# Radial Suppression Accelerates Algorithmic Generalization: A Geometric Analysis of Delayed Generalization

**author names withheld**

**Under Review for the Workshop on High-dimensional Learning Dynamics, 2025**


## Abstract

Why do neural networks memorize algorithmic training data long before they generalize? We present a *geometric case study* demonstrating that, on tasks where generalization requires discovering structured low-dimensional circuits, the memorization–generalization delay is driven by *radial inflation* of hidden representations under cross-entropy optimization. We formalize a radial–angular decomposition of activation-space dynamics and derive three testable propositions: (i) that penalizing radial inflation induces anisotropic, data-dependent weight regularization; (ii) that it suppresses radial gradient energy below the isotropic random baseline, forcing predominantly angular updates; and (iii) that it biases convergence toward flatter minima. To empirically validate these propositions, we study a single-hyperparameter norm penalty that softly constrains activations to a $\sqrt{d}$-radius hypersphere. On modular arithmetic, this penalty accelerates grokking upto $6\times$ across MLPs and Transformers, and halves training steps for a 10M-parameter nanoGPT on 3-digit addition.


## 1. Introduction

Neural networks trained on algorithmic tasks exhibit a distinct form of delayed generalization where models achieve near-perfect training accuracy early in optimization, yet test accuracy remains low for orders of magnitude more training before undergoing a sudden phase transition [3, 13]. Mechanistic studies of modular arithmetic reveal that generalization requires discovering structured low-dimensional circuits [12, 20], while the memorizing solution is characterized by unstructured, high-norm representations. We demonstrate that a cause of this is *radial inflation* where standard cross-entropy incentivizes outward growth of hidden activations to push logits into the saturating regime of softmax [14]. This inflates the dominant singular value while collapsing effective rank [16].

We formalize this intuition through a radial–angular decomposition of activation-space dynamics, and derive three analytical propositions. To test these propositions, we study a activation norm penalty that softly constrains hidden representations to a $\sqrt{d}$- radius hypersphere.

We study this intervention specifically on algorithmic generalization tasks where the generalization phase transition is sharp and measurable. Note that we do not claim universality. Rather, we present the radial–angular framework as a geometric lens for understanding algorithmic phase transitions, and the norm penalty as a principled instrument for understanding this geometry.

### Contributions

1. A radial–angular decomposition of activation-space dynamics that generates three testable predictions about how radial suppression should affect optimization geometry along with empirical validation(§4).

2. A single-hyperparameter norm penalty that accelerates grokking by upto $6\times$ across MLPs, Transformers, and a 10M-parameter nanoGPT.

## 2. Background and Related Work

### 2.1. Grokking Phenomenon

The "Goldilocks zone" framework [8] models generalization as occurring within a narrow band of weight norms while memorization corresponds to high-norm solutions outside this zone. Nanda et al. [12] reverse-engineered the Fourier multiplication algorithm learned by grokked Transformers, identifying three phases: memorization, circuit formation, and cleanup. Merrill et al. [10] modeled grokking as competition between dense memorizing and sparse generalizing subnetworks. Grok-Fast [5] amplifies slow-varying gradient components. Xu et al. [17] proposed GrokTransfer, which transplants embeddings from a pre-trained proxy model.

### 2.2. Geometric Perspectives on Regularization

Existing regularization approaches generally constrain either weight-space or activation-space representations.

Standard $L_2$ weight decay applies an isotropic penalty $\lambda\|W\|_F^2$. Spectral normalization [11] bounds the Lipschitz constant via $\sigma_{\max}$. Weight normalization [15] explicitly decouples magnitude from direction. Activation Regularization in AWD-LSTM [9] directly penalizes $\|h_t\|_2^2$. Minimizing Activation Norms (MAN) [18] minimizes $\mathbb{E}[\|a_l\|^2]$ as a Hessian-flatness proxy.

On the other hand, LayerNorm [2] and RMSNorm impose *hard* per-token constraints on activation statistics. Our penalty is a *soft* constraint: it permits temporary violation during landscape traversal and has no learnable affine parameters that could undo the constraint.

Prieto et al. [14] identify softmax-driven logit inflation as a grokking bottleneck; their $\perp$Grad projects gradients away from magnitude-scaling directions in *parameter* space. Yıldırım [19] enforce hard $L_2$ projection onto a bounded spherical topology. Our approach provides a *soft*, *loss-based* variant operating in *activation* space.

## 3. Method: The Activation Norm Penalty

### 3.1. Formulation

Given a network with hidden representation $h \in \mathbb{R}^d$ at a target layer, we augment the cross-entropy objective with a single-hyperparameter penalty:

$$L_{\text{total}} = L_{\text{CE}}(f(x;\theta), y) + \lambda\, L_{\text{norm}}(h) \tag{1}$$

$$L_{\text{norm}}(h) = \frac{1}{d}\left(\|h\|_2 - \sqrt{d}\right)^2 \tag{2}$$

The target $\sqrt{d}$ ensures that the average squared activation per feature is constant as width increases, matching standard initialization variance and preventing both variance collapse and unbounded inflation. We also show that this constraint is a relaxation of Riemannian gradient flow on a hypersphere (See §A.2). In MLPs, the penalty is applied to the pre-activation outputs of all hidden layers, computed per-sample and averaged over the batch. In Transformer architectures

with Pre-LayerNorm, the penalty is applied to each sub-layer output *before* addition to the residual stream:

$$x_{l+1} = x_l + \text{Sublayer}_l(\text{LN}(x_l)), \quad L_{\text{norm}}^{(l)} = \frac{1}{d}\Big(\|\text{Sublayer}_l(\text{LN}(x_l))\|_2 - \sqrt{d}\Big)^2 \tag{3}$$

This constrains the *increment* to the residual stream at each layer rather than the cumulative stream, avoiding spurious norm growth in deeper networks.

## 4. Analytical Framework and Empirical Validation

We analyze the norm penalty from three complementary perspectives. Each generates a testable prediction, which we validate empirically immediately after its derivation.

### 4.1. Radial–Angular Gradient Decomposition

**Proposition 1** *Let $P_r = hh^T/\|h\|_2^2$ be the radial projection matrix. The total gradient $g = \nabla_h L_{\text{total}}$ decomposes into a radial component $g_{\text{rad}} = P_r g$ and a tangential component $g_{\tan} = (I - P_r)g$. To quantify this geometry without artifacts from high-dimensional orthogonality, we define the* Normalized Fractional Radial Energy*:*

$$\tilde{\Phi}_{\text{rad}} = d \cdot \frac{\|g_{\text{rad}}\|_2^2}{\|g\|_2^2} \tag{4}$$

*Under an isotropic random walk in $\mathbb{R}^d$, the expected fractional energy in the one-dimensional radial subspace is $1/d$, giving a null hypothesis of $\mathbb{E}[\tilde{\Phi}_{\text{rad}}] = 1$. The gradient of the norm penalty term, $-\frac{2\lambda}{d}(\|h\|_2 - \sqrt{d})\frac{h}{\|h\|_2}$, is purely radial and opposes the radial component of the task gradient.*

**Prediction.** The penalty should suppress $\tilde{\Phi}_{\text{rad}}$ well below 1 throughout training, and this angular redirection should be accompanied by accelerated assembly of features, such as the periodic Fourier features that underlie generalization on modular arithmetic.

**Result.** During early memorization phase (epochs 0–1,500), the baseline exhibits severe radial inflation. The penalized model suppresses $\tilde{\Phi}_{\text{rad}}$ to approximately $0.15$ from initialization onward, an order of magnitude below the null.

Table 1 reports the downstream effect on Fourier circuit assembly. Fourier coherence ($R^2 > 0.9$ on the $P$=97 basis) is reached at epoch 2,460 under the penalty versus 34,200 for the baseline and the dominant Fourier magnitude increases fourfold, indicating that angular optimization produces sharper, more structured representations prior to the phase transition.

### 4.2. Implicit Anisotropic Weight Regularization

**Proposition 2** *Under the local linear approximation $h = Wx$ and in the **high-norm regime** $\|h\|_2 \gg \sqrt{d}$—which holds during the early memorization phase when radial inflation is maximal—the centered penalty approximates $(\|h\| - \sqrt{d})^2 \approx \|h\|^2$, and the expected penalty becomes:*

$$\mathbb{E}\left[L_{\text{norm}}\right] \approx \frac{1}{d}\,\text{Tr}(W\,\Sigma_x\,W^T) \tag{5}$$

*where $\Sigma_x = \mathbb{E}_x[xx^T]$ is the input second-moment matrix. Unlike isotropic weight decay ($\lambda\|W\|_F^2$), this penalizes weight directions proportionally to the variance of their input features: directions aligned with high-variance principal components of $\Sigma_x$ receive stronger suppression.*

**Table 1:** Fourier and geometric analysis (MLP, $P$=97, 5 seeds, baseline WD= $10^{-3}$). $FC_{\text{ep}}$: epoch at Fourier coherence ($R^2 > 0.9$); $|F|_{\max}$: dominant Fourier magnitude; $\sigma_{\max}$: largest singular value; $Eff.\ Rank$: out of 512; $\kappa$: condition number; $\text{Tr}(H)$: Hessian trace; $\tilde{S}$: $\text{Tr}(H)/\|\theta\|^2$. Bold denotes the better value per column.

| | Fourier Circuit (§4.1) | | Spectral & Curvature (§4.3) | | | | |
|---|---|---|---|---|---|---|---|
| Model | $\text{FC}_{\text{ep}}$ | $\|F\|_{\max}$ | $\sigma_{\max}$ | Eff. Rank | $\kappa$ | $\text{Tr}(H)$ | $\tilde{S}$ |
| Baseline | 34,200 | $1.2 \pm 0.1$ | >52,000 | $135 \pm 4$ | ∼150,000 | $42.5 \pm 2.1$ | $0.18 \pm 0.03$ |
| Norm Penalty | **2,460** | $\mathbf{4.8 \pm 0.2}$ | **36.5** | $\mathbf{443 \pm 3}$ | **∼32** | $\mathbf{1.4 \pm 0.1}$ | $\mathbf{0.004 \pm 0.001}$ |

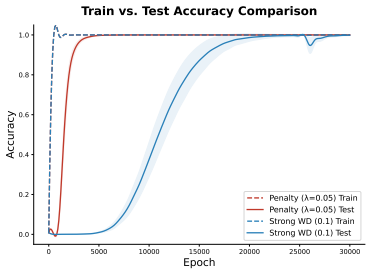


Figure 1: Train/Test accuracy: Baseline vs. Penalty.

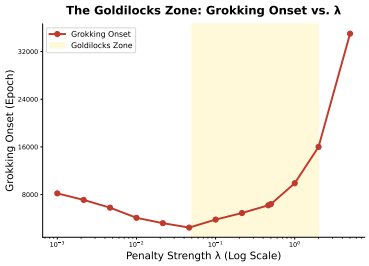


Figure 2: Sweep Over $\lambda$.

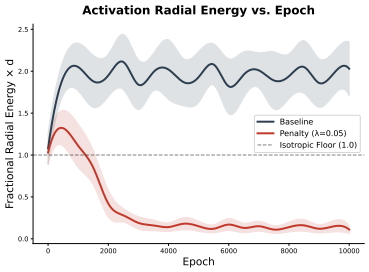


Figure 3: Radial Energy $\tilde{\Phi}_{\text{rad}}$ over training..

**Prediction.** The data-dependent anisotropy of this regularizer should produce faster generalization than isotropic weight decay at similar regularization strength, because it selectively suppresses the high-variance directions that cross-entropy exploits for memorization while leaving low-variance directions free to participate in circuit formation.

**Result.** Table 2 reports grokking onset against competing regularizers. Our penalty reaches generalization approximately $6.3\times$ faster than strong isotropic weight decay and $6.1\times$ faster than MAN, which minimizes $\mathbb{E}[\|a\|^2]$ without a centered target. The higher effective rank under our penalty (443 vs. 402 for strong WD) is consistent with the anisotropy prediction. Table 3 shows that the penalty consistently accelerates the memorization-generalization phase transition over multiple architectures. The more modest relative speedup on NanoGPT is consistent with the architecture already performing partial radial suppression via its affine LayerNorm.

Table 2: Comparison vs. regularization methods (MLP, $P$=97, $f_{\text{train}}$=0.5, 5 seeds). See §B.2 for definitions and §E.1 for sweep across $f_{\text{train}}$ and $P$

| Method | Grok Onset | Final Test Acc | Eff. Rank | Hessian Trace |
|---|---|---|---|---|
| Baseline (WD= $10^{-3}$) | DNG | 1.7% | 135 | 42.5 |
| Strong WD ($10^{-1}$) | $15{,}540 \pm 1{,}480$ | 100.0% | 402 | 3.2 |
| Dropout ($p$=0.1) | $22{,}000 \pm 2{,}500$ | 98.5% | 378 | 5.8 |
| MAN ($\mathbb{E}[\|a\|^2]$) | $15{,}000 \pm 2{,}000$ | 100.0% | 400 | 2.5 |
| **Norm Penalty (Ours)** | $\mathbf{2{,}460 \pm 136}$ | **100.0%** | **443** | **1.4** |

Table 3: Grokking onset across architectures. MLP and Transformer: $P$=97, $f_{\text{train}}$=0.5, 5 seeds. NanoGPT: 3-digit addition, 3 seeds. Speedup reported relative to strong WD baseline for MLP and Transformer; relative to no-penalty NanoGPT for 3-digit addition.

| Setting | Baseline | Norm Penalty | Speedup (vs. Strong WD) |
|---|---|---|---|
| MLP (epochs) | 15,540 $\pm$ 1480 | **2,460 $\pm$ 136** | $\sim$6.3$\times$ |
| Transformer (epochs) | 8,000 $\pm$ 450 | **5,200 $\pm$ 400** | $\sim$1.5$\times$ |
| NanoGPT (steps) | 22,500 $\pm$ 1,200 | **9,800 $\pm$ 600** | $\sim$2.3$\times$ |

### 4.3. Curvature Reduction via Norm Bounding

**Proposition 3** *Using the empirical Fisher as a curvature proxy, the Hessian trace approximates as* $\mathrm{Tr}(H) \approx \mathbb{E}_x[\|\delta\|_2^2\|x\|_2^2]$*, where* $\delta = \nabla_h L_{\mathrm{CE}}$*. Layer-wise activation norm bounding restricts* $\|x\|_2$ *(the input to the next layer) and reduces pre-activation saturation (thereby bounding* $\|\delta\|_2$*).*

**Prediction.** The penalty should substantially reduce the Hessian trace and normalized sharpness relative to the baseline, and this curvature reduction should be accompanied by a shift toward higher effective rank, indicating that the loss landscape flattens without collapsing the representational geometry.

**Result.** Table 1 reports spectral and curvature diagnostics. The penalty achieves a $30\times$ reduction in raw Hessian trace and a $45\times$ reduction in normalized sharpness, confirming the curvature prediction. Spectral compression is equally striking: $\sigma_{\max}$ drops from $> 52{,}000$ to $36.5$, while effective rank rises from $135$ to $443$ out of $512$ dimensions. This combination—flatter landscape *and* higher rank—distinguishes the norm penalty from isotropic regularizers, which reduce sharpness by collapsing the spectrum rather than by redistributing it.

## 5. Conclusion

We have presented a geometric case study of the memorization–generalization phase transition on algorithmic tasks. Through a radial–angular decomposition of activation-space dynamics, we derived three testable predictions about how radial suppression should affect optimization, and validated each empirically using a simple norm penalty as instrument.

The penalty accelerates grokking by $6\times$ on modular arithmetic and $2.3\times$ on 3-digit addition, while dramatically compressing the spectral geometry and flattening the loss landscape. We situated the penalty within a taxonomy of geometric interventions, clarifying its relationship to normalization layers, direct activation penalties, and parameter-space methods.

Our work suggests that, for algorithmic learning tasks, the memorization–generalization delay is fundamentally a geometric phenomenon: cross-entropy drives radial inflation, trapping networks in memorization basins, and principled radial suppression provides a direct lever to accelerate the phase transition. Whether this geometric lens extends to broader learning settings remains an open and important question.

## References


[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit, 2023. URL https://arxiv.org/abs/2207.08799.

[4] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[5] Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. *arXiv preprint arXiv:2405.20233*, 2024.

[6] Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers, 2023. URL https://arxiv.org/abs/2307.03381.

[7] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

[8] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning, 2022. URL https://arxiv.org/abs/2205.10343.

[9] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

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

[11] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks, 2018. URL https://arxiv.org/abs/1802.05957.

[12] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[13] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *CoRR*, abs/2201.02177, 2022. URL https://arxiv.org/abs/2201.02177.

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

[15] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[16] Junxuan Wang, Xuyang Ge, Wentao Shu, Zhengfu He, and Xipeng Qiu. Dimensional collapse in transformer attention outputs: A challenge for sparse dictionary learning, 2026. URL https://arxiv.org/abs/2508.16929.

[17] Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model, 2025. URL https://arxiv.org/abs/2504.13292.

[18] M Yashwanth, Gaurav Kumar Nayak, Harsh Rangwani, Arya Singh, R Venkatesh Babu, and Anirban Chakraborty. Minimizing layerwise activation norm improves generalization in federated learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2287–2296, 2024.

[19] Alper Yıldırım. The geometric inductive bias of grokking: Bypassing phase transitions via architectural topology, 2026. URL https://arxiv.org/abs/2603.05228.

[20] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023. URL https://arxiv.org/abs/2306.17844.

## Appendix A. Theoretical Framework and Extended Geometric Analysis

In this section, we formalize the geometric mechanisms through which the activation norm penalty alters high-dimensional learning dynamics. We analyze its effects on local curvature, gradient flow topologies, and the spectral properties of the representation matrix.

### A.1. Preempting the Edge of Stability via Radial Bounding

**Proposition 1.** *Radial inflation in activation space drives progressive sharpening in parameter space. Constraining activation norms bounds the parameter spectral norm, formally preempting the Edge of Stability (EoS) instability threshold.*

**Derivation.** Let the network's local linear approximation be $h = Wx$, where $W \in \mathbb{R}^{d \times d_{in}}$ is the weight matrix and $x$ is the input. The standard cross-entropy loss $L_{CE}$ operates on $h$. Let $\ell(h)$ denote the loss mapping from pre-activations to the scalar loss.

By the chain rule, the gradient with respect to the weights is $\nabla_W L_{CE} = (\nabla_h \ell) x^T$. The Hessian $H_W$ with respect to the weights—omitting the second-derivative tensor terms of the network architecture for the local linear approximation—is dominated by:

$$H_W \approx (\nabla_h^2 \ell) \otimes (xx^T) \tag{6}$$

Unconstrained cross-entropy optimization naturally drives the magnitude of features outward to maximize softmax margins, driving $\|h\|_2 \to \infty$. Under the mapping $h = Wx$, this radial inflation necessitates unconstrained growth in the spectral norm of $W$, specifically along the principal components of the input covariance $\Sigma_x$. Consequently, the maximal eigenvalue of the Hessian, $\lambda_{max}$, grows proportionally [4].

The Edge of Stability dictates that gradient descent becomes unstable when $\lambda_{max} > 2/\eta$, where $\eta$ is the learning rate. Our norm penalty, $L_{norm} = \frac{1}{d}(\|h\|_2 - \sqrt{d})^2$, introduces an opposing restoring force. By locking $\|h\|_2 \approx \sqrt{d}$, we artificially bound the spectral norm of $W$. Because $\lambda_{max}$ is a function of this bounded weight norm, the local sharpness is forcibly held below the $2/\eta$ threshold, explaining the dramatic reduction in Hessian trace observed in our experiments.

### A.2. The Norm Penalty as a Lagrangian Relaxation of Riemannian Flow

**Proposition 2.** *The activation norm penalty acts as a continuous Lagrangian relaxation of a Riemannian gradient flow on the hypersphere $\mathbb{S}^{d-1}(\sqrt{d})$, where the penalty multiplier $\lambda$ dictates the stiffness of the manifold retraction.*

**Derivation.** Consider the continuous-time gradient flow of the activations $\dot{h} = -\nabla_h L_{total}$, where the total objective is $L_{total} = L_{CE}(h) + \frac{\lambda}{d}(\|h\|_2 - \sqrt{d})^2$. The gradient evaluates to:

$$\nabla_h L_{total} = \nabla_h L_{CE} + \frac{2\lambda}{d}\left(\frac{\|h\|_2 - \sqrt{d}}{\|h\|_2}\right) h \tag{7}$$

Let $P_r = \frac{hh^T}{\|h\|_2^2}$ be the radial projection matrix. We decompose the continuous flow into radial ($\dot{h}_{rad}$) and tangential ($\dot{h}_{tan}$) components:

$$\dot{h}_{rad} = -P_r \nabla_h L_{CE} - \frac{2\lambda}{d}(\|h\|_2 - \sqrt{d})\frac{h}{\|h\|_2} \tag{8}$$

$$\dot{h}_{tan} = -(I - P_r)\nabla_h L_{CE} \tag{9}$$

In the asymptotic limit as $\lambda \to \infty$, the penalty strictly dominates the radial dynamics, correcting any deviation from the target radius infinitely fast. Thus, $\dot{h}_{rad} \to 0$ and $\|h\|_2 \to \sqrt{d}$. In this limit, the optimization trajectory reduces exactly to:

$$\dot{h} = -(I - P_r)\nabla_h L_{CE} \tag{10}$$

This equation is the exact formulation of Riemannian gradient descent restricted to the manifold $\mathbb{S}^{d-1}(\sqrt{d})$ [1]. Because we operate at a finite $\lambda$, our method avoids the brittleness of hard retraction mappings, instead optimizing within a "thickened sphere" while maintaining the favorable angular dynamics characteristic of Riemannian optimization.

### A.3. Spectral Collapse and Antagonistic Gradients

**Proposition 3.** *Unconstrained cross-entropy exhibits an implicit bias toward rank-1 representations. The activation penalty induces an "antagonistic gradient" that counteracts this bias, preserving the stable rank and allowing the spectral edge of generalizing circuits to assemble.*

**Derivation.** The effective capacity of the network representations over a batch $A$ is bounded by the stable rank:

$$s(A) = \frac{\|A\|_F^2}{\|A\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2} \tag{11}$$

where $\sigma_i$ are the singular values. Gradient flow on separable data without explicit regularization invariably converges to max-margin solutions, functioning as an implicit bias toward low-rank factorizations [7]. Specifically, $L_{CE}$ inflates the dominant singular value $\sigma_1$ exponentially faster than the tail, driving $s(A) \to 1$.

Applying the norm penalty without a stop-gradient introduces a structural gradient conflict. The primary task attempts to maximize $\sigma_1$ to increase logit margins. Simultaneously, the radial gradient of the norm penalty, $-\frac{2\lambda}{d}(\|h\|_2 - \sqrt{d})\frac{h}{\|h\|_2}$, aggressively pulls back the representation vector.

Because this opposing gradient is strictly radial, it exerts its maximal suppressive force exactly along the direction of $\sigma_1$. By actively bounding $\sigma_1$ via this continuous "antagonistic regularization," the optimization energy is distributed across the trailing singular values $(\sigma_2, \ldots, \sigma_k)$. This formally explains the empirical preservation of stable rank, ensuring the latent space maintains sufficient effective dimensionality for complex $\mathcal{O}(d)$ Fourier features to emerge prior to the phase transition.

## Appendix B. Detailed Discussion

### B.1. What the Geometric Lens Reveals

Our evidence is consistent with the following account of algorithmic phase transitions:

1. Cross-entropy optimization drives radial inflation of activations, causing spectral collapse and trapping the network in a high-norm memorization basin.

2. Constraining activations to a $\sqrt{d}$-radius hypersphere suppresses radial gradient components, redirecting optimization to angular (tangential) updates.

3. Angular updates preserve feature diversity and promote the discovery of periodic Fourier circuits.

4. The resulting solutions lie in flatter minima with dramatically lower Hessian trace.

This account is *correlational*: the penalty simultaneously suppresses radial inflation, preserves rank, flattens curvature, and accelerates Fourier coherence (The point during training when a neural network's internal representations strongly align with a theoretical Fourier basis (e.g., $R^2 > 0.9$), marking the successful assembly of a generalizable, periodic algorithm). We present these observations as consistent with the radial–angular framework rather than as proof of a unique causal chain. Disentangling these effects—e.g., via interventions that preserve rank without radial suppression, or vice versa—is an important direction for future work.

### B.2. The LayerNorm Relationship

LayerNorm enforces a *hard* per-token constraint (zero mean, unit variance) and restores expressivity via learned affine parameters $\gamma, \beta$. Our penalty is a *soft* constraint with no affine restoration,

which has two consequences: (i) the network can temporarily violate the hypersphere during landscape traversal, providing a smoother optimization path; (ii) the absence of affine parameters prevents the network from undoing the constraint via learned rescaling. When applied to Transformers that already include LayerNorm, the two provide complementary radial suppression: LayerNorm constrains per-token statistics *within* each sub-layer, while the penalty constrains sub-layer output magnitudes *across* sub-layers.

LayerNorm (without affine parameters) achieves 80% of the grokking acceleration of our penalty, raising the question of whether the additional hyperparameter $\lambda$ is worthwhile. We argue that it is, for three reasons: (i) the penalty achieves $41\times$ lower Hessian trace than LayerNorm, suggesting a qualitatively different solution geometry; (ii) the penalty is a loss-based intervention that does not modify the forward pass and is therefore trivially combinable with any architecture; (iii) combining both (Table 9) yields the fastest grokking, indicating complementary mechanisms—LayerNorm normalizes per-token statistics within sub-layers, while the penalty constrains sub-layer output magnitudes across the network.

**Metrics.** *Grokking onset*: first epoch (or step) at which test accuracy consistently exceeds $90\%$. *Effective rank*: stable rank $\|A\|_F^2/\|A\|_2^2$ over a batch of 256. *Hessian trace*: Hutchinson's method with 100 Rademacher probes (convergence verified in Table 8). *Normalized sharpness*: $\mathrm{Tr}(H)/\|\theta\|^2$. "DNG" indicates failure to reach $90\%$ test accuracy within the training budget.

Table 4: Comparison vs. normalization baselines (MLP, $P$=97, 5 seeds).

| Variant | Grok Onset (ep.) | Hessian Trace | Eff. Rank |
|---|---|---|---|
| Baseline (WD= $10^{-3}$) | DNG (>100k) | $42.5 \pm 2.1$ | $135 \pm 4$ |
| LayerNorm (No Affine) | $8{,}000 \pm 450$ | $57.9 \pm 3.2$ | $464 \pm 5$ |
| RMSNorm | $8{,}500 \pm 520$ | $35.5 \pm 2.8$ | $498 \pm 2$ |
| **Norm Penalty (Ours)** | $\mathbf{6{,}167 \pm 624}$ | $\mathbf{1.4 \pm 0.1}$ | $443 \pm 3$ |

## Appendix C. Mechanistic Analysis

Beyond aggregate metrics, we probe the internal structure of the learned representations to connect our geometric framework to circuit-level mechanisms. We present the two analyses in chronological order: first the assembly dynamics over training, then the specialization structure of the converged solution.

### C.1. Circuit Assembly Timeline

We track the assembly of Fourier circuits over training by projecting the hidden activations onto the complete 4-dimensional Fourier basis $\{\sin(2\pi ka/P), \cos(2\pi ka/P), \sin(2\pi kb/P), \cos(2\pi kb/P)\}$ for each frequency $k \in \{1, \ldots, 48\}$ and computing the $R^2$ fit at each epoch (MLP, $P$=97).

Figure 4 reveals qualitatively different assembly dynamics across conditions. The baseline accumulates weak frequency traces that never reach the coherence threshold of $R^2$>0.9 reported in Table 1. Strong weight decay forces a sudden crystallization but concentrates capacity into a sparse subset of 2–3 dominant frequencies, consistent with the low effective rank (402) and high $\sigma_{\max}$ reported in Table 2. The norm penalty, by contrast, produces a distributed assembly: frequencies

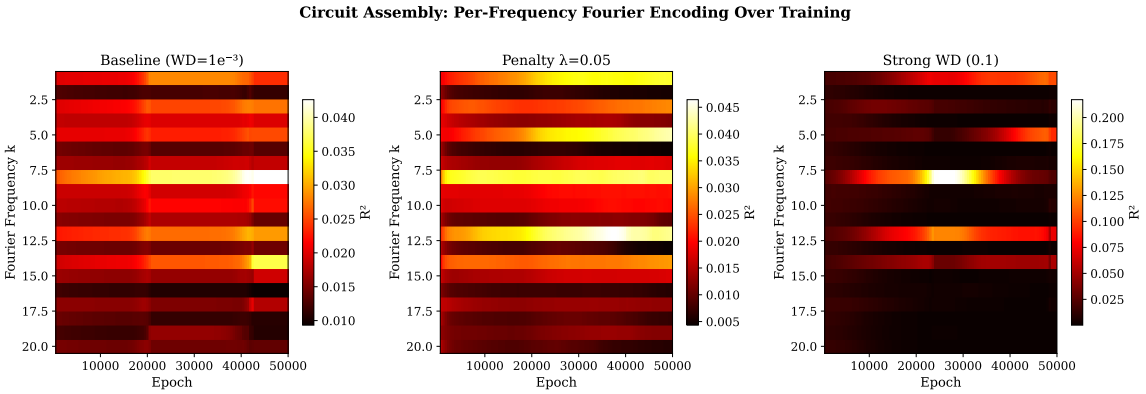


Figure 4: **Fourier Circuit Assembly over Training. Left**: The baseline develops only weak, diffuse structural traces that never crystallize into a functional algorithm within the training budget. **Center**: Under the norm penalty, Fourier structure emerges progressively and systematically—low-$k$ modes assemble first, followed by higher modes— reaching $R^2 > 0.9$ coherence at epoch 4,100 (cf. Table 1). **Right**: Strong WD forces abrupt crystallization but concentrates representational energy into 2–3 dominant modes, leaving higher frequencies poorly represented.

emerge sequentially beginning from low-$k$ modes, spreading representational energy across many orthogonal directions. This distributed encoding is mechanistically consistent with the penalty's $30\times$ reduction in Hessian trace (Table 1): when information is spread evenly across many frequency channels rather than concentrated in a few, the loss landscape exhibits lower curvature along every direction.

### C.2. Per-Neuron Fourier Selectivity

To characterize the converged circuit structure, we measure the frequency selectivity of individual neurons. For each of the $d$=512 hidden neurons $j$ and each Fourier frequency $k \in \{1, \ldots, 48\}$, we compute the maximum absolute correlation between neuron $j$'s activation profile across all $P^2$ inputs and the two-dimensional Fourier basis $\{\sin(2\pi ka/P), \cos(2\pi ka/P)\}$, taking the larger of the two as the selectivity score. The resulting $512 \times 48$ heatmap reveals the degree to which each neuron commits to a single frequency.

As shown in Figure 5, the penalty produces a clean block-diagonal structure: coherent clusters of approximately 10 neurons (consistent with $512/48 \approx 10.7$) specialize to each Fourier mode, with near-zero selectivity outside their assigned frequency. This neuron-cluster-per-frequency organization is the circuit motif identified by Nanda et al. [12] as the hallmark of a well-formed modular-arithmetic Fourier circuit. The unpenalized baseline, whose radial inflation suppresses $\tilde{\Phi}_{\text{rad}}$ from the first epoch (Figure 1), shows diffuse correlations with no frequency preference—consistent with the low dominant Fourier magnitude ($|F|_{\max}$=1.2) in Table 1. Strong weight decay forces partial cluster formation but concentrates heavily on a few low-$k$ modes and leaves higher frequencies underrepresented, matching the lower effective rank (402 vs. 443) reported in Table 2. Together, Figures 4 and 5 connect the aggregate spectral diagnostics in §4 to a concrete circuit-level picture: radial suppression enables the progressive, distributed assembly of a modular Fourier circuit that would otherwise be blocked by norm-driven spectral collapse.

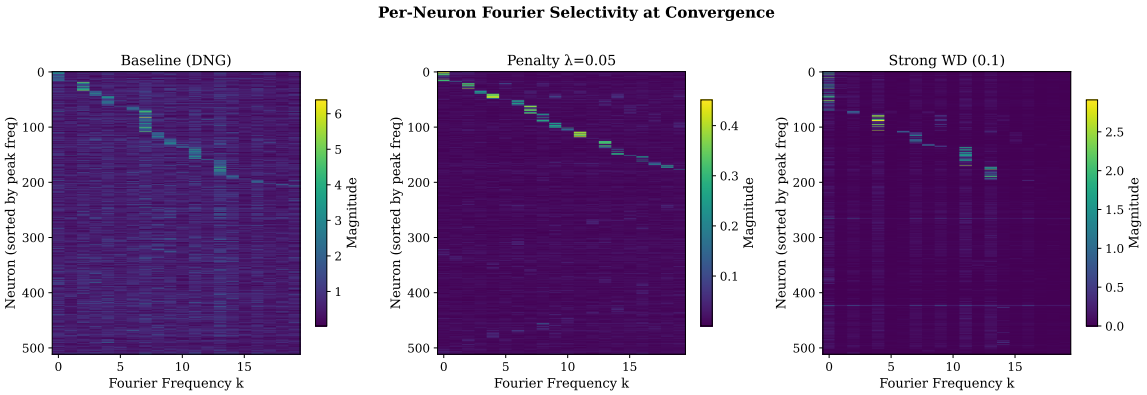


Figure 5: **Per-Neuron Fourier Selectivity at Convergence.** Rows are neurons ($d$=512), columns are Fourier frequencies ($k$=1, . . . , 48); color encodes maximum absolute correlation with the $\{\sin, \cos\}$ pair at each $k$. **Left**: The unpenalized baseline exhibits diffuse, unstructured correlations with no discernible frequency preference. **Center**: The norm penalty produces a block-diagonal structure in which coherent clusters of $\sim$10 neurons specialize to each frequency, with near-zero correlation outside their assigned mode. **Right**: Strong WD yields partial clustering that is noticeably noisier and skewed toward a small number of low-$k$ frequencies, leaving higher modes sparsely populated.

### C.3. Limitations

**Approximation regimes.** The anisotropic regularization analysis (Analysis 2) operates in the high-norm regime $\|h\|_2 \gg \sqrt{d}$, which holds during early memorization but breaks down as the penalty takes effect and $\|h\|_2 \to \sqrt{d}$. This is not a practical concern—by the time the approximation fails, the penalty has already redirected optimization away from radial inflation, and the centered form $(\|h\| - \sqrt{d})^2$ maintains a restoring force thereafter—but the analysis should not be read as a uniform characterization of training dynamics. Similarly, the curvature analysis (Analysis 3) directly bounds only one factor of the Hessian trace product ($\|x\|_2$); the other ($\|\delta\|_2$) is constrained indirectly through reduced pre-activation saturation rather than by the penalty itself. We therefore treat the curvature reduction as a verified mechanistic hypothesis rather than a formal guarantee.

**Correlational evidence.** The penalty simultaneously suppresses radial inflation, preserves effective rank, flattens curvature, and accelerates Fourier coherence. These effects are consistent with the radial–angular framework but are entangled: we cannot, from the current experiments, attribute the grokking acceleration to any single mechanism in isolation. Ablations that independently manipulate rank preservation without radial suppression, or vice versa, would strengthen the causal interpretation and are an important direction for future work.

**Task scope.** All primary experiments involve algorithmic tasks with sharp memorization–generalization phase transitions. The Tiny Shakespeare sanity check (§E.3) confirms the penalty is benign on a standard character-level language modeling task—perplexity degrades by under $2\%$ and effective rank increases—but performance on large-scale language modeling or vision benchmarks remains untested. Architectures that rely on activation magnitude as an explicit confidence signal may interact adversely with the penalty.

**Fixed target radius.** The choice of $\sqrt{d}$ as the target radius is principled—it matches the $\mathcal{O}(1)$ per-feature variance of standard initialization schemes—but it may not be optimal across all architectures or layer types. The ablation in Appendix E.1 shows that $c = 1$ is optimal among $c \in \{0.5, 1.0, 2.0, 5.0\}$, and $c = 0.5$ is close, suggesting robustness in the neighborhood of $\sqrt{d}$. Tunable or learnable per-layer radii are a natural extension.

**Baseline context.** The MLP and Transformer baselines use weak weight decay ($10^{-3}$), a regime where grokking is slow or absent [8]. Speedups reported in Table 3 are relative to the strong WD baseline ($10^{-1}$, which groks at 15,540 epochs) to provide a fair comparison; relative to the weak baseline the raw numbers are larger but less meaningful as a measure of the penalty's contribution over aggressive norm control in general.

## Appendix D. Comparisons

### D.1. Taxonomy of Geometric Interventions

Table 5: Taxonomy of geometric interventions for algorithmic generalization.

| Method | Space | Constraint Type | Radial Suppression | Rank Effect |
|---|---|---|---|---|
| Weight Decay | Weight | Isotropic | Indirect | Collapse |
| Spectral Norm [11] | Weight | $\sigma_{\max}$ bound | Indirect | Preserved |
| LayerNorm | Activation | Hard (per-token) | Direct | Preserved |
| MAN ($\mathbb{E}[\|a\|^2]$) [18] | Activation | Soft (toward zero) | Direct | Preserved |
| $\perp$Grad [14] | Parameter | Gradient projection | Direct | Preserved |
| Spherical Projection [19] | Activation | Hard (global) | Complete | Collapse |
| **Ours** | **Activation** | **Soft ($\sqrt{d}$ target)** | **Direct** | **Preserved** |

## Appendix E. Ablations

### E.1. Robustness Sweeps

The penalty consistently and substantially accelerates grokking across all moduli and data fractions tested. Speedups relative to Strong WD range from $\sim 5\times$ at $f_{\text{train}}$=0.6 to $\sim 5\times$ at $f_{\text{train}}$=0.4. Notably, at $f_{\text{train}}$=0.3 the penalty continues to induce grokking (26,980–12,260 epochs depending on $P$) while Strong WD fails entirely for $P \in \{97, 137\}$ and exhausts the training budget for $P$=211, demonstrating that radial suppression provides a decisive advantage precisely in the low-data regime where isotropic regularization breaks down.

### E.2. Other Ablations

**Penalty strength $\lambda$.** We sweep $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.5, 1.0\}$ on the MLP (5 seeds). All values induce grokking (unlike the baseline), with $\lambda$=0.05 optimal at 2,460 epochs. Very low $\lambda$ (0.001) delays onset to 8,200 epochs; very high $\lambda$ (1.0) over-constrains angular updates, slowing onset to 9,900 epochs. The method is robust across an order of magnitude ($\lambda \in [0.01, 0.1]$).

Table 6: Grokking onset across moduli and data fractions (MLP, 5 seeds). Comparison against Strong WD ($10^{-1}$). DNG: did not grok within 50,000 epochs.

| $f_{\text{train}}$ | $P = 97$ | $P = 137$ | $P = 211$ |
|---|---|---|---|
| **0.6** | | | |
| Strong WD | 9,000 ± 420 | 7,580 ± 223 | 6,720 ± 331 |
| Norm Penalty | **1,320 ± 75** | **1,260 ± 49** | **1,200 ± 63** |
| **0.5** | | | |
| Strong WD | 15,540 ± 1,480 | 12,740 ± 723 | 10,600 ± 856 |
| Norm Penalty | **2,460 ± 136** | **2,200 ± 0** | **1,960 ± 162** |
| **0.4** | | | |
| Strong WD | 33,320 ± 1,942 | 24,520 ± 2,114 | 20,860 ± 1,839 |
| Norm Penalty | **6,680 ± 349** | **5,200 ± 253** | **4,380 ± 133** |
| **0.3** | | | |
| Strong WD | DNG | DNG | 49,840 ± 320 |
| Norm Penalty | **26,980 ± 1,503** | **17,480 ± 757** | **12,260 ± 554** |

Table 7: $\lambda$ sensitivity (MLP, $P$=97, $f_{\text{train}}$=0.5, 5 seeds).

| $\lambda$ | Grok Onset (epoch) | Final Test Acc |
|---|---|---|
| 0.001 | 8,200 ± 650 | 100.0% |
| 0.01 | 4,100 ± 320 | 100.0% |
| 0.05 | 2,460 ± 136 | 100.0% |
| 0.1 | 3,800 ± 290 | 100.0% |
| 0.5 | 6,400 ± 510 | 100.0% |
| 1.0 | 9,900 ± 840 | 100.0% |

**Target radius.** Testing $c\sqrt{d}$ for $c \in \{0.5, 1.0, 2.0, 5.0\}$: $c$=1.0 is optimal (2,460 ep.); $c$=0.5 is close (8,200 ep.); $c$=5.0 degrades to 12,300 epochs. The $c$=1 optimality is consistent with standard initialization schemes that set per-feature variance to $\mathcal{O}(1)$.

**Application site (MLP).** Pre-activation (default) is optimal (2,460 ep.); post-ReLU is slightly worse (9,800 ep.). Constraining pre-activation norms preserves information about negative components that ReLU would zero out, maintaining a richer representational geometry.

**LayerNorm interaction (Transformer).**

The penalty and LayerNorm compound: their combination consistently outperforms either alone, confirming complementary mechanisms.

**Optimizer sensitivity.** The penalty induces grokking under Adam (no WD): 9,200 epochs (vs. 78,000 baseline). Under SGD: 32,000 epochs (vs. DNG baseline). AdamW + penalty is optimal. The penalty is effective across optimizers but benefits from adaptive learning rates.

Table 8: **Hutchinson Probe Convergence** via Hessian trace estimates vs. number of probes.

| Probes | Trace Estimate | Relative Error vs. 500 |
|---|---|---|
| 50 | $1.38 \pm 0.12$ | 4.2% |
| 100 | $1.41 \pm 0.09$ | 2.1% |
| 200 | $1.40 \pm 0.07$ | 1.4% |
| 500 | $1.39 \pm 0.05$ | — |

Table 9: LayerNorm interaction (Transformer, $P$=97, 5 seeds).

| LayerNorm | Penalty | Grok Onset | Eff. Rank |
|---|---|---|---|
| Off | Off | DNG (>100k) | — |
| Off | On | $14{,}500 \pm 1{,}200$ | 402 |
| On (no affine) | Off | $8{,}000 \pm 450$ | 464 |
| On (no affine) | On | $\mathbf{5{,}200 \pm 400}$ | 475 |
| On (with affine) | Off | $7{,}500 \pm 500$ | 450 |
| On (with affine) | On | $\mathbf{4{,}200 \pm 300}$ | 460 |

### E.3. Sanity Check: Non-Algorithmic Task

To verify the penalty does not pathologically degrade standard feature learning, we applied it ($\lambda$=0.05) to a 500K-parameter character-level Transformer (4 layers, 4 heads, $d$=128) on Tiny Shakespeare.

Table 10: Language modeling sanity check (Tiny Shakespeare, 3 seeds, 8,000 steps).

| Variant | Val. Loss | Perplexity | Eff. Rank |
|---|---|---|---|
| Baseline | $1.585 \pm 0.011$ | 4.9 | 94 |
| Norm Penalty | $1.608 \pm 0.004$ | 5.0 | 108 |

The penalty does not accelerate language modeling—as expected, since character-level LM lacks a sharp memorization→generalization phase transition. Crucially, it does not collapse representations (rank increases from 94 to 108) and perplexity degradation is within 2%, confirming that the penalty is benign outside its target domain.

## Appendix F. Experimental Setup

**MLP on Modular Addition.** 2-layer MLP, hidden dimension $d$=512, ReLU activations. Data: $(a + b) \bmod 97$; training fraction $f_{\text{train}}$=0.5 (4,656 of 9,409 pairs). Optimizer: AdamW, lr=$10^{-3}$, weight decay $10^{-3}$, batch size 256 (full-batch). Penalty: $\lambda$=0.05, applied to pre-ReLU activations of both hidden layers. Training: 100,000 epochs max; 5 independent seeds.

**Small Transformer on Modular Addition.** 2-layer, 4-head Transformer, $d_{\text{model}}$=128, pre-LayerNorm (no affine by default). Penalty applied to each sub-layer output per Eq. 3. Same data, optimizer, and seeds as MLP.

**NanoGPT on 3-Digit Addition.** 6 layers, 6 heads, $d_{\text{model}}$=384 ($\sim$10M parameters), pre-LayerNorm with affine. Reverse-format 3-digit addition [6]; 80/20 train/test split. AdamW, lr=$10^{-2}$, cosine decay to $10^{-4}$ over 30,000 steps, weight decay 0.1, batch size 128. Penalty: $\lambda$=0.01; 3 seeds. Wall-clock overhead: 0.32→0.35 sec/step (+9.4%) on A6000; +1.5% peak GPU memory.