclusion

Activation-spectrum regularization imposes a concrete compression bias on minibatch representations. Rényi-2 is cheap, spectral entropy is effective but costly, and Log-Det gives routable inverse-ridge pressure on weak directions. The experiments show resolved suite-wide threshold reductions, strong MLP and heterogeneous-task gains, robust Transformer S5Stream gains, and broad superconductivity improvements. ASR is a testable representation-level inductive bias, not a universal law.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 specify the ASR objectives, numerical conventions, architectures, optimization settings, regularization coefficients, random seeds, data splits, evaluation criteria, training budgets, and censoring rules. The appendices report the full-data coefficient sweeps, regularizer-interaction controls, paired seed-level and aggregate analyses, wall-clock measurements, and numerical validation (Appendices C–H and J). Complete derivations of the fixed-trace Rényi-2–Log-Det connection and routed Log-Det identity appear in Appendices B and I. The algorithmic datasets are generated by the procedures in Section 3; the public UCI Superconductivity dataset is cited and its split and evaluation protocol are specified there.

AI USE STATEMENT

We used generative AI tools to improve grammar and clarity, rewrite portions of the manuscript, organize and summarize experimental results, and suggest intermediate algebraic steps in some mathematical derivations. The research questions, hypotheses, ASR method, experimental design, implementation decisions, data collection, and execution of all experiments were developed and performed by the authors. The authors checked the AI-assisted text, analyses, and derivations against the experimental records and mathematical calculations and take responsibility for the complete contents of this work.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Branton DeMoss, Silvia Sapora, Jakob Foerster, Nick Hawes, and Ingmar Posner. The complexity dynamics of grokking. *Physica D: Nonlinear Phenomena*, 482:134859, 2025. doi: 10.1016/j.physd.2025.134859.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Maryam Fazel, Haitham Hindi, and Stephen P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the American Control Conference*, volume 3, pp. 2156–2162, 2003.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pp. 249–256, 2010.

Kamiran A. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, pp. 950–957, 1992.

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

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? A theoretical analysis on grokking modular addition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

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.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561, 1961.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pp. 606–610, 2007.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Ray J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1): 1–22, 1964.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12310–12320, 2021.

A Related work

Grokking and simplicity biases. Power et al. (Power et al., 2022) introduced grokking as delayed generalization on small algorithmic datasets. Such datasets make it practical to measure the gap between memorization and rule-like generalization. Our work uses them as a controlled setting for a different inductive bias: compression of the observed activation spectrum. The philosophical motivation is related to Occam's razor and MDL (Solomonoff, 1964; Rissanen, 1978), but ASR is neither a computable implementation of Solomonoff induction nor an estimate of neural network description length. Existing accounts instead emphasize changes in implicit bias and feature learning (Lyu et al., 2024; Mohamadi et al., 2024). These accounts are complementary to ASR: we intervene directly on a measured property of the learned representation and test the resulting tradeoff.

Compression-based accounts of grokking. DeMoss et al. (DeMoss et al., 2025) study grokking through lossy model compression and introduce a spectral-entropy penalty on weight matrices. This supports the broad hypothesis that compression and grokking are related, but the intervention differs from ours. ASR regularizes centered activation covariance rather than parameter spectra, couples its strength to predictive confidence, compares Rényi-2, spectral entropy, and normalized Log-Det, and tests minimum-data thresholds across MLPs, Transformers, heterogeneous tasks, and an untuned regression transfer. Our contribution is this activation-space control framework and its empirical evaluation, not the general proposition that lower spectral complexity can accompany generalization.

Weight-space versus representation-space complexity. Weight decay is often motivated as a simplicity bias, but it controls a parameter-space norm. Equivalent functions can be expressed by different parameterizations and can induce different activation geometries. We therefore study ASR as a complementary bias: it acts on the batchwise statistics of the computation rather than directly on its weights. This does not make ASR a superior universal form of MDL; it makes explicit

which aspect of complexity is being encouraged. This distinction matters because weight decay can improve generalization while remaining agnostic about the internal coordinates that carry the computation (Krogh & Hertz, 1992; Loshchilov & Hutter, 2019).

Spectral dimensionality and entropy. Effective rank summarizes the entropy of normalized singular values and gives a continuous notion of effective dimensionality (Roy & Vetterli, 2007). Rényi entropies generalize Shannon entropy (Rényi, 1961); the order-two quantity is particularly convenient because it depends only on a matrix trace. We use these quantities as regularizers, rather than only as post-hoc diagnostics. This places ASR alongside representation objectives that explicitly control feature covariance or redundancy, including Barlow Twins and VICReg (Zbontar et al., 2021; Bardes et al., 2022). The direction is deliberately different: those objectives protect feature variance and discourage decorrelation collapse, whereas ASR applies a likelihood-gated pressure toward a smaller reusable batchwise codebook.

Log-Det representation objectives. Log-Det is a classical smooth surrogate for matrix rank minimization (Fazel et al., 2003); coding-rate objectives also use log-determinants to construct diverse, discriminative representations across classes (Yu et al., 2020). ASR is neither formulation. We minimize a ridge Log-Det of the selected batch covariance, conditional on predictive confidence, rather than maximize a coding-rate difference or impose class-conditional diversity. The routed computation follows the standard determinantal identity $\det(I + AB) = \det(I + BA)$ (Horn & Johnson, 2012).

Information-theoretic interpretations. Information-theoretic language is useful here as an intuition for statistical compression, but it can be misleading. A concentrated activation covariance does not prove low mutual information with the input, causal abstraction, or algorithmic understanding. Our claims are deliberately restricted to an inductive bias and its measured effects on the considered tasks. In particular, the entropy quantities used here are spectral summaries rather than estimates of a communication channel's mutual information (Shannon, 1948; Cover & Thomas, 2006).

B Fixed-trace Taylor connection between Rényi-2 and Log-Det

The connection becomes exact in a small-α expansion. For a trace-normalized ρ with $\mathrm{Tr}(\rho) = 1$,

$$\log\det(I + \alpha\rho) = \mathrm{Tr}\log(I + \alpha\rho) \tag{15}$$

$$= \alpha\,\mathrm{Tr}(\rho) - \frac{\alpha^2}{2}\,\mathrm{Tr}(\rho^2) + \frac{\alpha^3}{3}\,\mathrm{Tr}(\rho^3) + O(\alpha^4) \tag{16}$$

$$= \alpha - \frac{\alpha^2}{2}\,\mathrm{Tr}(\rho^2) + \frac{\alpha^3}{3}\,\mathrm{Tr}(\rho^3) + O(\alpha^4). \tag{17}$$

The first-order term is constant. Consequently, minimizing the second-order approximation is equivalent to maximizing $\mathrm{Tr}(\rho^2)$, or equivalently to minimizing either $R_2(\rho)$ or its monotone proxy $r_2(\rho)$. This is a local, fixed-trace statement—not a claim that an arbitrary unnormalized Log-Det loss equals Rényi-2.

Equation 17 gives a clean interpretation of the comparison. The participation-ratio Rényi-2 proxy is a low-cost second-order spectral-concentration objective; Log-Det retains all higher trace moments and has the inverse-ridge response in Equation 8. Borrowing density-matrix terminology, increasing $\mathrm{Tr}(\rho^2)$ increases *purity*. In this paper, purity means spectral concentration of a covariance matrix only.

The participation-ratio proxy cannot reproduce the inverse-ridge tail. At fixed trace, let $q = \mathrm{Tr}(\rho^2)$, so the implemented Rényi-2 proxy is $r_2 = q^{-1}$. Its eigenvalue derivative is

$$\frac{\partial r_2}{\partial p_i} = -\frac{2p_i}{q^2}. \tag{18}$$

Because $q \in [1/d, 1]$, this response is finite and vanishes as $p_i \to 0$. It therefore cannot reproduce the inverse-ridge response $(\lambda + \epsilon)^{-1}$ near the tail. This distinction does not imply that Log-Det is always empirically superior; it identifies a precise optimization difference. The participation-ratio proxy supplies bounded local pressure near a vanishing eigenvalue, whereas ridge Log-Det supplies a tunably large pressure of order $1/\epsilon$.

C FULL-DATA SWEEP GEOMETRY

The full-data sweeps test whether the results depend on a single coefficient. Table 4 presents one 3×3 coefficient grid for each MLP task and ASR objective.

Log-Det gives the fastest displayed ModAdd cell, and every ASR objective has a fast S_5 region at weight decay 0.5. Several low-weight-decay S_5 cells are censored. Parity reaches the criterion rapidly throughout the grid, so its full-data timing leaves little room for improvement. Timeouts remain right-censored rather than being assigned a value of 100k steps.

Table 4: Full-data MLP sweeps. Within each mini-grid, rows are $w \in \{0, 0.01, 0.05\}$ from top to bottom and columns are $\lambda_{\text{wd}} \in \{0.05, 0.1, 0.5\}$ from left to right. Cells give mean time-to-grok in 10^3 steps. Unmarked cells contain five observed seeds; $x_n^\dagger$ denotes a mean over n observed seeds with the remaining trajectories right-censored at 100k steps. The $w = 0$ row is the shared canonical no-ASR baseline (the Rényi-2 configuration with coefficient zero), repeated across objective panels for comparison rather than treated as independent objective-specific trials.

	ModAdd			S_5			**Parity**		
Rényi-2	27.5	14.0	3.1	$39.9^\dagger_3$	30.7	6.8	0.9	**0.8**	1.0
	9.9	7.1	3.0	$7.9^\dagger_4$	$5.6^\dagger_4$	**5.6**	1.3	1.3	1.2
	10.2	7.5	**3.0**	$9.6^\dagger_4$	$6.8^\dagger_4$	8.2	1.5	1.5	1.6
Spectral	27.5	14.0	3.1	$39.9^\dagger_3$	30.7	6.8	0.9	**0.8**	1.0
	9.1	6.7	2.4	23.6	14.7	5.2	0.8	0.8	1.3
	7.4	5.6	**2.0**	28.3	15.6	**4.5**	1.5	1.4	1.2
Log-Det	27.5	14.0	3.1	$39.9^\dagger_3$	30.7	6.8	0.9	**0.8**	1.0
	7.3	4.0	2.0	$10.3^\dagger_4$	17.3	4.2	0.9	1.0	1.6
	3.6	2.6	**1.1**	$9.1^\dagger_4$	$5.5^\dagger_4$	**3.0**	1.4	1.4	1.5

D REGULARIZER-INTERACTION CONTROLS

Table 5 isolates how ASR interacts with two standard forms of regularization on full-data ModAdd. Dropout alone is not a replacement for weight decay: it reaches the grokking criterion in only one of five seeds. With weight decay, however, dropout reduces mean time-to-grok from 28.8 ± 13.1k to 11.0 ± 2.8k steps. Adding ASR reduces the mean further, to 7.7 ± 2.4k for Rényi-2, 7.1 ± 1.6k for spectral entropy, and 4.8 ± 0.9k for Log-Det.

Without weight decay, Rényi-2 never reaches the criterion, spectral entropy does so once, and Log-Det does so three times. All three objectives reach the criterion in every seed when combined with weight decay. The controls therefore support a complementary interaction: ASR improves the weight-decay and weight-decay-plus-dropout settings but is unreliable as the only regularizer at these coefficients.

Table 5: ModAdd regularizer-interaction controls over five matched seeds. All runs use the full dataset and the 100k-step, 99% test-accuracy criterion. Dropout probability is 0.1. Time is mean $\pm$ sample standard deviation over uncensored runs, in 10^3 steps; † denotes right-censoring.

Weight decay	Dropout	ASR	Grokked	Time to grok (10^3)
None	None	Rényi-2	$0/5$	$--^\dagger$
None	None	Spectral	$1/5$	$14.0^\dagger$
None	None	Log-Det	$3/5$	$20.1 \pm 10.9^\dagger$
None	0.1	None	$1/5$	$28.6^\dagger$
0.1	None	None	$5/5$	28.8 ± 13.1
0.1	None	Rényi-2	$5/5$	14.8 ± 2.5
0.1	None	Spectral	$5/5$	11.8 ± 3.9
0.1	None	Log-Det	$5/5$	9.7 ± 4.6
0.1	0.1	None	$5/5$	11.0 ± 2.8
0.1	0.1	Rényi-2	$5/5$	7.7 ± 2.4
0.1	0.1	Spectral	$5/5$	7.1 ± 1.6
0.1	0.1	Log-Det	$5/5$	$\mathbf{4.8 \pm 0.9}$

E PAIRED SEED-LEVEL THRESHOLD EFFECTS

Table 6: Paired effects on the minimum data threshold. Each entry is the mean baseline-minus-ASR difference in percentage points and its 95% percentile-bootstrap interval from 200,000 resamples (Efron & Tibshirani, 1993). S5Stream uses four observed matched seeds; other cells use five. Positive values favor ASR. Bold intervals exclude zero.

Architecture	Task	Rényi-2	Spectral	Log-Det
MLP	ModAdd	$\mathbf{+14.3}\,[\mathbf{+11.9}, \mathbf{+16.6}]$	$\mathbf{+17.2}\,[\mathbf{+14.2}, \mathbf{+19.4}]$	$+3.1\,[-1.5, +7.7]$
MLP	S_5	$\mathbf{+16.6}\,[\mathbf{+13.7}, \mathbf{+19.5}]$	$\mathbf{+18.1}\,[\mathbf{+15.6}, \mathbf{+20.7}]$	$\mathbf{+22.4}\,[\mathbf{+20.9}, \mathbf{+24.5}]$
MLP	Parity	$+0.2\,[-3.0, +4.2]$	$+1.3\,[-2.1, +4.5]$	$-0.9\,[-3.0, +1.0]$
Transformer	ModStream	$+2.4\,[-5.6, +12.6]$	$-0.3\,[-6.1, +4.2]$	$+0.3\,[-13.5, +13.4]$
Transformer	S5Stream	$\mathbf{+10.7}\,[\mathbf{+3.4}, \mathbf{+23.1}]$	$\mathbf{+6.9}\,[\mathbf{+2.7}, \mathbf{+11.3}]$	$\mathbf{+24.0}\,[\mathbf{+17.4}, \mathbf{+36.1}]$
Transformer	ParitySeq	$+0.4\,[-0.1, +0.9]$	$+0.1\,[-0.7, +0.8]$	$-0.2\,[-0.6, +0.3]$

F AGGREGATE MATCHED-PAIRS THRESHOLD ANALYSIS

Each ASR run shares its seed and baseline with the other objectives. For task t, seed s, and objective m, define the absolute and relative reductions

$$a_{t,s,m} = q_{t,s,\mathrm{None}} - q_{t,s,m}, \qquad r_{t,s,m} = \frac{a_{t,s,m}}{q_{t,s,\mathrm{None}}},$$

where q is the minimum required data percentage. Positive values favor ASR. One S5Stream baseline is censored, leaving 29 task–seed blocks and 87 matched differences.

Table 7 reports both estimands because they answer different questions. The absolute mean gives equal weight to a percentage point on every task. The relative mean gives more weight to changes when the baseline threshold is small; this makes ParitySeq influential because its baseline is only 2.68%. Both all-ASR intervals exclude zero. A two-sided Monte Carlo sign-flip test with 4,000,000 assignments gives $p = 1.62 \times 10^{-4}$ for the absolute mean and $p = 0.0095$ for the relative mean. Each flip is shared by the three objective differences within a task–seed block.

The architecture split is consistent across estimands. Exact enumeration of the 15 MLP blocks gives $p = 9.16 \times 10^{-4}$ for the absolute mean and $p = 0.00354$ for the relative mean. The corresponding Transformer values are $p = 0.091$ and $p = 0.391$. These inferences condition on the six evaluated tasks rather than treating them as a random sample of all tasks.

The primary analysis is complete-case: it excludes one S5Stream baseline that did not reach criterion at the maximum 90% exposure, although all three matched ASR runs did. As a conservative sensitivity analysis, assigning that censored baseline exactly 90% gives 69 wins in 90 comparisons, an absolute reduction of 8.84 points ($[6.31, 11.83]$), and a relative reduction of 12.93% ($[7.29, 18.70]\%$). The corresponding sign-flip estimates are $p = 8.9 \times 10^{-5}$ and $p = 0.0049$. Under this lower-bound assignment, the Transformer absolute effect is 7.43 points ($[2.59, 13.23]$; $p = 0.046$), while its relative effect remains unresolved. We retain the complete-case analysis as the headline result.

Table 7: Aggregate paired analysis of Table 1. Entries are means with 95% seed-bootstrap intervals stratified by task. Positive values favor ASR. Wins count observed baseline-minus-ASR differences above zero.

Comparison	Reduction (percentage points)	Relative reduction (%)	Wins
All ASR	$\mathbf{7.37}\,[\mathbf{5.71}, \mathbf{9.27}]$	$\mathbf{11.40}\,[\mathbf{5.87}, \mathbf{17.03}]$	$66/87$
Rényi-2	$\mathbf{7.32}\,[\mathbf{5.17}, \mathbf{9.78}]$	$\mathbf{12.80}\,[\mathbf{6.13}, \mathbf{19.45}]$	$22/29$
Spectral	$\mathbf{7.22}\,[\mathbf{5.78}, \mathbf{8.57}]$	$\mathbf{11.13}\,[\mathbf{3.54}, \mathbf{18.00}]$	$24/29$
Log-Det	$\mathbf{7.58}\,[\mathbf{4.85}, \mathbf{10.39}]$	$\mathbf{10.28}\,[\mathbf{1.18}, \mathbf{17.91}]$	$20/29$
MLP ASR family	$\mathbf{10.26}\,[\mathbf{8.91}, \mathbf{11.73}]$	$\mathbf{16.57}\,[\mathbf{12.57}, \mathbf{20.52}]$	$38/45$
Transformer ASR family	$\mathbf{4.28}\,[\mathbf{1.21}, \mathbf{7.94}]$	$5.87\,[-4.72, 16.80]$	$28/42$

G PAIRED SUPERCONDUCTIVITY EFFECTS

For the primary validation-selected test R^2, paired bootstrap gains over the baseline were 0.343 percentage points for Rényi-2 ($[0.071, 0.628]$), 0.290 for spectral entropy ($[0.076, 0.525]$), and 0.328

for Log-Det ($[0.112, 0.573]$). The intervals use 500,000 paired-seed resamples. The corresponding objectives improved seven, seven, and eight of ten paired runs, respectively.

H END-TO-END TIMING DETAIL

Table 8: End-to-end time to the first observed 99% test-accuracy threshold over five matched seeds. N is the number of uncensored runs. Steps and seconds are mean $\pm$ sample standard deviation over those N successful runs. Lower is better.

Arch.	Task	Objective	N	T_{grok} (10^3 steps)	Wall time (s)
MLP	ModAdd	None	5/5	33.7 ± 12.2	4.16 ± 1.54
		Rényi-2	5/5	13.5 ± 2.1	$\mathbf{2.17 \pm 0.42}$
		Log-Det	3/5	9.5 ± 3.3	$5.20 \pm 1.81^{\dagger}$
		Spectral	5/5	12.4 ± 4.8	34.87 ± 13.35
MLP	S_5	None	5/5	60.3 ± 15.0	12.30 ± 3.06
		Rényi-2	5/5	8.95 ± 2.60	$\mathbf{3.02 \pm 0.91}$
		Log-Det	5/5	39.8 ± 35.2	27.37 ± 24.35
		Spectral	5/5	22.1 ± 9.0	66.18 ± 26.98
MLP	Parity	None	5/5	1.09 ± 0.16	$\mathbf{0.15 \pm 0.02}$
		Rényi-2	5/5	2.29 ± 0.85	0.49 ± 0.19
		Log-Det	5/5	1.55 ± 0.48	0.93 ± 0.28
		Spectral	5/5	1.08 ± 0.09	3.09 ± 0.25
Transformer	ModStream	None	5/5	13.6 ± 10.8	52.46 ± 41.65
		Rényi-2	5/5	2.66 ± 2.69	$\mathbf{11.16 \pm 11.27}$
		Log-Det	4/5	2.96 ± 3.18	$13.77 \pm 14.74^{\dagger}$
		Spectral	5/5	2.56 ± 1.61	17.45 ± 10.91
Transformer	S5Stream	None	4/5	41.1 ± 5.2	$82.29 \pm 10.07^{\dagger}$
		Rényi-2	5/5	36.2 ± 6.0	83.51 ± 13.81
		Log-Det	5/5	31.9 ± 11.1	81.68 ± 28.35
		Spectral	5/5	34.6 ± 7.1	164.30 ± 33.06
Transformer	ParitySeq	None	5/5	1.27 ± 0.67	$\mathbf{26.17 \pm 13.78}$
		Rényi-2	5/5	3.21 ± 2.70	74.11 ± 62.14
		Log-Det	5/5	1.73 ± 1.04	40.94 ± 24.64
		Spectral	5/5	4.08 ± 2.72	100.34 ± 66.84

† Mean over successful runs; remaining runs are right-censored at the 100k-step budget.

I DERIVATION OF THE ROUTED LOG-DET IDENTITY

For $B < d$, choose $A = \widetilde{H}^{\top}/(\sqrt{B\epsilon})$ and $B' = \widetilde{H}/\sqrt{B\epsilon}$. Sylvester's identity gives

$$\det\left(\epsilon I_d + \frac{1}{B}\widetilde{H}^{\top}\widetilde{H}\right) = \epsilon^d \det(I_d + AB') \tag{19}$$

$$= \epsilon^d \det(I_B + B'A) \tag{20}$$

$$= \epsilon^{d-B} \det\left(\epsilon I_B + \frac{1}{B}\widetilde{H}\widetilde{H}^{\top}\right), \tag{21}$$

which yields Equation 13 after taking logarithms.

J NUMERICAL VALIDATION OF THE ROUTED COMPUTATION

We compared the routed Gram implementation with direct feature-covariance Log-Det on the same activation tensors, checking both values and gradients with respect to the activations. The double-precision routed checks validate the algebraic identity and its gradients to tight tolerances. In float32, the direct $B \geq d$ path agrees exactly in these checks, but the routed $B < d$ path becomes numerically sensitive at the chosen ridge $\epsilon = 10^{-5}$, especially for small batches. Thus the identity itself is exact, whereas the present float32 routed implementation should not be treated as a numerically interchangeable gradient computation in that regime. No reported training run uses that regime.

Table 9: Direct-versus-routed Log-Det validation. Entries are maximum absolute differences. Float32 value checks use `rtol`$= 10^{-5}$, `atol`$= 10^{-6}$; float32 gradient checks use `rtol`$= 10^{-4}$, `atol`$=$

10^{-5}. Float64 value checks use both tolerances 10^{-10}, and float64 gradient checks use $\texttt{rtol}= 10^{-8}$, $\texttt{atol}= 10^{-9}$.

dtype	B	d	route	$\lvert\Delta V\rvert_{\max}$	$\lvert\Delta\nabla_H V\rvert_{\max}$	value / gradient
float32	17	64	routed	$1.221\text{e}{-4}$	$1.422\text{e}{-1}$	pass / fail
float32	3	11	routed	$1.221\text{e}{-4}$	$2.629\text{e}0$	pass / fail
float32	64	17	direct	0	0	pass / pass
float32	128	128	direct	0	0	pass / pass
float32	11	3	direct	0	0	pass / pass
float64	17	64	routed	$1.592\text{e}{-12}$	$6.285\text{e}{-13}$	pass / pass
float64	3	11	routed	$2.615\text{e}{-12}$	$1.677\text{e}{-12}$	pass / pass