Understanding Learning Dynamics via Symmetry-Aware Spectra

Anonymous authors
Paper under double-blind review

Abstract

Overparameterized neural networks in certain settings have been observed to generalize long after perfect memorization, a phenomenon known as grokking. What drives this transition, and how to detect it during training, remains only partially understood. Common diagnostics are defined on the model weights, but parameter symmetries allow the same function to be represented by weight matrices of different norms and ranks. We study instead the singular-value spectrum of the parameter Jacobian $\partial f_\theta / \partial \theta$, called the *neural spectrum*, as a symmetry aware measure of the locally accessible functional directions. We find that across different architectures, grokking is accompanied by *spectral polarization*: sensitivity becomes concentrated in a small effective subspace. Comparing standard, NTK, and maximal-update (μP) parameterizations shows that the timing and form of this restructuring depend strongly on the feature-learning regime. Moreover, regularizing to encourage spectral concentration accelerates grokking, while encouraging a diffuse spectrum deters grokking. Beyond grokking, harder learnable tasks across four supervised-learning benchmarks consistently produce broader, higher-rank spectra. These results support the neural spectrum as a common diagnostic of feature-learning dynamics and learned functional complexity.

1 Introduction

A central question in deep learning is why large neural networks eventually learn comparatively simple solutions (Belkin et al., 2019; Power et al., 2022). One way to formalize this question is through the *functional dimension*. For a model f_θ evaluated on inputs Z, the parameter Jacobian $J_Z(\theta) = \frac{\partial f_\theta(Z)}{\partial \theta}$ is the differential of the parameter-to-function map. Its rank is the batch functional dimension (Grigsby et al., 2025): the number of independent ways infinitesimal parameter perturbations can change the model's outputs on Z. Thus, functional dimension measures local complexity in function space rather than in parameter space. Prior work has connected functional dimension to parameter-space symmetries (Stock and Gribonval, 2023; Grigsby et al., 2025; Bona-Pellissier et al., 2022; Zhao et al., 2025; Marchetti et al., 2025), but rank alone is too coarse to describe how the local parameter-to-function geometry evolves during learning.

In overparameterized models, the Jacobian may remain numerically full rank even when most directions barely change the output. We therefore study the full singular-value spectrum of J_Z, which called the *neural spectrum* Kopitkov and Indelman (2020). Its singular values quantify the strength of locally accessible functional directions, while its rank recovers functional dimension. The spectrum thus provides a graded notion of *effective functional dimension*, distinguishing diffuse sensitivity across many directions from concentration into a small effective subspace. Since the squared singular values are the eigenvalues of the empirical NTK $K = JJ^\top$, spectral evolution also provides a direct view of the transition from lazy training, where the Jacobian remains near initialization, to rich feature learning. Because the parameter Jacobian is defined for any differentiable model, its spectrum provides a common diagnostic for comparing training dynamics and learned complexity across architectures.

Grokking (Power et al., 2022) provides a particularly sharp setting in which to study this geometry. In grokking, networks memorize the training set quickly but generalize only after

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.

a long delay. Prior work has associated grokking with changes in weight norm, sparsity, low-rank parameter structure, and the emergence of simple algorithmic or architectural representations (Nanda et al., 2023; Doshi et al., 2024; Furuta et al., 2024; Yunis et al., 2024). These results often describe grokking in terms of the weights, but parameter-space symmetries allow the same function to be realized by weights of very different norms and ranks, making these less useful as diagnostics. We therefore ask a complementary question for function space: *how does the local space of functions accessible to the network reorganize as memorization gives way to generalization?*

Across Transformers, MLPs, LSTMs, and CNNs trained on modular multiplication, we find that grokking is accompanied by a pronounced restructuring of the neural spectrum: a small number of singular values become dominant. We call this *spectral polarization*. In functional terms, the model's sensitivity becomes concentrated in a smaller effective subspace.

We then test whether this phenomenon reflects feature learning. Under standard parameterization (SP), polarization and grokking occur together; under maximal update parameterization (μP), generalization precedes full polarization, suggesting that polarization reflects convergence toward a low effective functional dimension rather than a prerequisite for generalization; and under NTK parameterization, the spectrum remains close to initialization and grokking does not occur. These observations alone do not establish that spectral restructuring contributes to the transition. We therefore intervene directly on the spectrum during training. Under SP, encouraging concentration of Jacobian spectral mass causes polarization to occur earlier and accelerates grokking, whereas encouraging a diffuse spectrum preserves training-set memorization but prevents grokking within the training budget.

Finally, we test whether spectral rank captures functional complexity on four families of tasks with systematically varying difficulty: Hermite polynomial regression, DCT-MNIST, noisy MNIST, and QM9 molecular property prediction. Across these settings, harder tasks consistently produce higher spectral rank at convergence, supporting the neural spectrum as a way to understand learning dynamics and learned functions.

Contributions.

1. We develop the *neural spectrum* as a graded, architecture-agnostic measure of effective functional dimension, providing a common way to quantify and compare learned functional complexity across models.
2. We identify *spectral polarization*, the concentration of parameter-to-function sensitivity into a small effective subspace, and show that it accompanies grokking across four different architectures (§4.1).
3. We show that the timing and form of polarization depend on the feature-learning regime. Comparisons among SP, NTK, and μP, and interpolation between them, reveal different relationships between spectral restructuring and generalization (§4.2).
4. We directly manipulate spectral polarization during training. Encouraging concentration accelerates grokking, while encouraging a diffuse spectrum deters grokking, providing evidence that spectral restructuring can influence the transition (§4.4).
5. We show across four supervised-learning benchmark families that harder learnable tasks generally produce broader, higher-rank neural spectra, supporting spectral rank as a diagnostic of learned functional complexity beyond grokking (§4.5).

2 Related Work

Parameter space symmetries and loss landscape geometry. Prior work has studied neural-network function spaces and loss landscapes through parameter-space symmetries and non-identifiability (Trager et al., 2020; Kileel et al., 2019; Stock and Gribonval, 2023; Grigsby et al., 2025; Bona-Pellissier et al., 2022; Zhao et al., 2025; Marchetti et al., 2025). The rank of the parameter Jacobian is the local functional dimension (Grigsby et al., 2025); prior work has connected this viewpoint to generalization and optimization (Grigsby and Lindsey, 2024; Bona-Pellissier et al., 2024). We study the full Jacobian spectrum as a graded

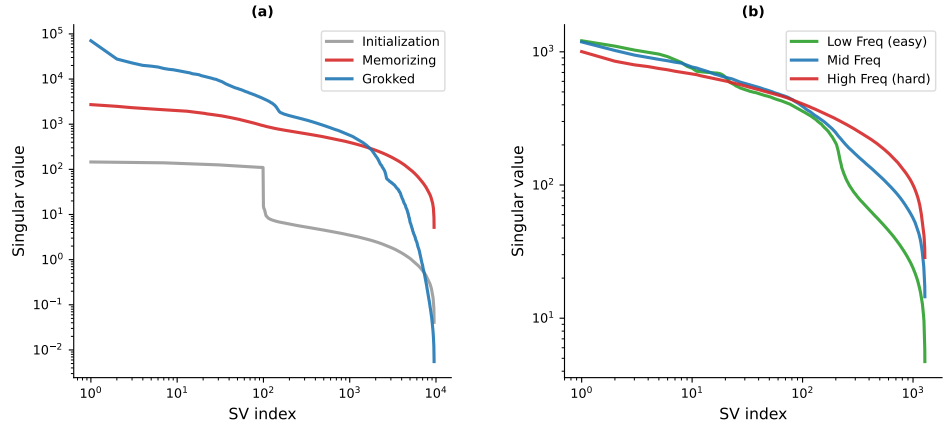

Figure 1: **The neural spectrum encodes feature learning.** Neural spectrum - Jacobian singular values (log-log). **(a)** SP Transformer (width 128) on mod-97 multiplication, 3 seeds. Gray: initialization; the cliff at O=99 is consistent across architectures (Fig. 27). Red: memorization. Blue: post-grokking *spectral polarization*. **(b)** DCT-MNIST converged spectra (MLP, 5 seeds). Harder tasks yield broader spectra.

refinement, using Rényi-rank measures that interpolate among numerical, effective (Roy and Vetterli, 2007), and stable rank (Rudelson and Vershynin, 2007).

Spectral evolution during optimization. Prior spectral analyses have focused primarily on weight matrices or the input-output Jacobian (Martin and Mahoney, 2018; Martin et al., 2021; Yunis et al., 2024; Pennington et al., 2017). We instead study the parameter Jacobian, whose squared singular values are the eigenvalues of the empirical NTK. Most closely, Lin (2026) analyzed empirical-NTK eigenspaces and spectral cliffs in toy models and modular addition. Theoretical work has also studied neural and NTK spectra in several regimes (Choromanska et al., 2015; Xie et al., 2017; Nguyen and Mondelli, 2020; Nguyen et al., 2021; Murray et al., 2023). We track the full spectrum continuously through training across architectures and parameterizations, and directly intervene on its concentration.

Delayed generalization. Grokking was introduced by Power et al. (2022) and has since been studied through Fourier-circuit formation, modular-polynomial structure, internal circuits, scaling laws, and effective theories (Nanda et al., 2023; Doshi et al., 2024; Furuta et al., 2024; Tian, 2025; Liu et al., 2022; 2023). Mallinar et al. (2025) further connected delayed generalization to feature learning beyond neural networks. These works identify solution structure or feature learning mechanisms, but do not provide a continuous spectral diagnostic that tracks the transition during training across architectures.

Feature learning and the NTK. NTK and μP theory distinguish lazy from feature-learning regimes (Jacot et al., 2020; Chizat et al., 2020; Yang and Hu, 2022; Yang et al., 2022). Our parameterization experiments test this axis directly, complementing prior empirical and theoretical evidence connecting empirical-NTK effective rank to generalization (Baratin et al., 2021).

3 Methodology

We find that the neural spectrum restructures at the memorization-to-generalization transition. This section defines the objects and metrics we use to characterize this restructuring.

Neural Spectrum. For a model $f_\theta : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^O$ with parameters $\theta \in \mathbb{R}^d$, evaluated on a fixed batch of N inputs $Z = \{z_i\}_{i=1}^N$, the parameter Jacobian is $J_Z(\theta) \in \mathbb{R}^{NO \times d}$ with entries $\partial f_\theta(z_i)/\partial \theta_j$. The neural spectrum is the set $\{\sigma_i\}_{i=1}^{\min\{NO,d\}}$ of singular values of this matrix which by convention are arranged in decreasing order. Their squares σ_i^2 are the eigenvalues

of the empirical NTK $K = JJ^\top$ (Jacot et al., 2020). We relate the shape of the neural spectrum to learning dynamics and generalization, cf. Figure 1.

Spectral polarization. We define spectral polarization as the concentration of a model's output sensitivity onto a small number of parameter-space directions: a small number of singular values grow large while the rest collapse. We measure it by the polarization index $\mathcal{P} = \log_{10}(\sigma_1/\sigma_k)$ for $k = 50$, which is the log ratio between the largest singular value and a value in the tail. Results are qualitatively unchanged for $k \in \{20, 50, 100, 1000\}$ (Figure 8).

Spectral ranks. The rank of J_Z is precisely the batch functional dimension of (Grigsby et al., 2025) on the probe set Z – the dimension of the manifold of functions reachable by local parameter perturbations. Spectral polarization therefore has a direct geometric interpretation: grokking corresponds to a collapse of functional dimension, as the model converges to a solution where all nearby parameters either approximately preserve the function or move it within a low-dimensional subspace. In practice, numerical rank cannot be used directly: it is sensitive to the tolerance threshold and can saturate at $\min(NO, d)$.

Let $\lambda_i = \sigma_i^2 / \sum_j \sigma_j^2$ denote the normalized squared singular values. The Rényi entropy of order α, $H_\alpha(\lambda) = \frac{1}{1-\alpha} \log \sum_i \lambda_i^\alpha$, yields a family of rank-like quantities $\exp(H_\alpha)$. At $\alpha = 0$ this is the numerical rank (count of nonzero singular values); at $\alpha = 1$ it gives the effective rank (Roy and Vetterli, 2007) (entropy-weighted mode count); at $\alpha = \infty$ it gives the stable rank ($\|J\|_F^2/\sigma_1^2$). In the overparameterized regime ($d \gg NO$), numerical rank is sensitive to tolerance and can both saturate and undercount due to its discrete nature, while stable rank is dominated by σ_1. We use $\alpha = 0.05$: the function $x^{0.1}$ approximates the indicator function that is 0 at 0 and 1 elsewhere, so $\exp(H_{0.05})$ gives a soft mode count that avoids the threshold sensitivity of numerical rank while remaining sensitive to modes far below σ_1. This quantity approximates functional dimension and the Rényi formulation applies to any architecture without requiring a hard threshold.

Spectral decay shape. We quantify the spectral decay shape using two complementary adaptations of classical spectral variation measures (see Appendix C for full details).

The first, which we call *(truncated) log-spectral flatness*, adapts the spectral flatness measure from signal processing (e.g. Johnston (2002), Dubnov (2004)) to the singular value setting: Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ denote the ordered list of singular values of the parameter Jacobian J. Fix a truncation threshold $p \in (0, 1)$ (we use $p = 0.95$) and set $k = \lfloor p \cdot r \rfloor$. We define the *truncated log-spectral flatness* as

$$\mathrm{F}_p := \frac{\left(\prod_{i=1}^{k-1} \log(\sigma_i/\sigma_k)\right)^{1/(k-1)}}{\frac{1}{k-1} \sum_{i=1}^{k-1} \log(\sigma_i/\sigma_k)}. \tag{1}$$

$\mathrm{F}_p \in (0, 1]$ by the arithmetic–geometric mean inequality, with equality if and only if $\log(\sigma_i/\sigma_k)$ is constant for $i < k$. Smaller values indicate greater nonuniformity in the log-spectrum relative to σ_k, and in particular are sensitive to the emergence of sharp spectral cliffs.

The second, which we introduce as the *Dirichlet energy* of the log-spectrum,

$$\mathrm{TV}_2(\log \sigma) := \frac{1}{\log r} \sum_{i=1}^{r-1} \frac{(\log \sigma_{i+1} - \log \sigma_i)^2}{\log(i{+}1) - \log i}, \tag{2}$$

adapts the classical Dirichlet energy to $\log \sigma_i$ viewed as a function of $\log i$, normalized by $\log r$. Note that the simpler alternative, standard total variation $\sum |\Delta \log \sigma_i|$, telescopes to the log condition number $\log(\sigma_1/\sigma_r)$ for any sorted (monotone) sequence and is therefore insensitive to whether the decay between the extremes is gradual or cliff-like. These two measures are complementary: F_p is sensitive to large-scale cliff structure in the spectrum, while TV_2 captures finer-grained variation in the decay shape.

4 Experiments

4.1 Spectral restructuring during grokking

We first ask how the neural spectrum evolves during grokking. Across four different architectures—Transformer, MLP, LSTM, and CNN—trained on the same modular multiplication task, the memorization-to-generalization transition is accompanied by a pronounced restructuring of the neural spectrum.

Setup. We train four models on modular multiplication $a \cdot b \bmod 97$, encoded as 5-token sequences with vocabulary size 99. The dataset contains 9,312 input-output pairs, with a 50/50 train/test split. We define the *grokking step* as the first train step where test accuracy exceeds 95%. All architectures are matched at approx. 420k parameters: a Transformer (2 layers, 4 heads, dimension 128), MLP (3 hidden layers, dimension 89), LSTM (3 cells, dimension 128), and CNN (3 1D convolutional layers, kernel size 3, dimension 61). All use GELU activations. Full architecture and training details are given in Tables 2 and 3.

Results. Figures 2 and 3 show a common pattern across architectures. After the models have fit the training data but before they generalize, the neural spectrum remains relatively close to its initialization shape. Near the grokking transition, the spectrum becomes increasingly concentrated in its leading singular directions. As validation accuracy rises, the leading singular values grow by one to two orders of magnitude while the tail collapses, producing spectral polarization. Correspondingly, the polarization index increases and the effective rank decreases around the same time that validation accuracy rises.

The architectures differ, however, in both the timing and form of this restructuring. The Transformer groks fastest and exhibits the strongest polarization, while the CNN groks slowest and retains the most diffuse spectrum; the MLP and LSTM lie in-between. The detailed spectral shape also differs: for the Transformer, MLP, and CNN, grokking sharpens the spectral cliff and decreases log-spectral flatness $\mathrm{F}_{0.95}$, whereas for the LSTM the cliff disappears and $\mathrm{F}_{0.95}$ increases. Thus $\mathrm{F}_{0.95}$ captures architecture-dependent changes in spectral shape even though the common signature across architectures is increasing polarization and decreasing effective rank. In this high-weight-decay regime, $\mathrm{TV}_2(\log \sigma)$ tracks the transition less cleanly, as the dominant change is the large-scale movement of the spectral cliff.

4.2 The feature-learning regime controls spectral polarization

We next ask how the feature-learning regime affects spectral polarization during grokking. We track the parameter-Jacobian spectrum throughout training for the same Transformer under three parameterizations—standard parameterization (SP), NTK parameterization, and maximal update parameterization (μP)—which place the model in different feature-learning regimes. We then interpolate continuously between these parameterizations to test whether the observed differences extend across a continuum of feature-learning regimes.

Setup: three parameterizations We train Transformers under three parameterizations to investigate whether feature learning dynamics drive grokking (Kumar et al., 2024). *Standard parameterization* (SP) uses default PyTorch scaling (Glorot and Bengio, 2010; He et al., 2015); at finite width, the model begins in an approximately lazy regime and transitions to rich feature learning during training. *NTK parameterization* scales updates such that the empirical NTK evolves at a rate of $\Theta(1/n)$, suppressing feature learning as width $n \to \infty$ (Jacot et al., 2020). *Maximal update parameterization* (μP) ensures layer pre-activations receive $\Theta(1)$ updates regardless of width (Yang et al., 2022), placing the model in the rich regime from initialization. Both NTK and μP maintain $\Theta(1)$ output changes per update across widths, enabling learning rate transfer, whereas SP's output changes are width-dependent. We sweep over all three parameterizations at widths 64, 128, and 256. The width-128 experiments use 3 seeds; other widths use a single seed. We implemented μP from Yang et al. (2022), with additional help from the implementation in Dey et al. (2024), and NTK parameterization for transformers as described in Everett et al. (2024). These experiments use weight decay $\lambda = 0.01$, an order of magnitude lower than the

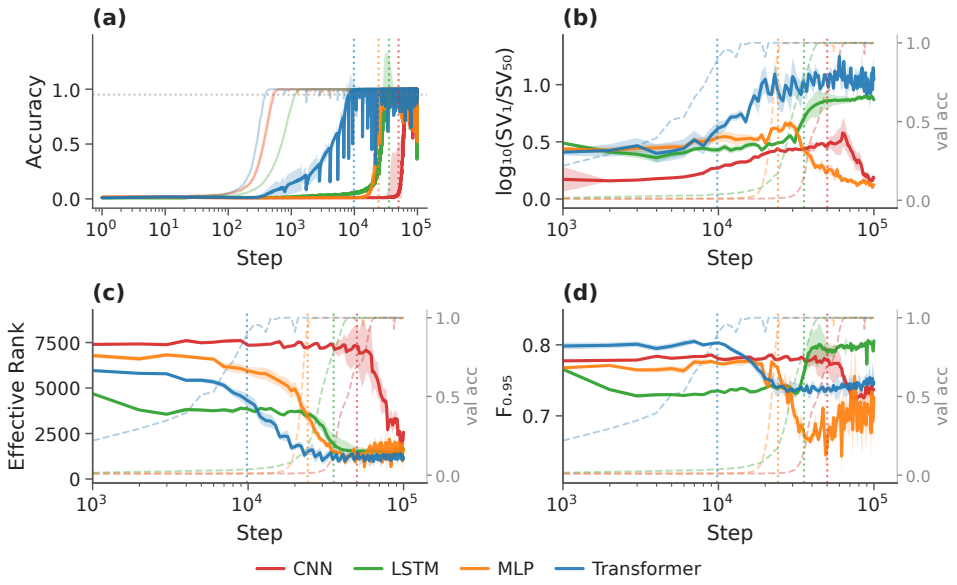

Figure 2: **Spectral metrics track grokking across four architectures.** Dotted vertical lines mark each architecture's grokking step (val acc $> 95\%$). **(a)** Train (faint) and validation (bold) accuracy. **(b)** Polarization index $\log_{10}(\sigma_1/\sigma_{50})$. **(c)** Effective rank. **(d)** Log-spectral flatness $F_{0.95}$ (Eq. 1). Panels (b)–(d) overlay validation accuracy (dashed) for temporal reference. Seed 42 line; shaded ± 1 std from 3 seeds.

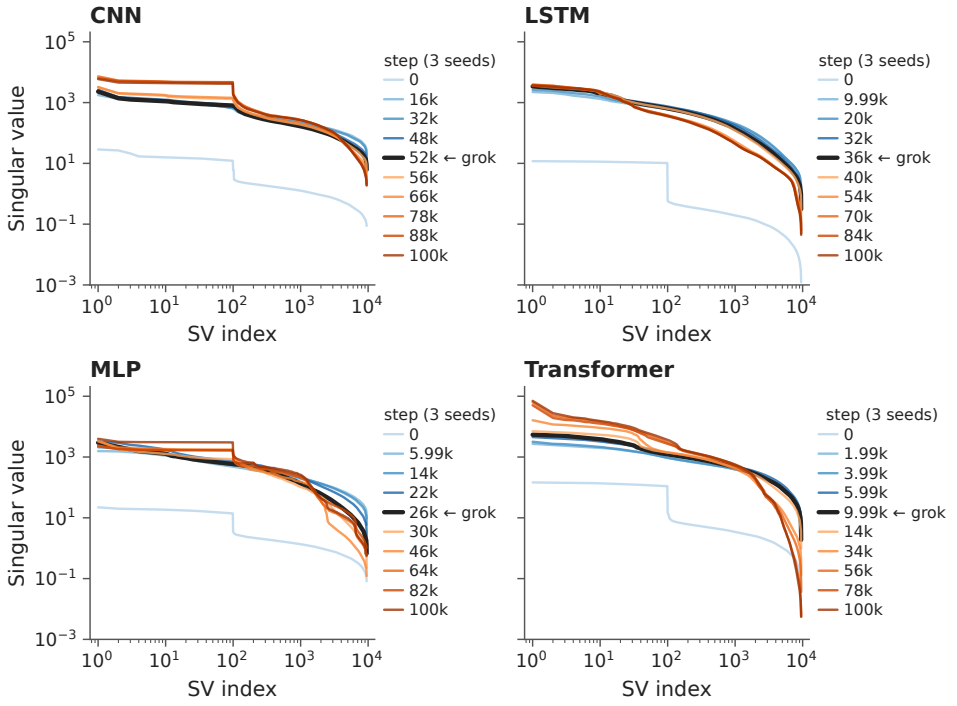

Figure 3: **SV spectra evolve from diffuse to polarized.** Each panel shows one architecture's sorted SV spectrum (log-log) at ten snapshots, colored from early (blue) to grokking (black) to late (orange). Early spectra decay gradually; after grokking, a few large SVs separate from the rest by a steep drop. Averaged over 3 seeds.

multi-architecture experiments ($\lambda = 0.1$), because μP with higher weight decay groks too rapidly to observe the spectral dynamics of interest.

Results. Figure 4 shows the width-128 results; similar patterns hold across widths (Figures 6 and 7). The three parameterizations exhibit markedly different relationships between spectral polarization and generalization. Under SP, polarization and grokking occur together: effective rank drops sharply at approximately the same time that validation accuracy rises. Under μP, generalization occurs earlier and precedes full polarization, while under NTK the spectrum remains close to initialization and the model does not grok within the training horizon. Thus polarization is not a prerequisite for generalization. Rather, its timing depends on the feature-learning regime: μP begins in a rich regime and generalizes before substantial spectral compression, whereas under SP the onset of polarization coincides with the transition out of the lazy regime.

Because these experiments use weaker weight decay ($\lambda = 0.01$) than the multi-architecture experiments ($\lambda = 0.1$) of §4.1, the form of spectral restructuring is also different. Under stronger weight decay, grokking is accompanied by a large-scale sharpening and movement of the spectral cliff, which is captured clearly by log-spectral flatness $F_{0.95}$. Here, the cliff remains comparatively stable while finer-scale variation develops across the spectrum, so $TV_2(\log \sigma)$ tracks the transition more clearly (see Figure 37 and panel (d) of Figure 4). Thus the two metrics capture complementary aspects of spectral restructuring.

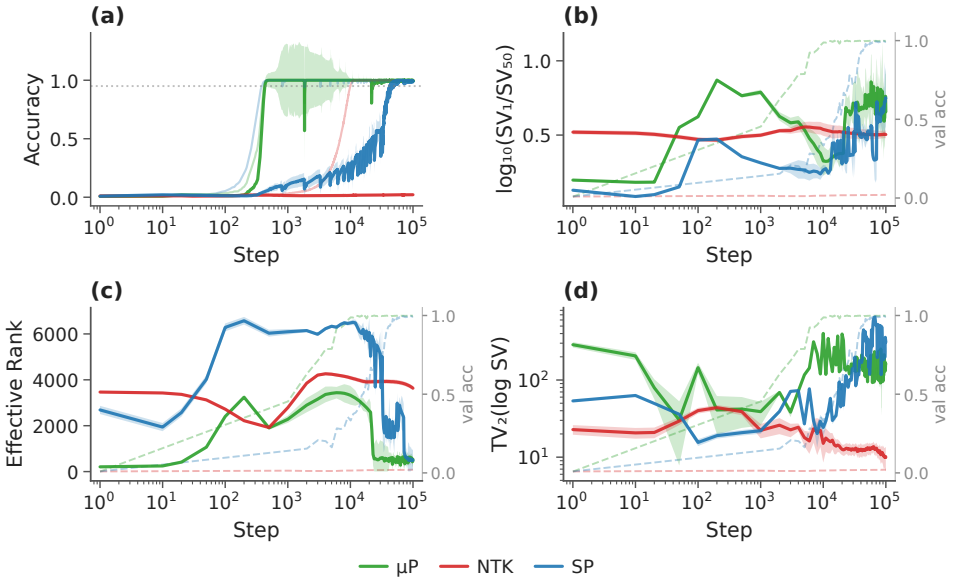

Figure 4: **Spectral dynamics under three parameterizations** (Transformer, width 128; seed 42 line, ±1 std from 3 seeds). **(a)** Validation accuracy. **(b)** Polarization (k=50). **(c)** Effective rank. **(d)** $TV_2(\log \sigma)$. μP (green) groks fastest and shows early spectral restructuring. SP (blue) groks after a delay, with restructuring coinciding with the generalization transition. NTK (red) never groks; its spectrum remains near its initialization throughout training.

Further Experiments: Finer interpolation across feature learning regimes To test whether the differences among SP, NTK, and μP are specific to these three parameterizations, we evaluate seven parameterizations along a continuous interpolation between NTK to μP, with SP at the midpoint. For more details on setup and results, see the Appendix C.3. Table 10 shows that none of the three tested points between NTK and SP grok within the training budget, while all seeds grok at SP and at each tested point between SP and μP. Polarization increases in all of the regimes, but its timing and accompanying rank dynamics differ across the interpolation. At SP, polarization is sharply concentrated around the grokking transition. Moving toward μP, the model starts low rank and thus can polarize later after grokking, while effective rank no longer undergoes the pronounced collapse seen at SP and can instead increase over training. Thus the interpolation supports the view that spectral polarization accompanies grokking across the feature-learning regimes tested, while the timing and form of the restructuring depend on the regime.

4.3 The neural spectrum is resilient to parameter symmetry

To separate changes in the parameter representative from changes in the function, we apply exact attention change-of-basis and positive ReLU-rescaling symmetries to six grokked checkpoints across seeds and feature learning parameterizations. At condition number $\kappa = 10^4$, weight Frobenius and nuclear norms change by 3.59–20.1× and 9.67–15.1×, respectively, while predictions are preserved. For a smooth invertible symmetry g, with $T = Dg_\theta$, $J_Z(g(\theta)) = J_Z(\theta)T^{-1}$: Jacobian rank and its output-space image are invariant, but its singular values need not be. Empirically, Jacobian effective rank changes by $-11.8\% \pm 7.6\%$, versus a $-33.2\% \pm 1.5\%$ change in weight effective rank. Thus weight summaries can shift substantially without a functional change; the Jacobian spectrum is less affected, though it is not completely symmetry-invariant. More details can be found in Appendix C.3.

4.4 Manipulating spectral polarization accelerates or prevents grokking

The preceding experiments show that spectral polarization accompanies grokking and that its timing depends on the feature-learning regime, but they remain observational. We next ask whether directly encouraging or suppressing spectral polarization changes the grokking transition. We find that in the SP setting, encouraging polarization accelerates grokking, while encouraging a diffuse spectrum prevents grokking within the training budget.

Setup. We augment the training objective with a differentiable approximation to the fraction of parameter-Jacobian spectral mass lying outside its leading eight directions. Minimizing this quantity encourages spectral concentration, while reversing its sign encourages a more diffuse spectrum. The approximation is computed from random projections of the Jacobian during training; all reported spectra are still measured by exact SVD on the fixed probe set. We additionally train with a gradient-norm penalty as a nonspectral control for generic gradient pressure. Full implementation details and controls are given in Appendix C.3

Results. Encouraging spectral concentration causes polarization to occur earlier and accelerates the 95% validation-accuracy transition by 5.2×, while reversing the intervention preserves training-set memorization but prevents grokking in all three seeds within the training budget (Table 11). A gradient-norm penalty also accelerates grokking, but less strongly, and its measured polarization occurs only after the generalization transition, showing that polarization is not the only route to generalization and that generic gradient pressure does not reproduce the behavior of the polarizing intervention. The effect also persists without weight decay: the polarizing intervention accelerates the 95% validation transition by 6.6× across two seeds (Table 13). Thus, in the SP setting, directly advancing or suppressing spectral polarization changes the grokking transition in the predicted direction.

4.5 Spectral rank tracks task complexity

We next ask whether the neural spectrum reflects functional complexity beyond the grokking setting. Across four families of supervised learning tasks with systematically varied difficulty, we compare the converged neural spectra of models trained on easier and harder targets. Across these benchmarks, harder learnable tasks generally produce broader spectra with higher effective and Rényi rank, suggesting that features at similar scales increase for harder tasks as reflected in spectral rank that is modality, task and architecture agnostic.

Setup. We evaluate whether spectral rank tracks task complexity across four supervised-learning benchmarks in which difficulty is varied in a controlled way. In Hermite polynomial regression, complexity is increased by raising the polynomial degree; in DCT-MNIST, by restricting the input to progressively higher-frequency components; in noisy MNIST, by increasing the magnitude of input perturbations; and in QM9 molecular property prediction, by moving from more local to more globally dependent target properties. Figure 18 summarizes these four benchmark constructions. For each, we train the same architecture across difficulty levels and compare the converged neural spectra using effective rank and Rényi rank ($\alpha = 0.05$). Full task constructions, architectures, and training details are given in Appendix A.6.

Results. Across the four benchmarks, harder learnable tasks generally produce broader neural spectra and higher spectral rank at convergence (Figure 5 and table 1). This pattern persists across substantially different data modalities and across both regression and classification tasks. Rényi rank is consistently higher than effective rank, reflecting its greater sensitivity to small but nonzero singular values. The main exception occurs beyond the learnable frontier: in the Hermite benchmark, rank increases through degree 20 but decreases slightly at degree 60, where the model no longer fits the target successfully. Thus spectral rank tracks increasing task complexity while the model continues to recruit additional functional directions, but need not increase once the task exceeds the model's effective capacity.

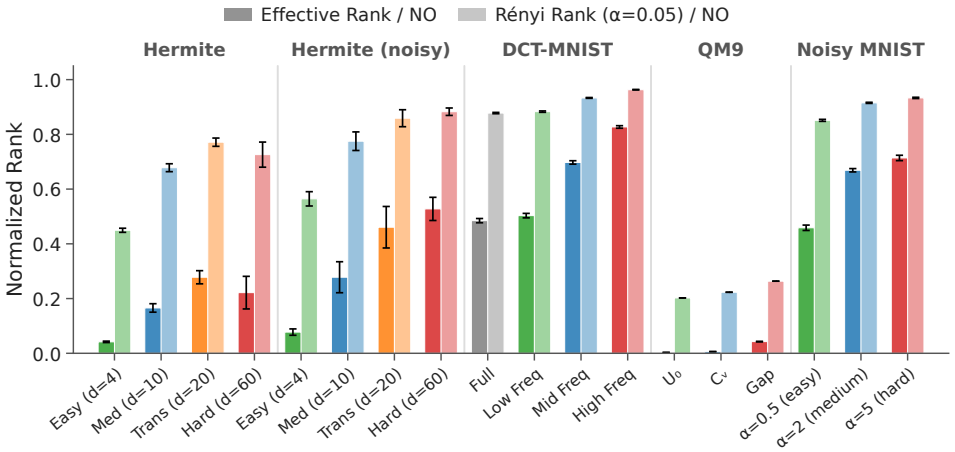

Figure 5: **Normalized spectral rank at convergence across all task complexity benchmarks.** Each group of bars shows one benchmark (Hermite, Hermite with label noise, DCT-MNIST, QM9, Noisy MNIST), with tasks ordered by increasing difficulty within each group. Solid bars: effective rank / NO; lighter bars: Rényi rank ($\alpha = 0.05$) / NO, where N is the number of samples and O is the Output Dimension. Within every benchmark, both rank metrics increase with task difficulty up to learnability, confirming our hypothesis. Colors encode difficulty (green = easy, red = hard) across benchmarks. Error bars: ± 1 std over 3–5 seeds.

5 Conclusion

We study the parameter Jacobian's singular-value spectrum as a window into the local function-space geometry of learning. Across four architectures, grokking coincides with spectral polarization, concentrating output sensitivity in fewer functional directions. Its timing depends on the feature-learning regime: polarization peaks near grokking under SP, restructuring begins earlier toward μP, and both spectral restructuring and grokking are suppressed toward NTK. Encouraging or suppressing polarization changes the SP transition, while harder learnable tasks consistently yield broader, higher-rank spectra, reflecting a larger number of active functional modes in the learned solution. On grokked checkpoints, parameter symmetries perturb spectral summaries less than weight summaries, supporting the neural spectrum as a useful account of learning dynamics.

Limitations and future work. The neural spectrum may be more resilient to parameter symmetry, but it is not truly invariant, which motivates further study of functional or spectral quantities. There is also a connection to mechanistic interpretability through the top singular vectors as functional modes, on which we show preliminary work for the known Fourier circuits of modular multiplication (Appendix A.14). Finally, exact SVD limits the present experiments to moderate-scale models, making scalable approximations to the neural spectrum an important direction for future work, although initial checkpoint-scale tests are promising (Appendix C.3).

AI use statement

In this work, we used generative AI tools for polishing text, producing figures, and writing code. We have not used generative AI tools to generate the core research idea, the method, or the theoretical formulations of this work, and the remaining disclosure categories are not applicable to this work. We have reviewed all AI-assisted work. All AI-assisted text was read and edited by the authors, figures were checked against the underlying data, and AI-written code was verified and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work studies neural network training dynamics using synthetic tasks and publicly available datasets (MNIST, QM9), none of which contain personally identifiable information. It involves no human subjects, and we do not foresee negative societal impacts beyond those common to foundational deep-learning research. All authors have read and adhere to the ICLR Code of Ethics.

Reproducibility statement

All spectral quantities are defined in Section 3 and Appendix C. Architectures, hyperparameters, seeds, and probe sets are listed in Tables 2 to 7, task constructions are described in Appendix A.6, and the additional experiments are detailed in Appendix C.3. All spectra are computed by exact SVD on fixed probe sets, and results are reported as mean $\pm$ one standard deviation across seeds. We plan to release the source code publicly upon acceptance. This will ensure that anyone will be able to reproduce our work.

References

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *ICML*, volume 32, pages 1908–1916. PMLR, 2014. 17

Aristide Baratin, Thomas George, César Laurent, R. Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2269–2277. PMLR, 2021. URL http://proceedings.mlr.press/v130/baratin21a.html. 3

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. 1

Joachim Bona-Pellissier, Francois Malgouyres, and Francois Bachoc. Local identifiability of deep ReLU neural networks: The theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b0ae046e198a5e43141519868a959c74-Abstract-Conference.html. 1, 2

Joachim Bona-Pellissier, Fran cois Malgouyres, and Fran cois Bachoc. Geometry-induced implicit regularization in deep ReLU neural networks. *Preprint arXiv:2402.08269*, 2024. 2

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning, 2020. URL https://arxiv.org/abs/1912.01198. 17

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020. URL https://arxiv.org/abs/1812.07956. 3

Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks, 2022. URL https://arxiv.org/abs/2202.13473. 17

Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, JMLR Workshop and Conference Proceedings. JMLR.org, 2015. URL http://proceedings.mlr.press/v38/choromanska15.html. 3

Nolan Dey, Quentin Anthony, and Joel Hestness. The practitioner's guide to the maximal update parameterization. https://www.cerebras.ai/blog/the-practitioners-guide-to-the-maximal-update-parameterization, September 2024. URL https://www.cerebras.ai/blog/the-practitioners-guide-to-the-maximal-update-parameterization. 5

Darshil Doshi, Tianyu He, Aritra Das, and Andrey Gromov. Grokking modular polynomials, 2024. URL https://arxiv.org/abs/2406.03495. 2, 3, 20

S. Dubnov. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, 11(8):698–701, 2004. doi: 10.1109/LSP.2004.831663. 4

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers, 2024. URL https://arxiv.org/abs/2407.05872. 5, 38

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials, 2024. URL https://arxiv.org/abs/2402.16726. 2, 3

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html. 5

J. Elisenda Grigsby and Kathryn Lindsey. On functional dimension and persistent pseudodimension, 2024. URL https://arxiv.org/abs/2410.17191. 2

J. Elisenda Grigsby, Kathryn Lindsey, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward ReLU neural networks. *Adv. Math.*, 482:Paper No. 110636, 62, 2025. ISSN 0001-8708,1090-2082. doi: 10.1016/j.aim.2025.110636. URL https://doi.org/10.1016/j.aim.2025.110636. 1, 2, 4

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. URL https://arxiv.org/abs/1502.01852. 5

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572. 3, 4, 5

James D Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on selected areas in communications*, 6(2):314–323, 2002. 4

Andrej Karpathy. nanogpt. https://github.com/karpathy/nanoGPT, 2022. 39

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://arxiv.org. 2

Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study, 2020. URL https://arxiv.org/abs/1910.08720. 1

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics, 2024. URL https://arxiv.org/abs/2310.06110. 5

Jennifer Lin. Feature identification via the empirical ntk, 2026. URL https://arxiv.org/abs/2510.00468. 3

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. 3

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data, 2023. URL https://arxiv.org/abs/2210.01117. 3

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product, 2025. URL https://arxiv.org/abs/2407.20199. 3

Giovanni Luca Marchetti, Vahid Shahverdi, Stefano Mereta, Matthew Trager, and Kathlén Kohn. An invitation to neuroalgebraic geometry. *CoRR*, abs/2501.18915, 2025. doi: 10.48550/ARXIV.2501.18915. URL https://doi.org/10.48550/arXiv.2501.18915. 1, 2

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning, 2018. URL https://arxiv.org/abs/1810.01075. 3

Charles H. Martin, Tongsu Peng, and Michael W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1), July 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24025-8. URL http://dx.doi.org/10.1038/s41467-021-24025-8. 3

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montúfar. Characterizing the spectrum of the NTK via a power series expansion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Tvms8xrZHyR. 3

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. 2, 3

Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8abfe8ac9ec214d68541fcb888c0b4c3-Abstract.html. 3

Quynh Nguyen, Marco Mondelli, and Guido F. Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8119–8129. PMLR, 2021. URL http://proceedings.mlr.press/v139/nguyen21g.html. 3

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice, 2017. URL https://arxiv.org/abs/1711.04735. 3

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177. 1, 3

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *EUSIPCO*, 2007. 3, 4

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):1–19, 2007. doi: 10.1145/1255443.1255447. 3

Pierre Stock and Rémi Gribonval. An embedding of ReLU networks and an analysis of their identifiability. *Constructive Approximation*, 57:853–899, 2023. 1, 2

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking, 2025. URL https://arxiv.org/abs/2509.21519. 3

Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and spurious critical points: a geometric study of linear networks. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. URL https://openreview.net. 2

Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1216–1224. PMLR, 2017. URL http://proceedings.mlr.press/v54/xie17a.html. 3

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2022. URL https://arxiv.org/abs/2011.14522. 3

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466. 3, 5

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights, 2024. URL https://arxiv.org/abs/2408.11804. 2, 3

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *CoRR*, abs/2506.13018, 2025. doi: 10.48550/ARXIV.2506.13018. URL https://doi.org/10.48550/arXiv.2506.13018. 1, 2

A Additional Figures

A.1 Feature Learning Regimes: Other Widths

A.2 Polarization Robustness

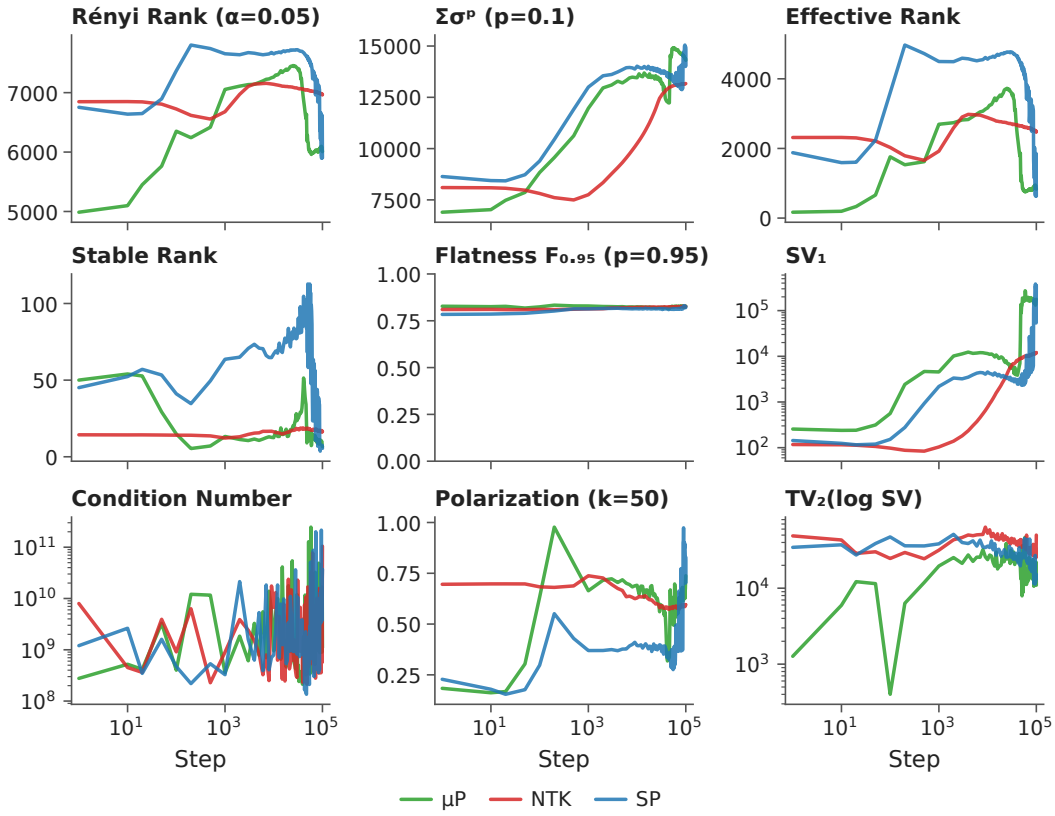

Figure 6: **Width 64:** μP groks first, NTK does not grok but does show some spectral movement at smaller width.

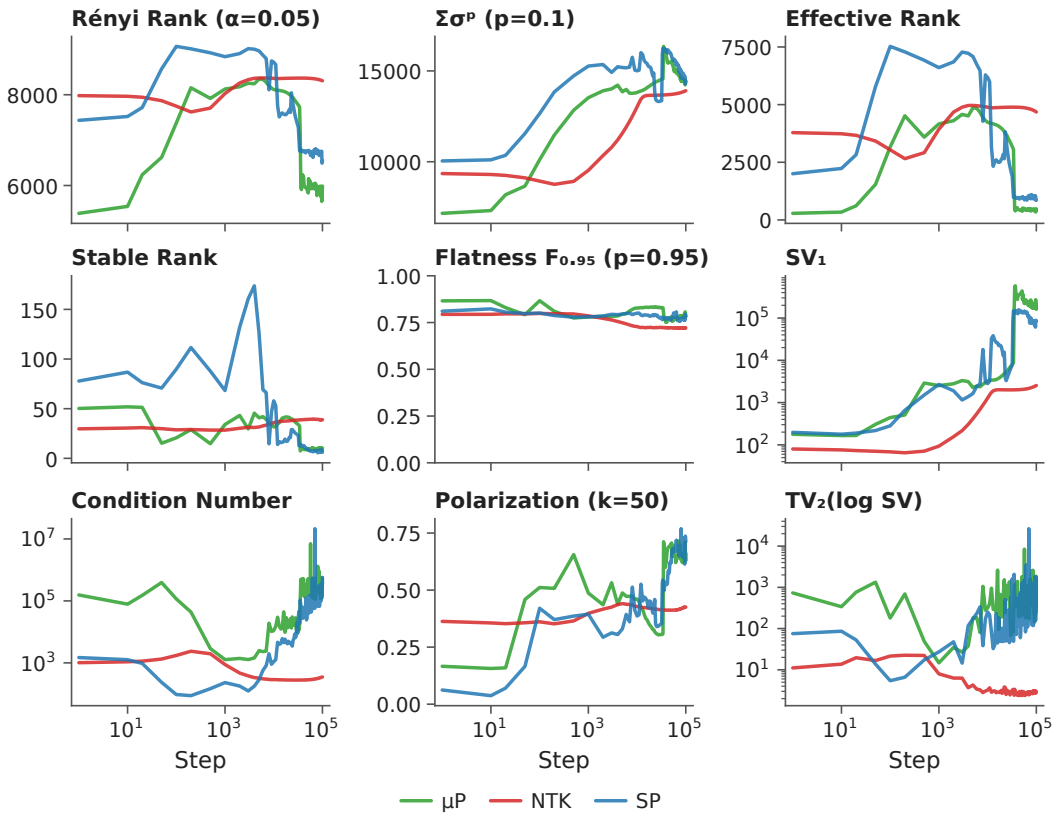

Figure 7: **Width 256:** NTK remains in a suppressed lazy regime at larger widths and does not grok while SP and μP each grok progressively faster in comparison.

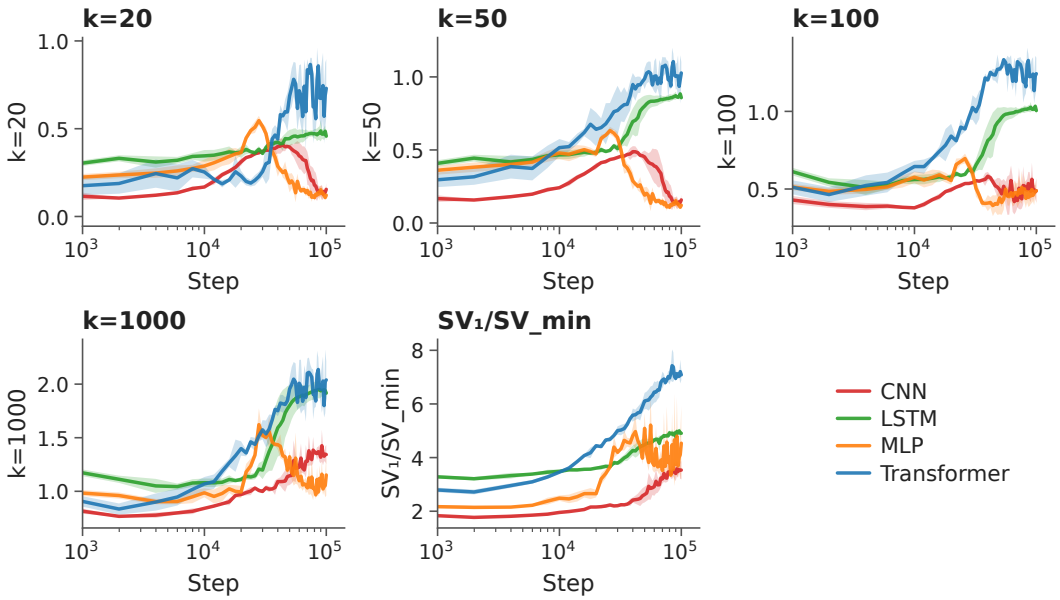

Figure 8: **Robustness of the polarization index.** $\log_{10}(\sigma_1/\sigma_k)$ for $k \in \{20, 50, 100, 1000\}$ across all four architectures. The qualitative pattern is unchanged.

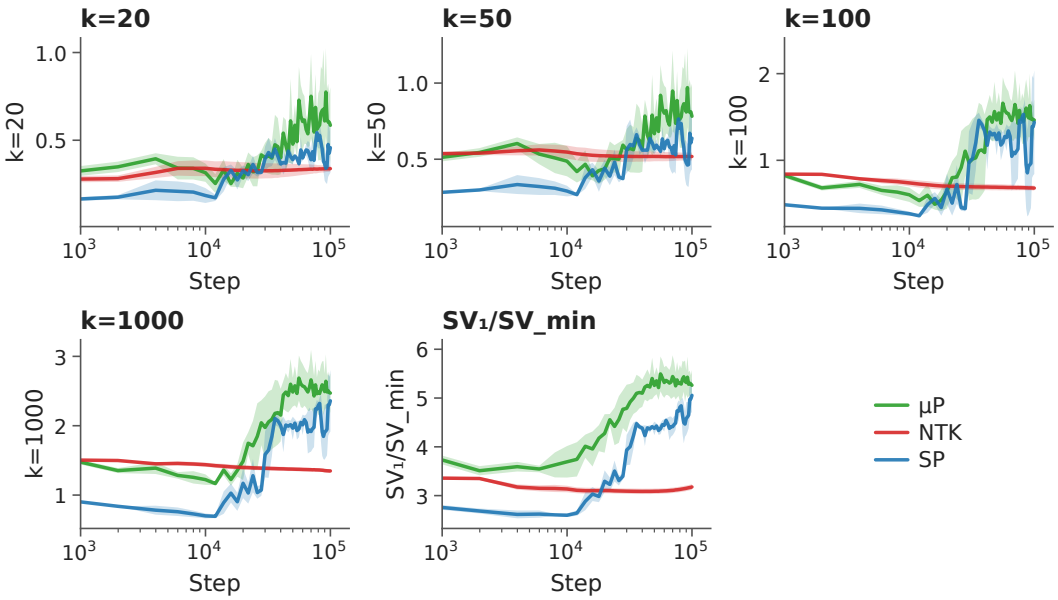

Figure 9: **Polarization robustness (feature parameterization study).** $\log_{10}(\mathrm{SV}_1/\mathrm{SV}_k)$ for $k \in \{20, 50, 100, 1000\}$ comparing SP, NTK, and μP at width 128. The qualitative pattern: μP polarizes most, NTK least, thus, polarization index is robust to the choice of k.

A.3 Complexity Table

Table 1: Effective rank and Rényi rank ($\alpha = 0.05$) measured at the final training epoch across benchmark tasks of varying complexity.

Task	**Effective Rank**	**Rényi Rank** ($\alpha = 0.05$)	**Final R^2 / Accuracy**
Hermite Polynomial Regression			
Easy ($d = 4$)	86 ± 5	921 ± 15	0.982
Medium ($d = 10$)	339 ± 32	1389 ± 30	0.157
Transition ($d = 20$)	569 ± 49	1580 ± 31	-0.035
Hard ($d = 60$)	454 ± 122	1487 ± 94	-0.182
DCT-MNIST Classification			
Full	620 ± 10	1124 ± 3	98.5%
Low Frequency	644 ± 10	1130 ± 3	98.6%
Mid Frequency	893 ± 8	1194 ± 2	87.8%
High Frequency	1059 ± 6	1233 ± 1	75.7%
QM9 Molecular Property Prediction			
U_0 (local)	2 ± 0	104 ± 0	0.9871
C_v (semi-local)	3 ± 0	114 ± 0	0.7330
Gap (global)	22 ± 1	135 ± 0	0.4241
Noisy MNIST Classification			
$\alpha = 0.5$ (easy)	4696 ± 101	8716 ± 37	94.6%
$\alpha = 2$ (medium)	6849 ± 62	9371 ± 21	87.7%
$\alpha = 5$ (hard)	7309 ± 101	9556 ± 24	66.5%

A.4 Full Spectral Metrics

A.5 SV Evolution

A.6 Task complexity benchmark details

Figure 18 summarizes the four benchmark constructions.

Hermite polynomial regression. We need a benchmark where task difficulty is analytically controlled. Probabilist's Hermite polynomials ($\mathrm{He}_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$) form an orthonormal basis for $L^2(\mathbb{R}^d, \gamma)$ under Gaussian inputs, so a target $f(x) = \sum_j c_j H_{\alpha_j}(x)$ of exact degree m has precisely controlled frequency content. Higher degree means more oscillations per unit interval, connecting to the spectral bias of neural networks and polynomial approximation theory (Cao et al., 2020; Choraria et al., 2022; Andoni et al., 2014). We select four difficulty levels (degree 4, 10, 20, 60, where coefficient distributions chosen are Sparse, Geom (exponentially decaying but not sparse), Geom, and Uniform, from a systematic ablation (Figure 26), spanning fully learnable to beyond the network's capacity. A 4-layer MLP (dim 512, ∼794k params) is trained for 500 epochs across 5 seeds. Effective rank and Rényi rank ($\alpha = 0.05$) at convergence increase monotonically with task difficulty up to the learnable frontier (Figure 5).

DCT-MNIST. To test on real images with controlled difficulty, we use the Discrete Cosine Transform to decompose MNIST digits into frequency bands. Low-frequency DCT coefficients encode coarse shape; high-frequency coefficients encode edges and fine texture. We retain only specific bands, producing four tasks of increasing difficulty: Full (all coefficients), Low (bottom 25%), Mid (middle 50%), and High (top 25%). An MLP (∼728k params) is trained across 5 seeds. The high-frequency task produces the highest effective rank (∼950 vs. ∼650 for full), consistent with harder tasks requiring more independent parameter directions (Figure 5).

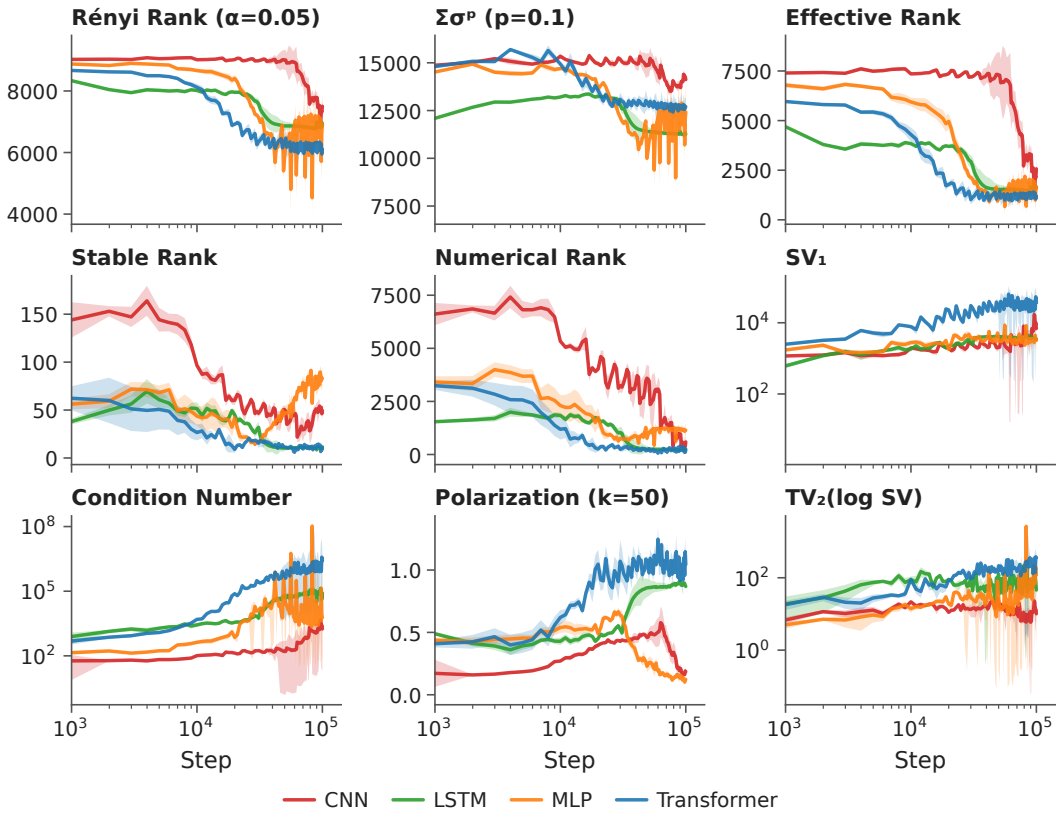

Figure 10: **Multi-architecture grokking: all spectral metrics** including Rényi rank ($\alpha = 0.05$), effective rank, stable rank, largest SV (σ_1), condition number, and polarization. Hybrid plotting: seed 42 fine line, ± 1 std from 3 seeds.

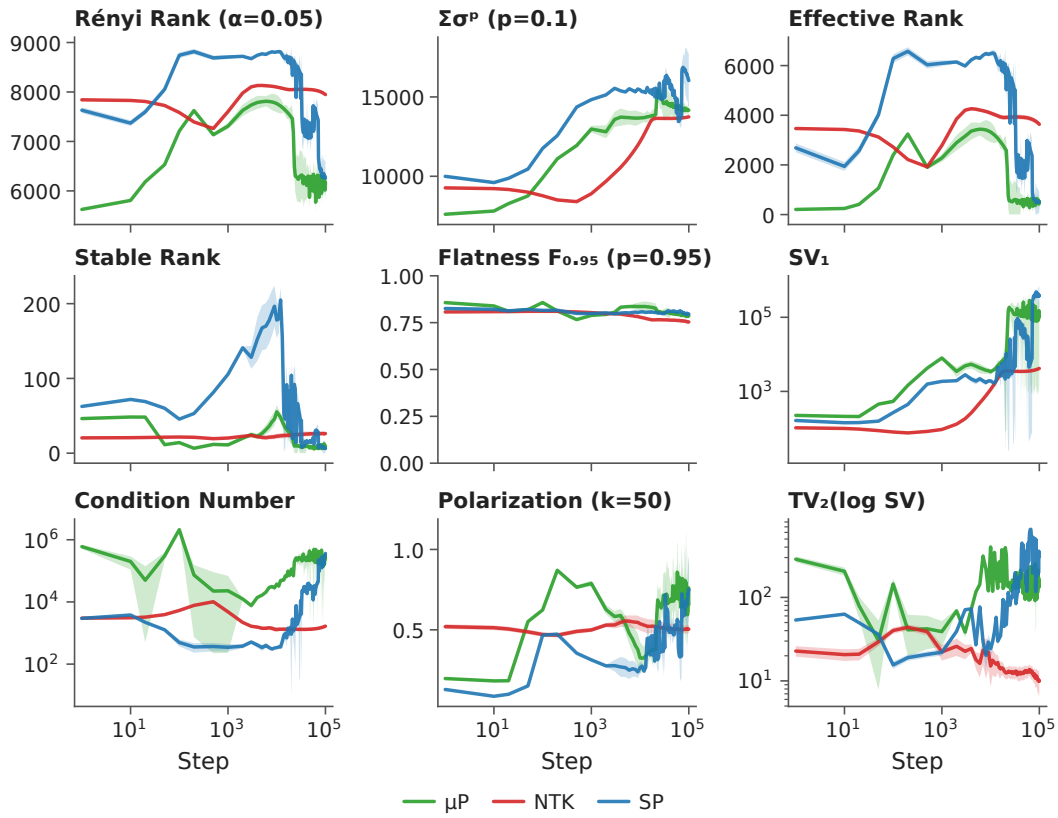

Figure 11: **Feature parameterization study (width 128): all spectral metrics.** Same layout as Multi-architecture experiments but comparing SP, NTK, and μP parameterizations. Uses hybrid plotting (seed 42 fine line, ± 1 std from 3 seeds).

Noisy MNIST Another common way to introduce complexity to MNIST classification is to add noise to the MNIST dataset. For each MNIST image x_i, we add noise by computing r_i, the distance to its nearest neighbor within the same minibatch, and add a perturbation drawn uniformly from the ball of radius $\alpha \cdot r_i$ around x_i (sampled via a normalized Gaussian rescaled by $U^{1/d}$, with $d = 784$). We generate three variants with $\alpha \in \{0.5, 2, 5\}$, which yield classification tasks of increasing difficulty as the radius of noise increases relative to the local neighborhood scale (Figure 25). We train an MLP ($\sim$26k params) for five different seeds and compare the effective rank and Rényi rank ($\alpha = 0.05$) for different task complexities. We see that as the difficulty of the task increases, the effective rank and the Rényi rank ($\alpha = 0.05$) also increase monotonically (Figure 5).

QM9 molecular property prediction. To test whether the neural spectra can capture task difficulty arising naturally from physics, we use QM9 ($\sim$130k small molecules). We select three targets of increasing non-locality and hence increasing complexity: atomization energy U_0 (local bond energies), heat capacity C_v (semi-local vibrational modes), and HOMO-LUMO gap (globally delocalized electronic structure). An MLP ($\sim$794k params) is trained across 3 seeds. Spectral rank at convergence increases monotonically with target difficulty (Figure 5).

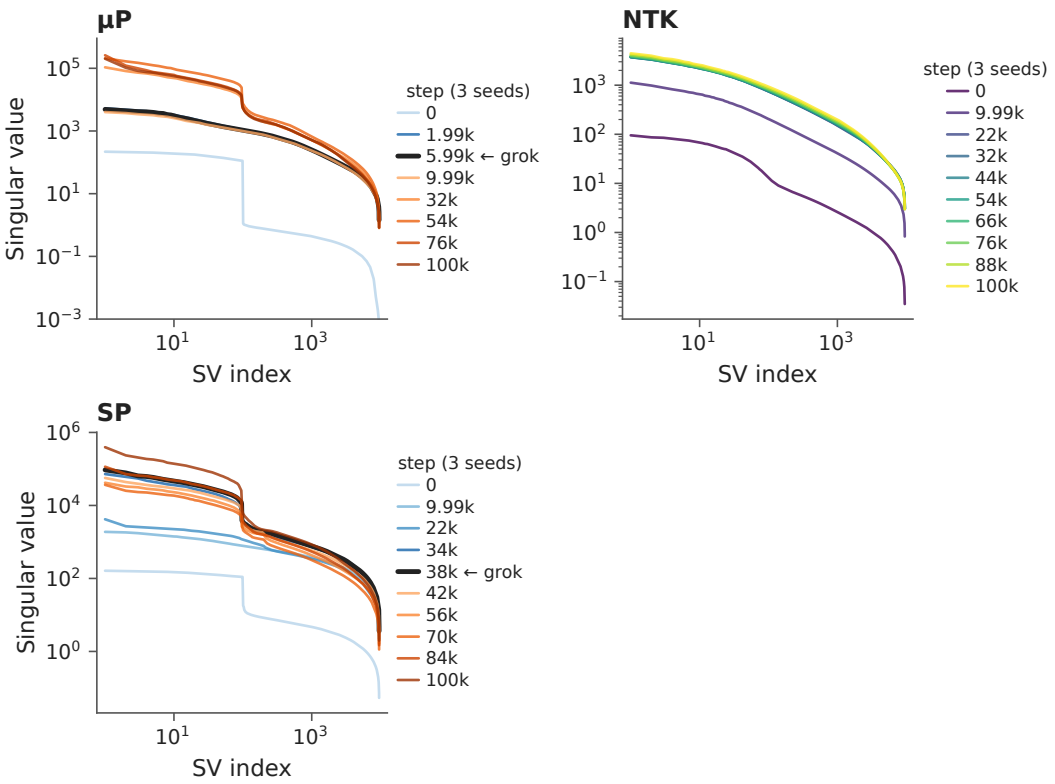

Figure 12: **SV spectra per parameterization at width 128.** SP and μP show spectral polarization; NTK remains near initialization.

A.7 Task Complexity Training Dynamics

A.8 QM9 (GNN)

A.9 Hermite with Noise

A.10 DCT-MNIST Task Construction

A.11 Noisy MNIST Task Construction

A.12 Hermite Task Selection

A.13 Spectral Cliff at Initialization

A.14 Fourier Alignment of Singular Vectors

We measure the alignment between the parameter Jacobian's left singular vectors and the discrete-logarithm Fourier basis vectors of the known multiplication circuit (Doshi et al., 2024). Figure 29–Figure 35 show per-frequency, per-singular-vector alignment heatmaps at several training stages. Models that grok develop structured alignment in the top singular vectors; NTK models do not. A full analysis of these results is deferred to future work.

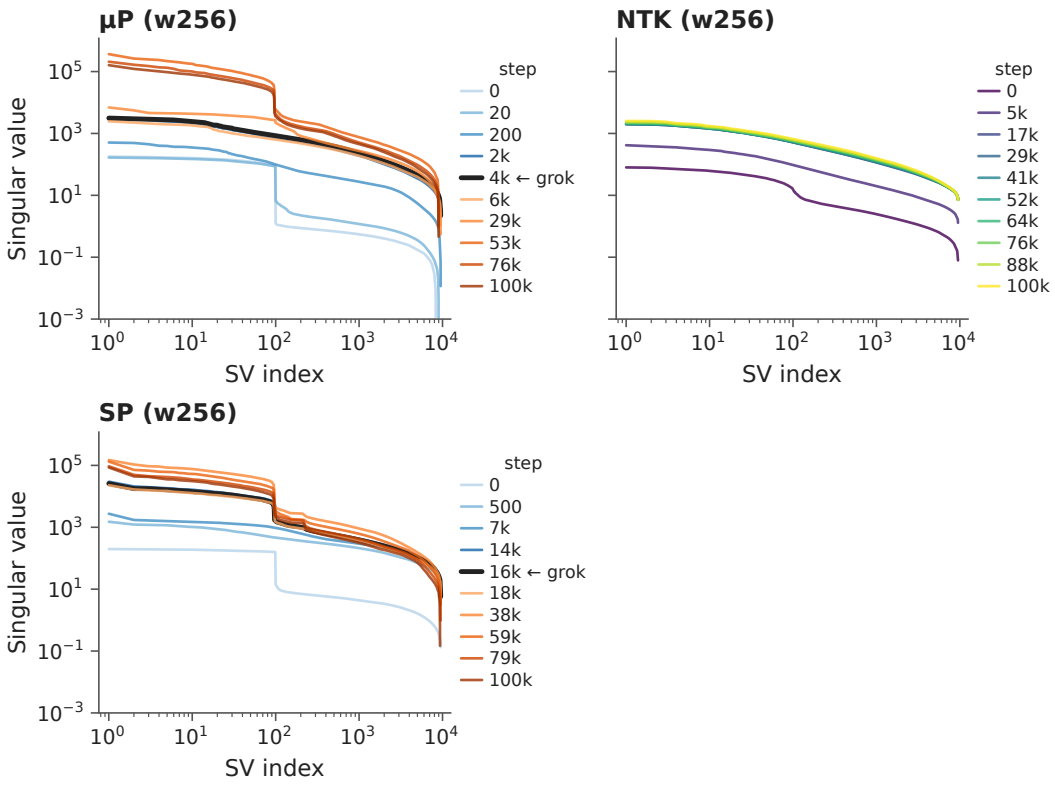

Figure 13: **SV evolution under three parameterizations (width 256).** Same layout as Figure 3. Blue: pre-grok snapshots. Black: grokking transition. Orange: post-grok. Single seed.

Table 2: **Grokking Architecture configurations.**

Architecture	Structure	Parameters	Details
Transformer	2 layers, 4 heads, dim 128	423k	Causal attention, GELU, LayerNorm
MLP	3 hidden layers, dim 89	425k	Input tokens concatenated, GELU
LSTM	3 cells, dim 128	420k	Manual implementation (jacrev-compatible)
CNN	3 causal 1D conv, kernel 3, dim 61	433k	Left-padded, GELU

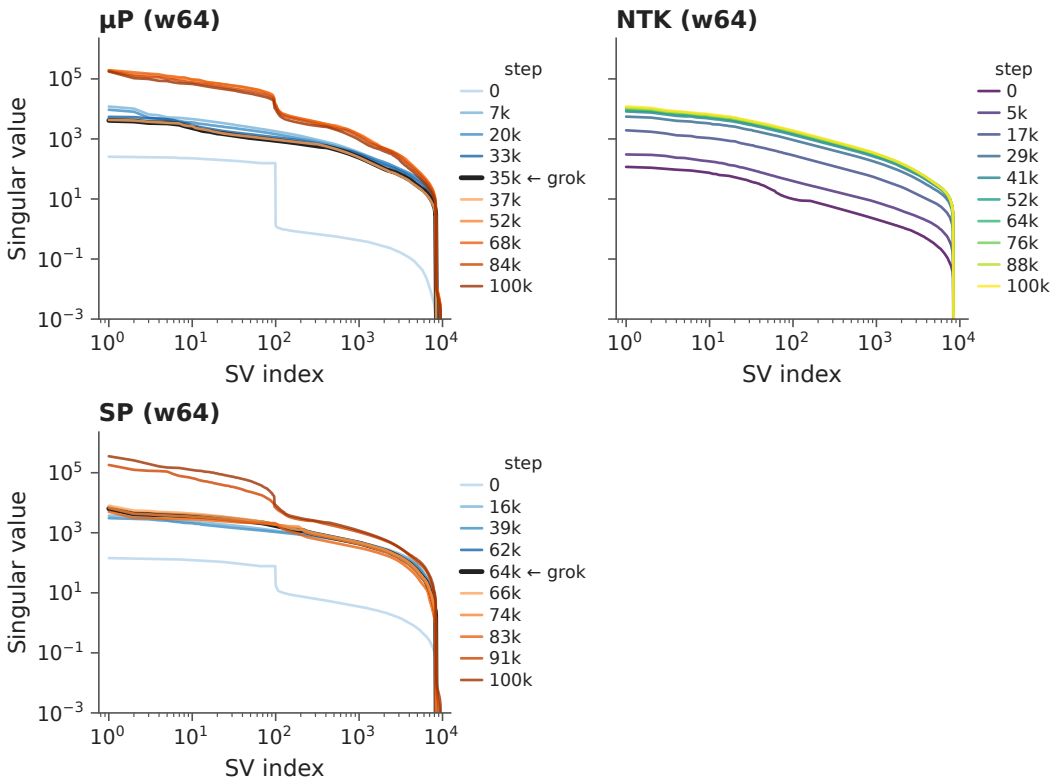

Figure 14: **SV evolution under three parameterizations (width 64).** Same layout as Figure 13. Single seed.

Table 3: **Grokking training configuration.**

Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	0.1 (multi-arch), 0.01 (feature parameterization)
Betas	$(0.9, 0.98)$
Batch size	512
Total steps	100,000
Total Epochs	10,000
Seeds	42, 67, 1234 (multi-arch and width-128 param); single seed (widths 64, 256)
Probe set	96 fixed training examples
Spectral method	Exact thin SVD (QR + SVD, float32)
Hardware	1 × A100

B Architecture and Training Details

B.1 Grokking

B.2 Task Complexity

B.2.1 Hermite polynomial regression configuration

B.2.2 DCT-MNIST configuration

B.2.3 QM9 molecular property prediction configuration

B.2.4 Noisy MNIST configuration

C Spectral variation measures: log-spectral flatness and Dirichlet energy

C.1 Log-spectral flatness

Spectral flatness (also known as Wiener entropy) is a classical measure from signal processing, defined for a sequence of positive values $x_1, \ldots, x_k$ as the ratio of the geometric mean to the

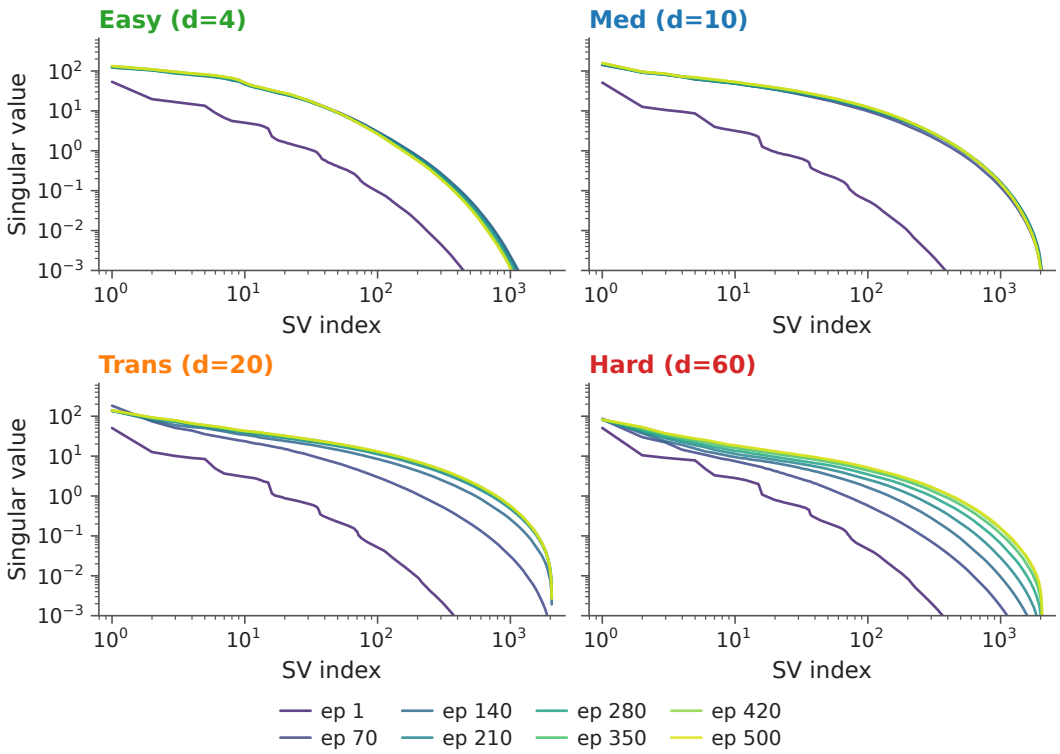

Figure 15: **Hermite SV spectra.** Easier tasks produce steeper spectra with few dominant modes; harder tasks produce broader spectra.

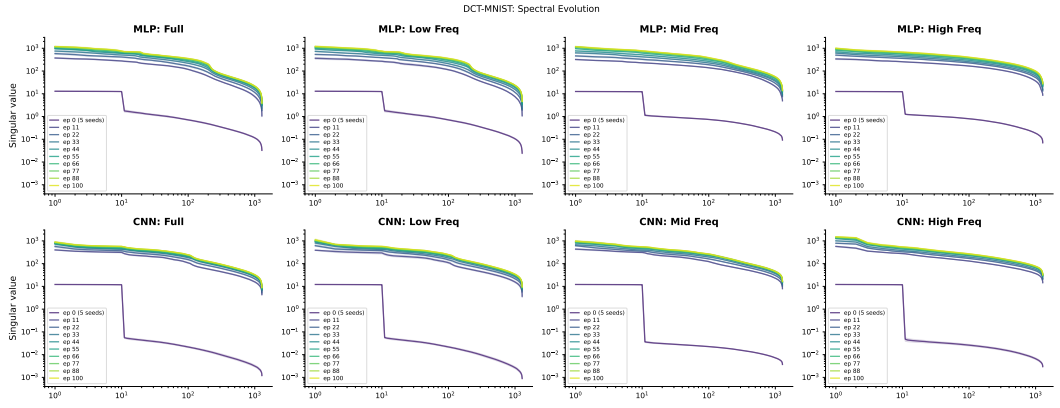

Figure 16: **DCT-MNIST SV spectra.** CNN spectra are smoother than MLP spectra across tasks.

arithmetic mean:

$$F(x_1, \ldots, x_k) = \frac{\mathrm{GM}(x_i)}{\mathrm{AM}(x_i)} = \frac{\left(\prod_{i=1}^{k} x_i\right)^{1/k}}{\frac{1}{k}\sum_{i=1}^{k} x_i}.$$

By the AM-GM inequality, $F \in (0, 1]$, with $F = 1$ if and only if all x_i are equal. Intuitively, F measures how uniformly the values are distributed: a flat, uniform distribution yields F close to 1, while a distribution concentrated on a few large values yields F close to 0, because the geometric mean is collapsed by any small values while the arithmetic mean is dominated by the large ones.

Applying spectral flatness directly to the singular values σ_i or their squares σ_i^2 is problematic in the neural spectrum setting, because singular values span many orders of magnitude (in

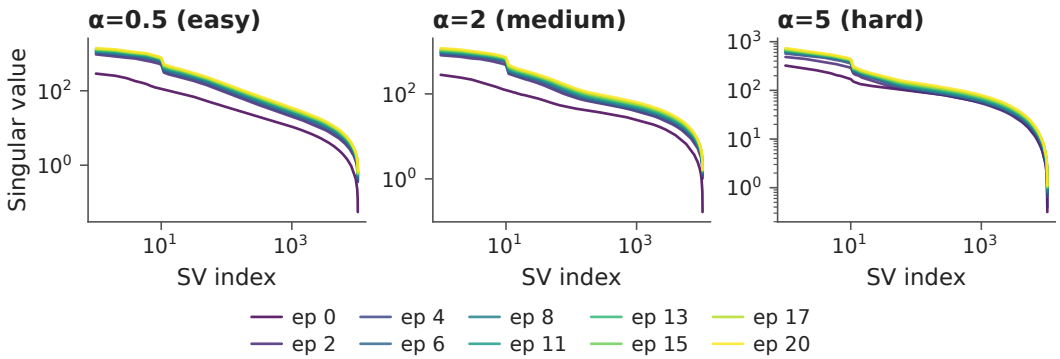

Figure 17: **Noisy MNIST: SV evolution across noise levels.** Each panel shows the sorted SV spectrum (log-log) at training snapshots for one noise level, averaged over seeds. Higher noise produces broader spectra with less polarization at convergence.

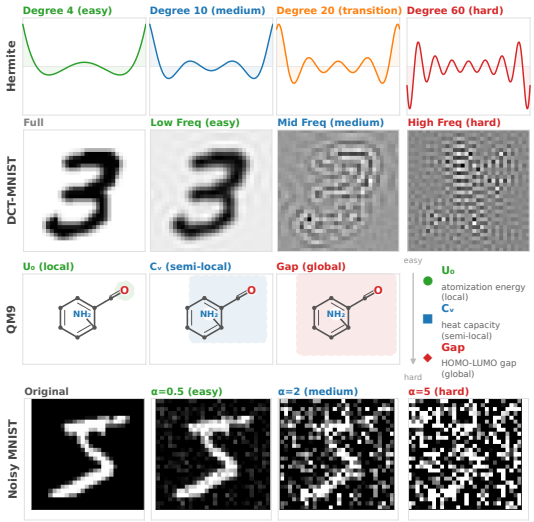

Figure 18: **Task complexity benchmarks.** Difficulty is varied through Hermite degree, DCT frequency band, MNIST noise level, and QM9 target non-locality.

our experiments, from 10^{-3} to 10^{5}). In this regime, the geometric mean is dominated by the smallest singular values, which are numerically unreliable, while the arithmetic mean is dominated by the largest. The ratio GM/AM therefore collapses to near zero throughout training and has little discriminating power.

Two further observations guide our definition. First, the neural spectrum span many orders of magnitude, and the logarithm is the natural way to handle this dynamic range ; it compresses the spectrum into a scale where structure at all levels is visible. The qualitative notion of roughness we wish to capture, whether the log-spectrum decays uniformly or contains sharp cliffs, is therefore a property of the log-transformed values, and it is natural to apply flatness to $\log \sigma_i$ rather than to σ_i directly. Second, $\log \sigma_i$ can be negative (when $\sigma_i < 1$), and geometric mean requires positive inputs. We address both issues through truncation and normalization, described next.

Definition 1. *Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ be the singular values of the parameter Jacobian J, or the neural spectrum. We define the* truncated log spectral flatness *as follows.*

Step 1: Truncation. *Fix a percentile $p \in (0, 1)$ (we use $p = 0.95$). Let $k = \lfloor p \cdot r \rfloor$. Retain only the top k singular values $\sigma_1 \geq \cdots \geq \sigma_k$, discarding the bottom $(1 - p)$ fraction of the spectrum.*

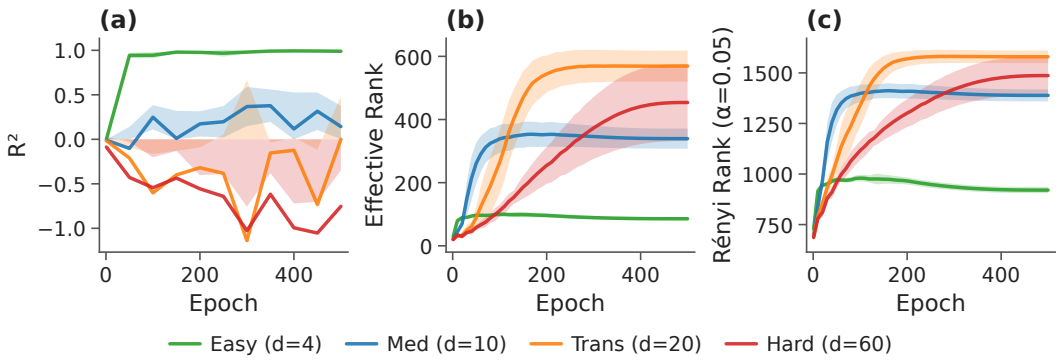

Figure 19: **Hermite training dynamics:** spectral rank over 500 epochs for four difficulty levels.

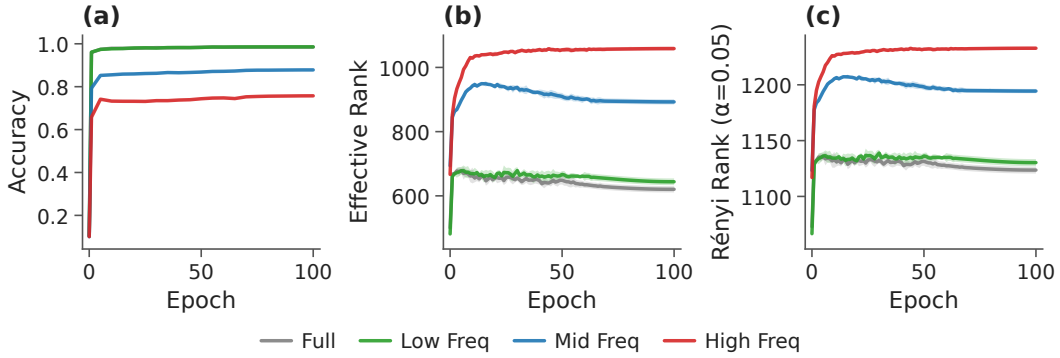

Figure 20: **DCT-MNIST training dynamics (MLP):** accuracy and effective rank over 100 epochs.

This removes the numerically unreliable tail while, for $p = 0.95$, reliably capturing the experimentally observed cliff that separates large from small singular values.

Step 2: Normalization. *Let σ_k denote the smallest retained singular value, which serves as the normalization reference. Define log-ratios:*

$$x_i = \log(\sigma_i/\sigma_k), \quad i = 1, \ldots, k-1.$$

By construction, $x_i > 0$ for all $i = 1, \ldots, k-1$ (since $\sigma_i > \sigma_k$ for $i < k$ in practice), so the geometric mean is well-defined. Note that σ_k itself is excluded from the flatness calculation — it serves purely as the reference point for normalization. The log-ratios are scale-invariant: multiplying all singular values by a constant leaves each x_i unchanged.

Step 3: Log-spectral flatness. *We define the (truncated) log-spectral flatness with truncation threshold p to be:*

$$\boxed{\mathrm{F}_p := \frac{\mathrm{GM}(x_1, \ldots, x_{k-1})}{\mathrm{AM}(x_1, \ldots, x_{k-1})} = \frac{\left(\prod_{i=1}^{k-1} \log(\sigma_i/\sigma_k)\right)^{1/(k-1)}}{\frac{1}{k-1}\sum_{i=1}^{k-1} \log(\sigma_i/\sigma_k)}.}$$

Interpretation. F_p measures how uniformly the log-spectrum is distributed among the top $k-1$ singular values, each measured relative to the reference value σ_k:

- Smooth spectrum (F_p close to 1): The log-ratios $\log(\sigma_i/\sigma_k)$ are approximately equal across $i = 1, \ldots, k-1$, meaning the log-spectrum decays gradually and uniformly among the retained values. This corresponds to a spectrum where many singular values contribute comparably, as we see prior to grokking.

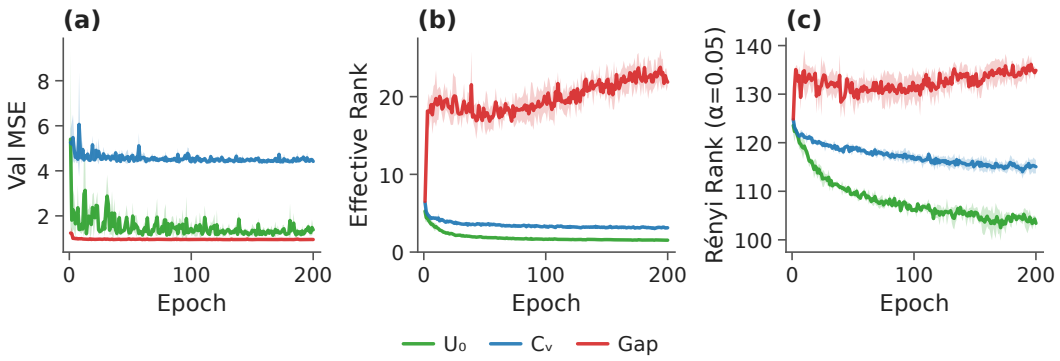

Figure 21: **QM9 training dynamics (MLP):** loss and spectral rank over 200 epochs.

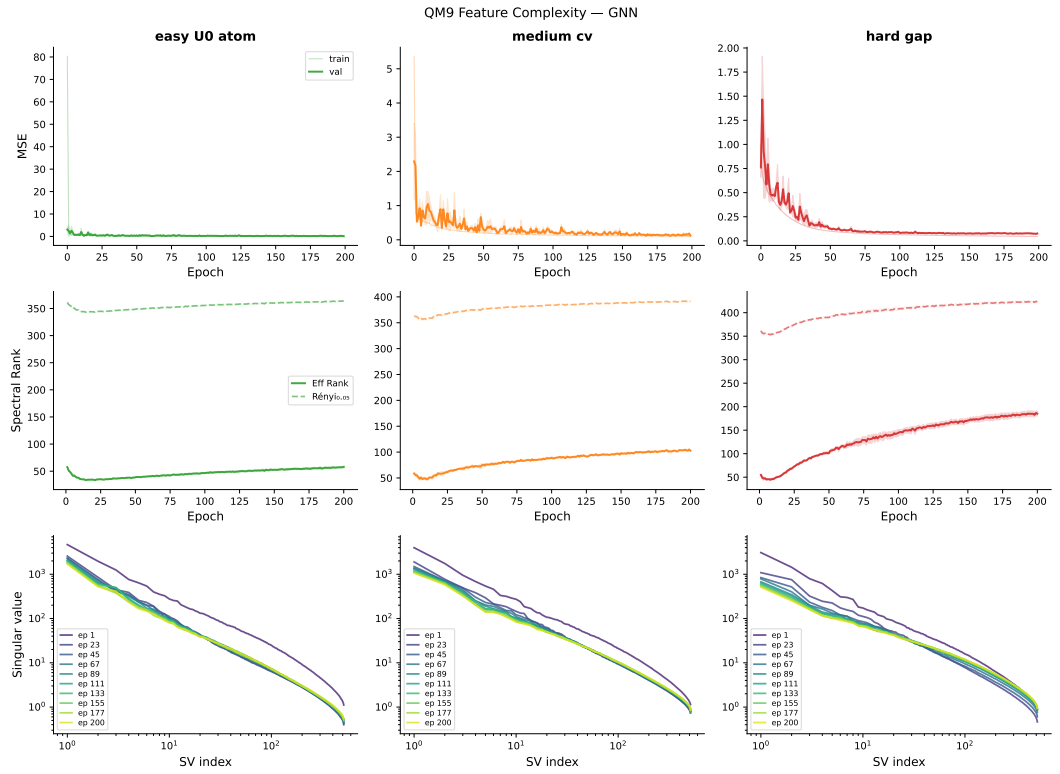

Figure 22: **QM9 (GNN):** same spectral rank ordering as MLP. GNN spectra show smoother power-law decay, consistent with message-passing weight sharing.

- Polarized spectrum (F_p close to 0): A small number of singular values are much larger than σ_k, while most retained values are only slightly larger than σ_k. The geometric mean of the log-ratios is collapsed by the many near-zero values $\log(\sigma_i/\sigma_k) \approx 0$, while the arithmetic mean is pulled up by the few large log-ratios. This corresponds to a polarized spectrum, as is seen after grokking.

Robustness. The choice of $p = 0.95$ is a hyperparameter. We find qualitatively similar results for $p \in \{0.90, 0.95, 0.99\}$; results are presented in Figures 38 and 39. The measure is undefined when all retained log-ratios are zero (i.e., when $\sigma_1 = \cdots = \sigma_k$, so that $x_i = 0$ for all $i = 1, \ldots, k-1$), but this degenerate case does not arise in practice. Note also that $k - 1 \geq 1$ requires $k \geq 2$, which is satisfied for any reasonable choice of p given the large number of singular values in our experiments.

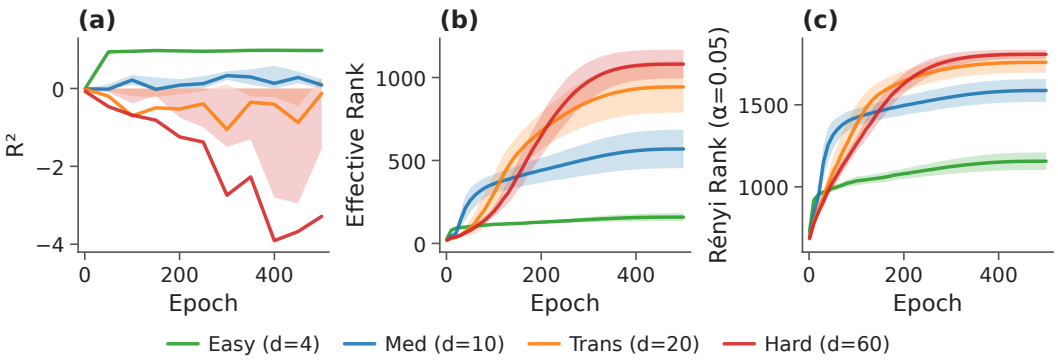

Figure 23: **Hermite with noise** (SNR = 25). Difficulty ordering preserved; rank shifted upward.

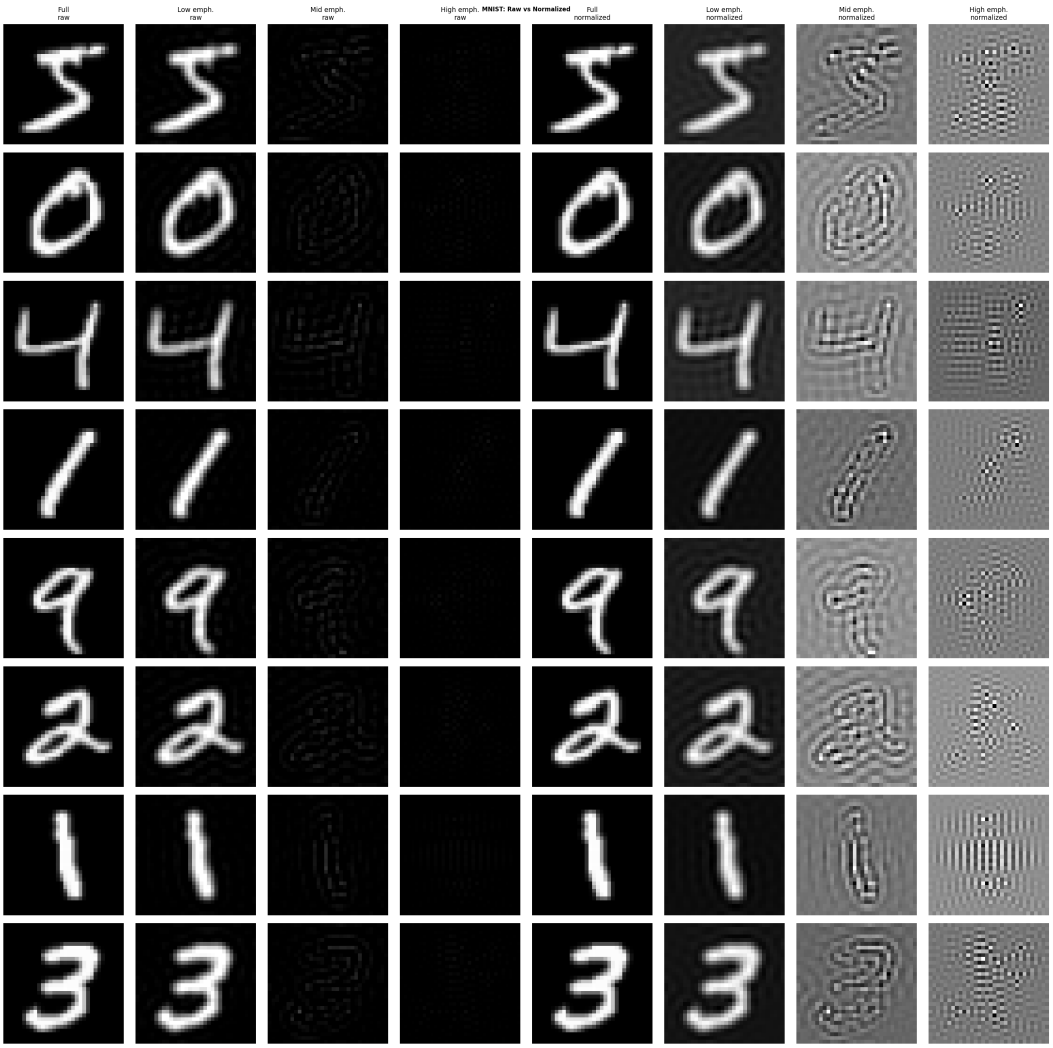

Figure 24: **DCT-MNIST task construction.** Low-frequency coefficients capture coarse shape; high-frequency coefficients capture edges and texture.

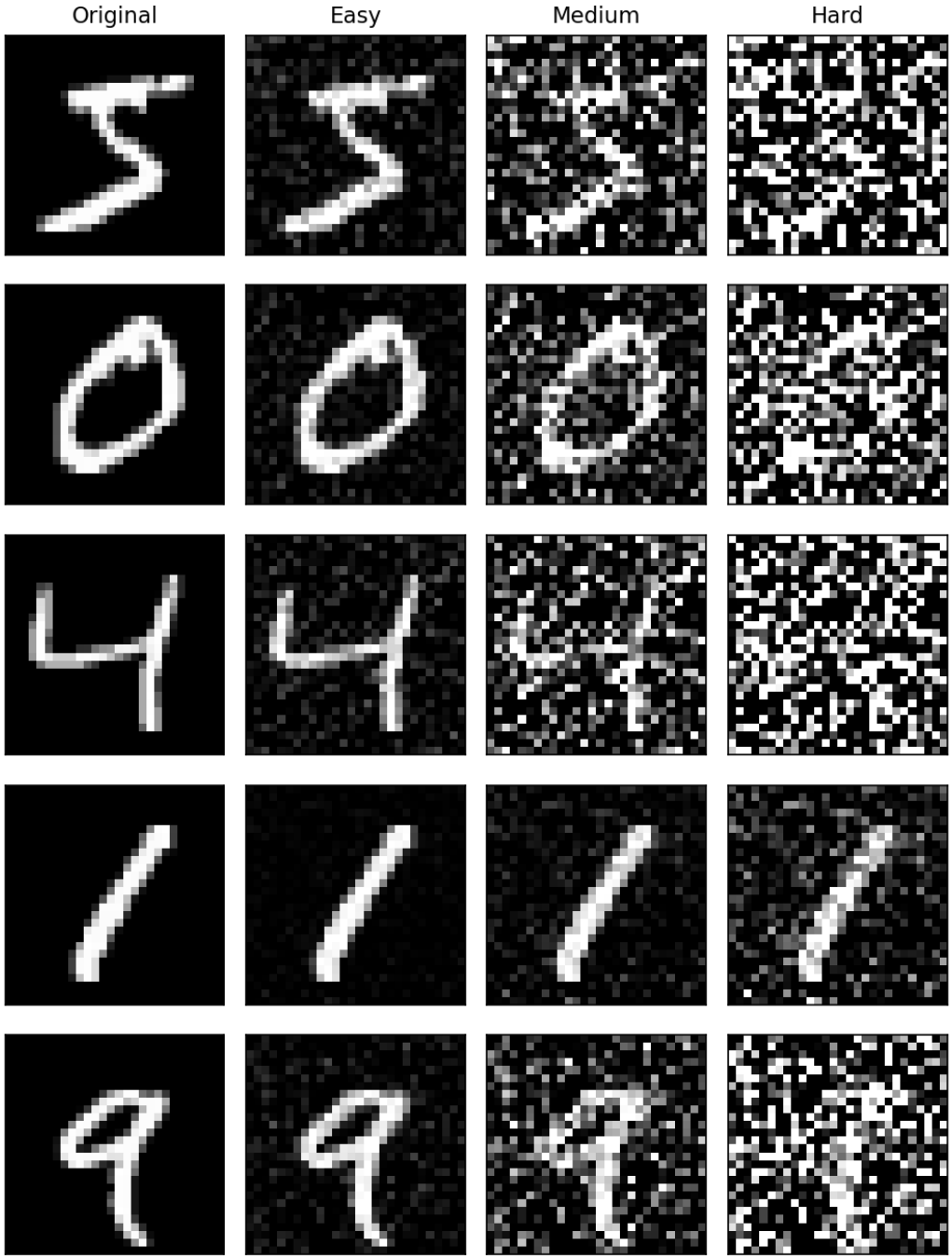

Figure 25: **Noisy MNIST data samples** Original MNIST images (left) alongside noisy variants at the Easy ($\alpha = 0.5$), Medium ($\alpha = 2$), and Hard ($\alpha = 5$) noise levels.

C.2 Dirichlet energy of the log spectrum

We define the Dirichlet energy of the log-spectrum as the normalized sum of squared differences of consecutive log-singular-values on a log-index scale:

$$\mathrm{TV}_2(\log \sigma) := \frac{1}{\log r} \sum_{i=1}^{r-1} \frac{(\log \sigma_{i+1} - \log \sigma_i)^2}{\log(i+1) - \log i}. \tag{3}$$

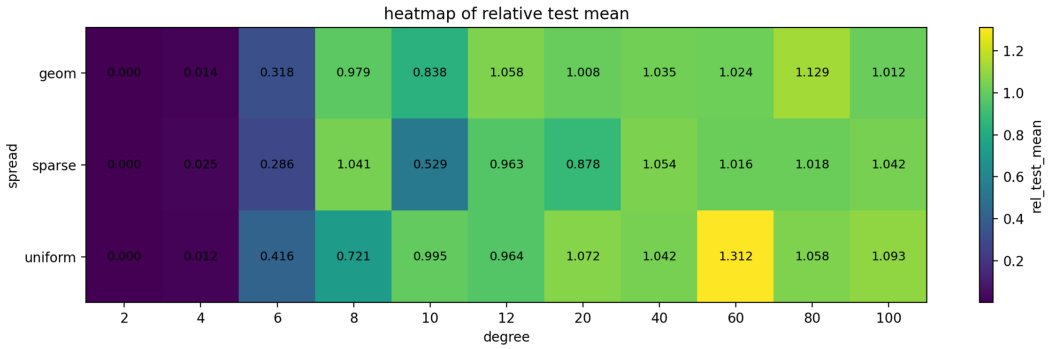

Figure 26: **Hermite task selection.** Relative test MSE vs. polynomial degree and coefficient distribution, guiding selection of difficulty levels.

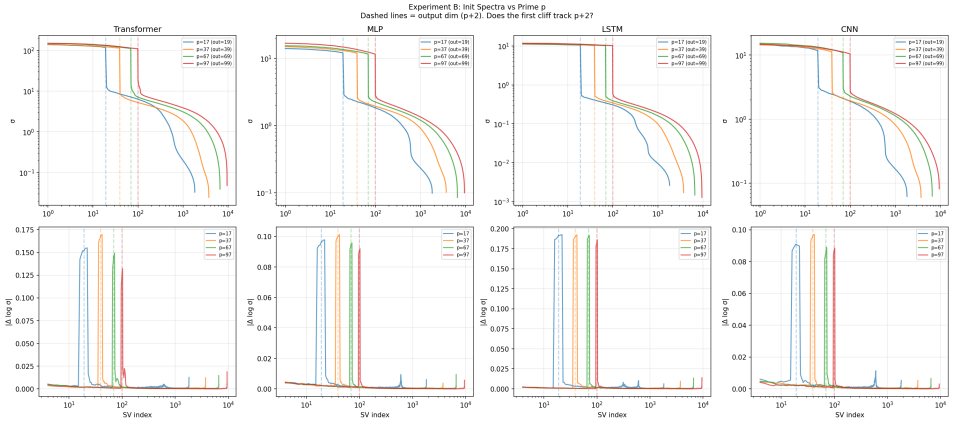

Figure 27: **The spectral cliff tracks the output dimension.** Spectra at initialization for varying p; the cliff moves with the output vocabulary size.

This adapts the classical Dirichlet energy $\int |f'|^2$ – which measures how rapidly a function varies, and is minimized by maximally smooth (harmonic) functions – to the discrete setting of a log-spectrum viewed as a function of log-index. The change of variables to log-index is necessary because overparameterized spectra contain thousands of near-zero singular values whose small random fluctuations would otherwise dominate the sum, obscuring the spectral shape in the informative head region. Normalizing by $\log r$ makes the quantity comparable across spectra of different lengths.

Why not standard total variation? The simpler alternative,

$$TV(\log \sigma) := \sum_{i=1}^{r-1} |\log \sigma_{i+1} - \log \sigma_i|$$

fails for monotonically sorted sequences: since $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, the absolute values are redundant and the sum telescopes to

$$\mathrm{TV}(\log \sigma) = \log \sigma_1 - \log \sigma_r = \log(\sigma_1/\sigma_r),$$

the log condition number. This is insensitive to whether the decay between the extremes is gradual or cliff-like, and is dominated by the numerically unreliable smallest singular value σ_r. TV_2 avoids both problems by measuring local squared differences rather than the global spread. The simpler alternative, the standard total variation $\sum |\Delta \log \sigma_i|$, does not work here because for a sorted sequence it telescopes to the log condition number $\log(\sigma_1/\sigma_r)$, which is the same regardless of whether the decay is smooth or jagged.

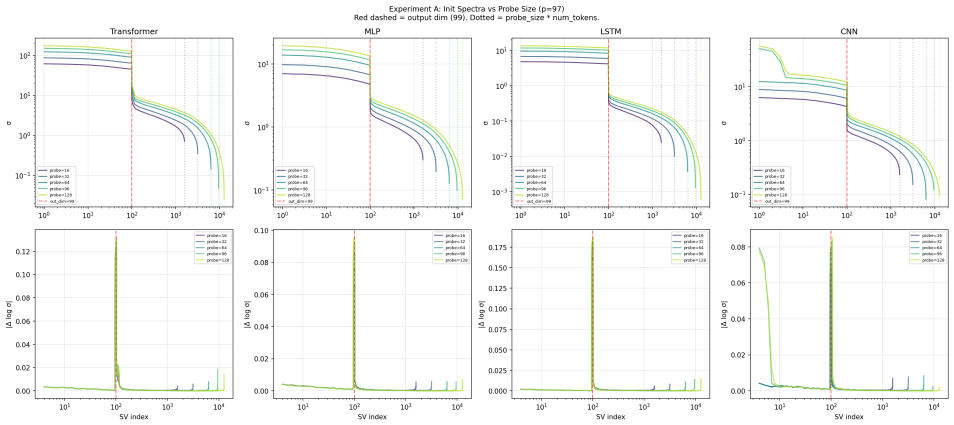

Figure 28: **Varying probe size does not move the cliff.** The cliff remains at the output dimension.

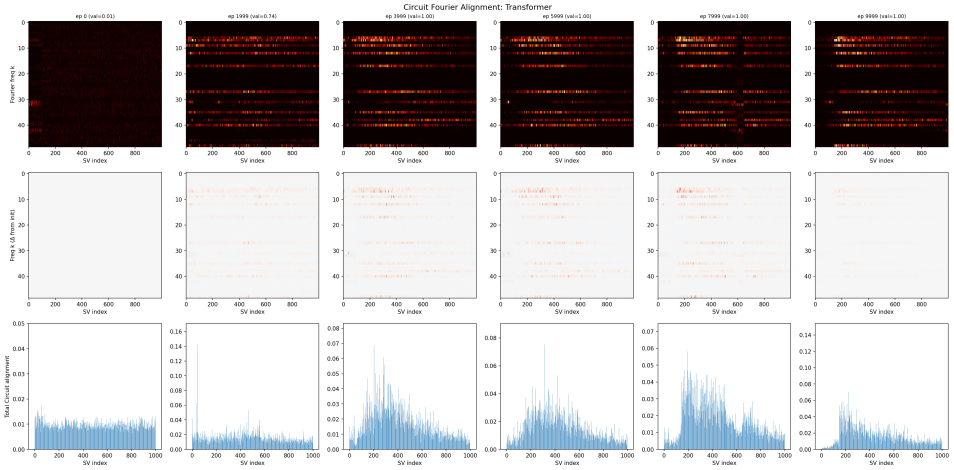

Figure 29: Transformer: Fourier circuit alignment of top 1000 left singular vectors at six training stages.

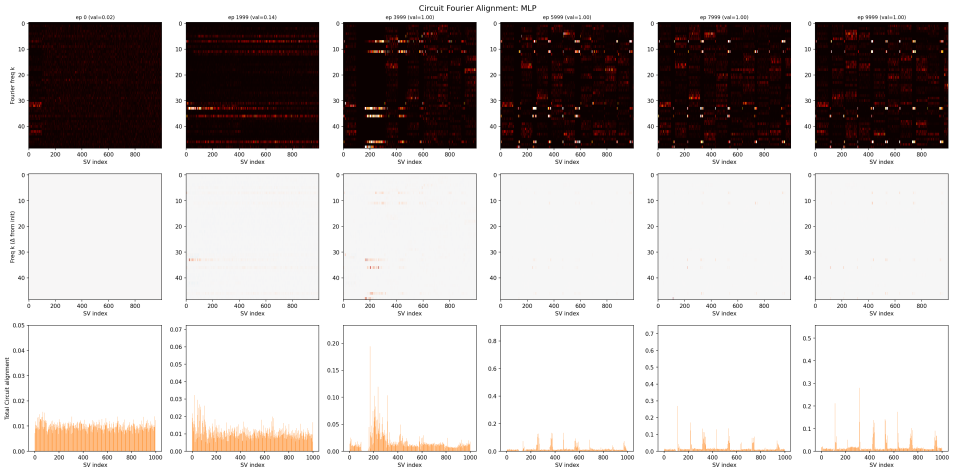

Figure 30: MLP: Fourier circuit alignment at six training stages.

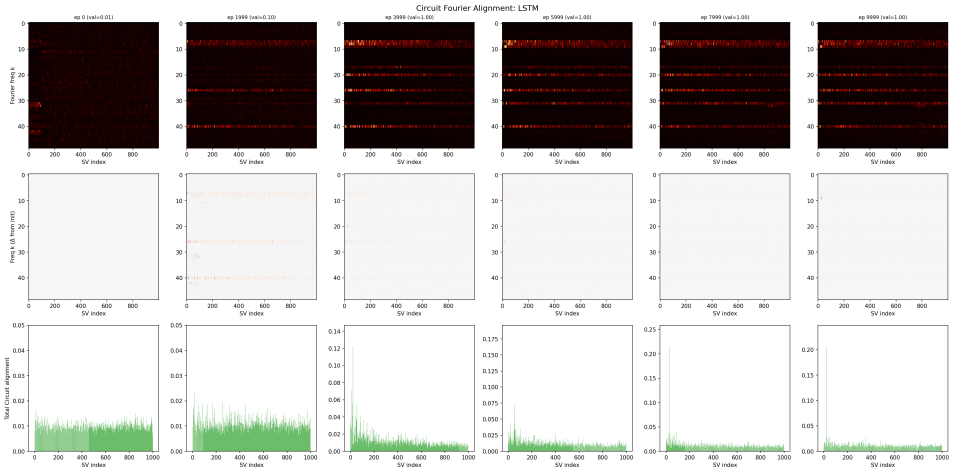

Figure 31: LSTM: Fourier circuit alignment at six training stages.

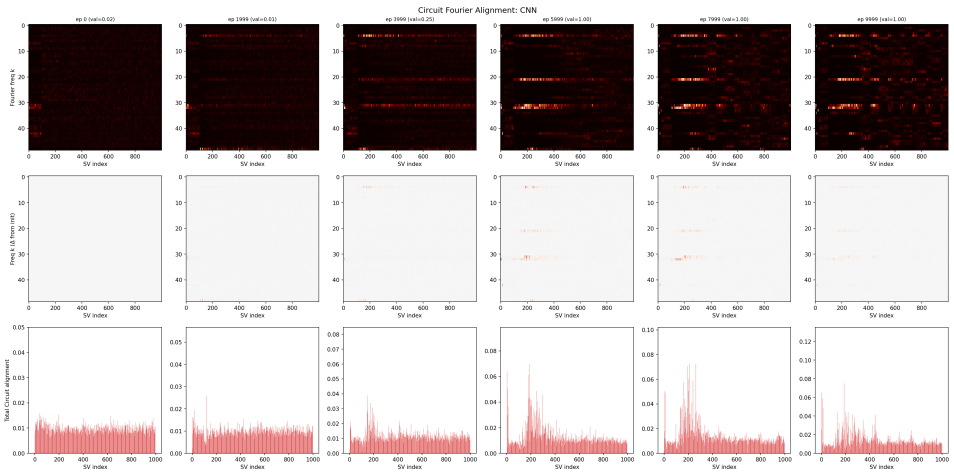

Figure 32: CNN: Fourier circuit alignment at six training stages.

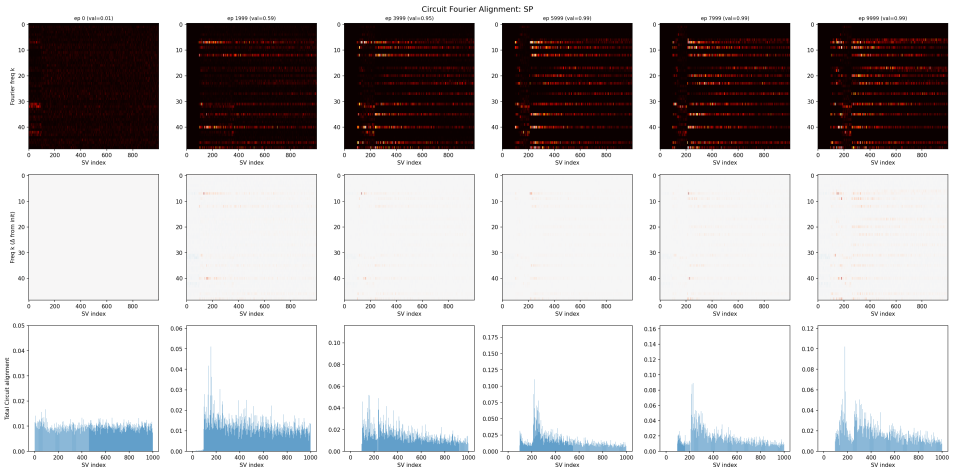

Figure 33: SP Transformer: Fourier circuit alignment at six training stages.

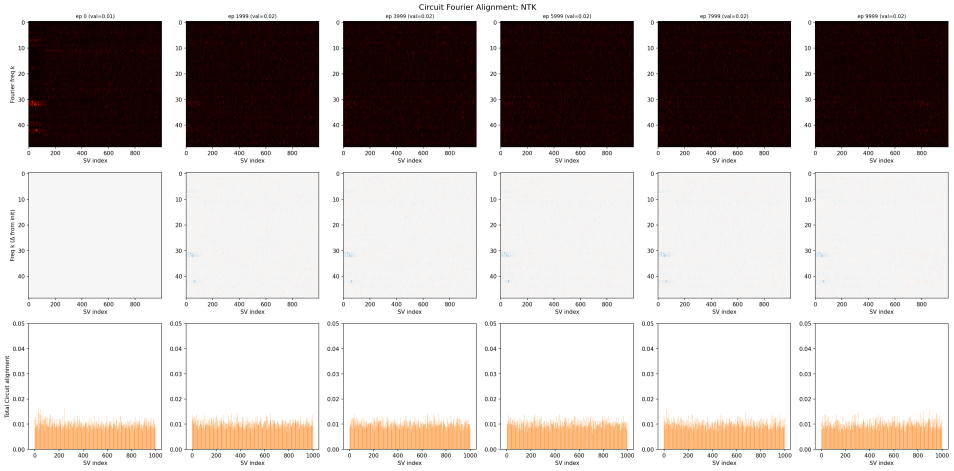

Figure 34: NTK Transformer: Fourier circuit alignment. No structured alignment develops.

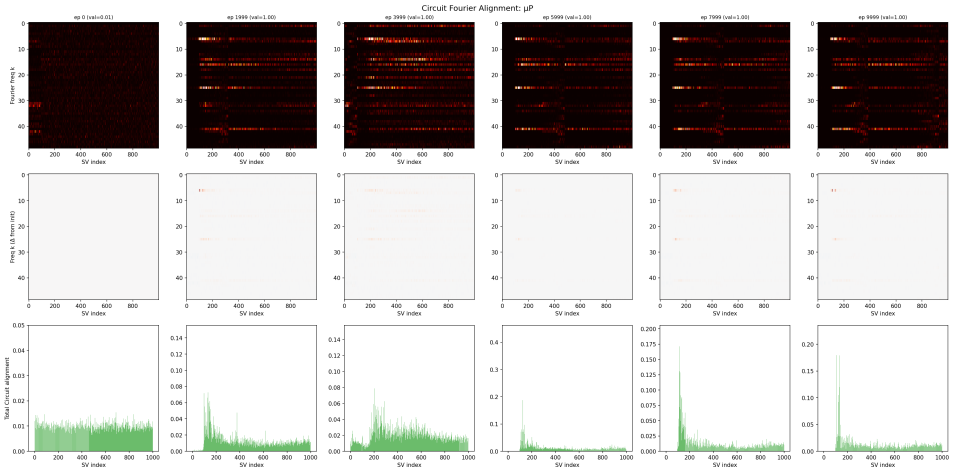

Figure 35: μP Transformer: Fourier circuit alignment at six training stages.

Table 4: **Hermite polynomial regression configuration.**

Architecture	MLP (4 hidden layers, width 512, GELU, 794k params)
Task difficulties	Degree 4 (easy), 10 (medium), 20 (transition), 60 (hard)
Optimizer	AdamW, lr $= 10^{-3}$
Epochs	500
Batch size	512
Probe set	2048 fixed samples
Seeds	42, 67, 420, 1234, 9999
Spectral method	Exact thin SVD (QR + SVD, float32), geometry every 10 epochs
Hardware	1 × A100

Table 5: **DCT-MNIST configuration.**

Architecture	MLP (bottleneck 128, 3×512, 728k params); CNN (6 conv, 651k params)
Task difficulties	Full, Low Freq (easy), Mid Freq (medium), High Freq (hard)
Optimizer	AdamW, lr $= 10^{-3}$, wd $= 0.01$
Epochs	100
Batch size	128
Probe set	128 fixed samples
Seeds	42, 67, 420, 1234, 9999
Spectral method	Exact thin SVD (QR + SVD, float32), geometry every epoch
Hardware	1 × A100

Table 6: **QM9 molecular property prediction configuration.**

Architecture	MLP (4 hidden layers, width 512, 794k params); GNN (3-layer GCN, width 256)
Targets	U_0 (local), C_v (semi-local), Gap (global)
Optimizer	AdamW, lr $= 10^{-3}$
Epochs	200
Batch size	128
Probe set	512 fixed samples
Seeds	42, 67, 1234
Spectral method	Exact thin SVD (QR + SVD, float32), geometry every epoch
Hardware	1 × A100

Table 7: **Noisy MNIST configuration.**

Architecture	MLP (2 hidden layers, width 32, 26k params
Task difficulties	Easy($\alpha = 0.5$), Medium($\alpha = 2$), Hard($\alpha = 5$)
Optimizer	Adam, lr $= 10^{-3}$
Epochs	20
Batch size	768
Probe set	1024 fixed training examples
Seeds	42, 67, 420, 1234, 9999
Spectral method	Exact thin SVD (QR + SVD, float32), geometry every 10 epochs
Hardware	4 x RTX 2080 Ti

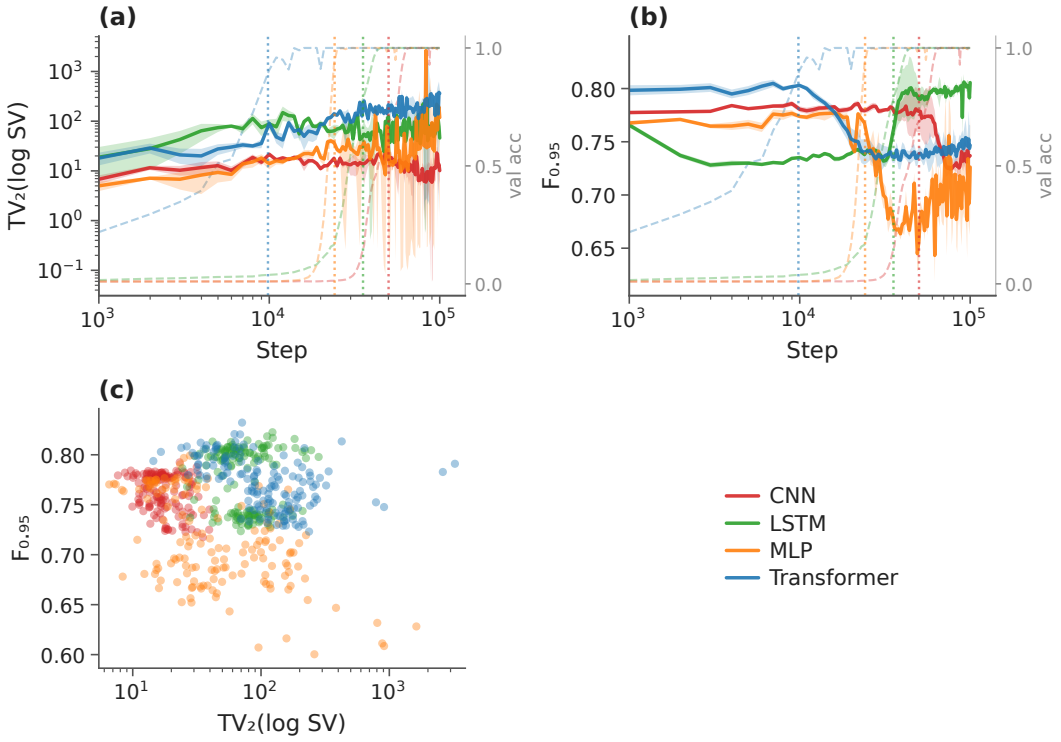

Figure 36: **Smoothness metric comparison (multi-architecture). (a)** TV_2(log SV) over training:the Dirichlet energy of the log-spectrum in log-index coordinates, which equals α^2 for a pure power law $\sigma_i \sim i^{-\alpha}$. Higher values indicate rougher, less uniform spectral decay. **(b)** Truncated log-spectral flatness $F_{0.95}$ ($p = 0.95$). Values near 1 indicate uniform log-spacing among the top 95% of singular values; values near 0 indicate a polarized or cliff-like spectrum. **(c)** Scatter of TV_2 vs. $F_{0.95}$ across all training snapshots: the two metrics seem to be somewhat contrary, but its also clear they encode different things. Dashed lines: validation accuracy overlay. Vertical dotted lines: mean grokking step.

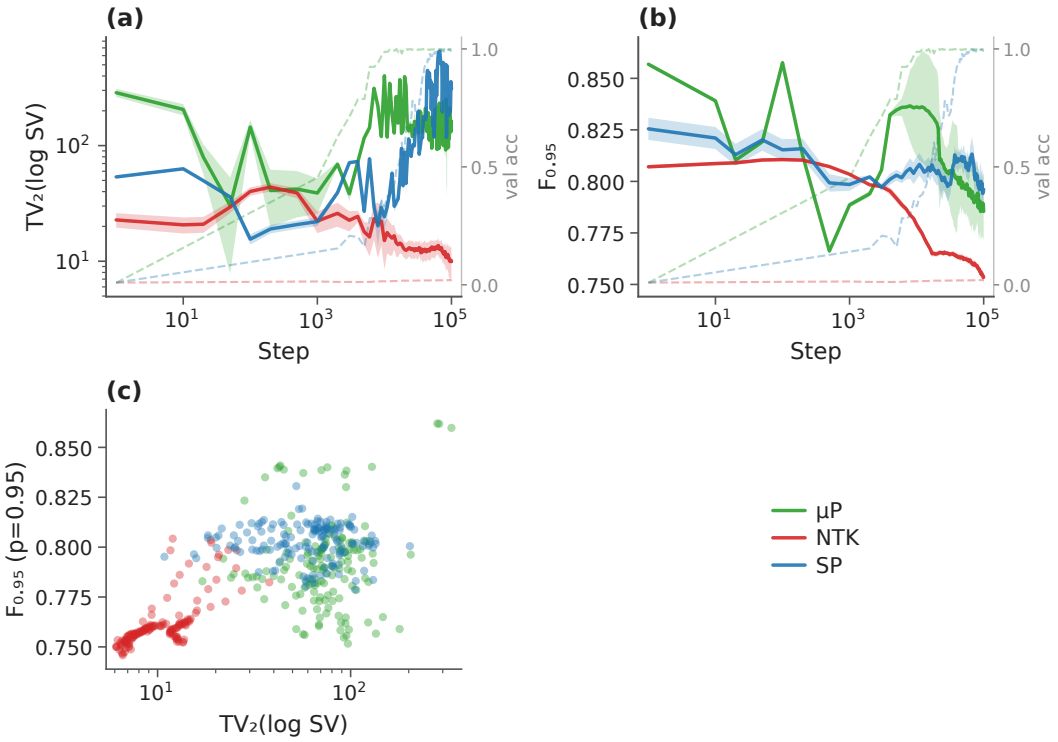

Figure 37: **Smoothness metric comparison (Feature Parameterizations, Width 128).** Same layout as fig. 36 but comparing SP, NTK, and μP parameterizations on the Transformer at width 128. μP shows the fastest spectral restructuring (TV_2 drops earliest), consistent with its faster grokking. NTK retains rough spectra throughout, reflecting its failure to grok. Here we also see the effect of weight decay, because of the lower weight decay for this experiment (0.01), the spectral cliff is largely unchanged, meaning flatness doesn't capture the grokking transition, but TV_2 does because the model is still locally changing its spectra.

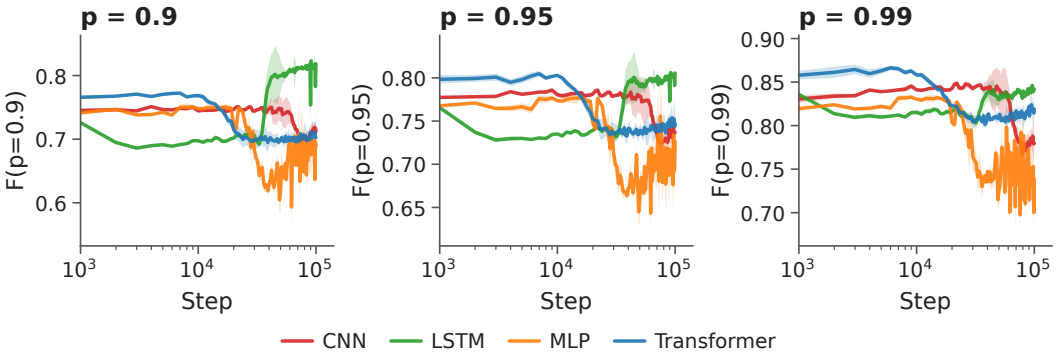

Figure 38: **Robustness of log-spectral flatness to truncation parameter p (multi-architecture).** F_p computed at $p \in \{0.90, 0.95, 0.99\}$. The qualitative ordering across architectures and the timing of flatness changes are stable across all three choices, confirming that $F_{0.95}$ is not sensitive to the specific truncation threshold.

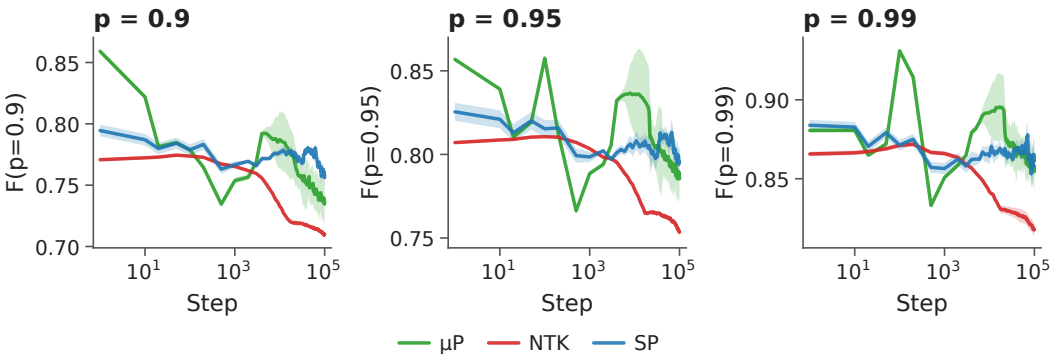

Figure 39: **Robustness of log-spectral flatness to truncation parameter p (parameterizations, w128).** Same as fig. 38 but for SP/NTK/μP. The relative separation between parameterizations is mostly preserved at all truncation levels.

Table 8: **Weight and Jacobian statistics from initialization to the final checkpoint.** Mean ± standard deviation over three ReLU Transformer runs. Weight effective rank is averaged over 11 matrices. Absolute weight and Jacobian ranks are not compared because they aggregate differently sized objects; the relevant quantities are their evolution and consistency across regimes.

Parameterization	Groks	Weight Frobenius	Weight eff. rank	Jacobian eff. rank	Jacobian pol.
SP	3/3	$119.2 \pm 0.7 \to 360.4 \pm 63.4$	$102.0 \pm 0.1 \to 85.1 \pm 0.6$	$2442 \pm 208 \to 1007 \pm 159$	$0.14 \pm 0.00 \to 0.60 \pm 0.02$
NTK	0/3	$647.9 \pm 0.7 \to 592.6 \pm 0.7$	$101.9 \pm 0.1 \to 101.8 \pm 0.0$	$3526 \pm 59 \to 3510 \pm 33$	$0.49 \pm 0.02 \to 0.54 \pm 0.05$
μP	3/3	$10.9 \pm 0.0 \to 180.6 \pm 55.5$	$101.9 \pm 0.0 \to 74.5 \pm 4.8$	$207 \pm 1 \to 985 \pm 353$	$0.19 \pm 0.01 \to 0.61 \pm 0.06$

C.3 Testing the neural spectrum: additional details

Our central distinction is between a parameter representation and the local space of functions accessible from it. For the evaluation map $\Phi_Z(\theta) = f_\theta(Z)$, the parameter Jacobian $J_Z = D\Phi_Z(\theta)$ maps infinitesimal parameter changes to changes in the model outputs. Parameter symmetries make Φ_Z many-to-one: movement along a symmetry orbit changes the weights without changing the represented function, and tangent directions to this orbit lie in $\ker J_Z$. We test 3 concrete predictions: Jacobian statistics should organize feature-learning dynamics more consistently than weight norms or weight-matrix ranks, especially when the parameter representative is changed without changing the function, the sharp SP polarization should vary systematically across the lazy-rich continuum if it reflects a lazy-to-rich transition rather than a universal signature of generalization and spectral reorganization is causal, so directly encouraging or suppressing it should alter grokking even after controlling for generic gradient regularization and weight decay. We test each prediction below.

Definitions. A step is one optimizer update, and "(k)" denotes thousands of steps. The memorization step is the first step with training accuracy ≥ 0.99. With $\mathrm{pol}(t) = \log_{10}(\sigma_1/\sigma_{50})$ measured on the fixed probe, $\Delta\mathrm{pol} = \max_t \mathrm{pol}(t) - \mathrm{pol}(0)$ is the peak rise in polarization, and the onset t_{pol} is the first step at which $\mathrm{pol}(t) - \mathrm{pol}(0) \geq 0.5$ (a further half order of magnitude in σ_1/σ_{50}), sustained over consecutive measurements after a brief warm-up transient shared by all runs. Speedup in Table 11 is the baseline's mean grokking step divided by the arm's.

Test 1: do Jacobian statistics add information beyond the weights? We repeat the width-128 SP, NTK, and μP experiments using three seeds and the same 58 checkpoints over 100k steps. These Transformers use ReLU in their feed-forward blocks so that the positive-rescaling symmetry used below is exact; their grokking times therefore need not match the GELU runs in Section 4.2. We compare Jacobian effective rank and polarization with the global weight Frobenius norm and the mean effective rank of the model's 11 matrix-valued weight tensors.

The two families of statistics separate immediately (Table 8). At initialization, weight effective rank is 101.9–102.1 across all nine runs, a spread of 0.2%, whereas Jacobian effective rank ranges from approximately 206 under μP to 3579 under NTK. This is not an intrinsic complexity ordering of the initial functions, since the Euclidean Jacobian spectrum also depends on parameter coordinates. Rather, it shows that J registers the tangent geometry that distinguishes the lazy and rich update regimes, while normalized layerwise weight spectra do not. Weight Frobenius norm distinguishes the parameterizations only through their different prescribed scales (648, 119, and 10.9 at initialization).

The training trajectories give a stronger comparison. When the three SP runs are aligned to their individual grokking steps, Jacobian effective rank makes its largest move in the 1k steps immediately before grokking, falling from 5025 to 2703 (46%), while validation accuracy rises from 0.61 to 0.84. Polarization makes its steepest increase over the same interval. Weight effective rank changes by only 1.5% and continues declining at a similar rate before and after the transition. Across all six SP and μP runs that grok, Jacobian effective rank changes by 36–77% in five runs, compared with 1.2–9.4% for weight effective rank; the weight norm rises by anywhere from 25% to 568%, reaching values from 39 to 314 at 99% validation accuracy. The exception is the earliest-grokking μP seed, whose Jacobian effective rank rises by 19% over the transition. We retain this case because it is informative: μP begins in a rich regime, so spectral compression need not coincide with generalization. Overall, the weight statistics

Table 9: **Moving along a parameter-symmetry orbit.** One grokked SP checkpoint is shown across the full sweep. The function is exactly preserved in real arithmetic and prediction agreement is 1.0000 throughout. Numerical J rank uses the paper's fixed threshold and becomes unstable under severe ill-conditioning; exact rank is invariant by Equation (4).

κ	Max $\lvert\Delta\text{logit}\rvert$	× weight Frob.	× weight nuc.	Weight eff. rank	Numerical J rank	J eff. rank	J pol.
Original	0	1.00	1.00	90.14	227	2743.6	0.497
3	2.44×10^{-3}	1.00	1.05	88.43	226	2744.0	0.495
10	2.44×10^{-3}	1.02	1.21	83.70	245	2688.3	0.458
10^2	2.93×10^{-3}	1.17	1.97	74.61	179	2477.7	0.481
10^3	5.86×10^{-3}	1.96	4.15	66.55	262	2322.6	0.379
10^4	2.93×10^{-2}	4.76	9.92	60.47	159	2160.7	0.493

are correlated with training, but no common weight-norm scale or weight-rank transition organizes SP, NTK, and μP.

Weight-space comparison and parameter symmetries. We test whether the parameter Jacobian provides a useful account of feature learning beyond summaries of the weights due to its functional nature. For the evaluation map $\Phi_Z(\theta) = f_\theta(Z)$, its Jacobian $J_Z = D\Phi_Z(\theta)$ maps local parameter changes to changes in model outputs. We examine three predictions: Jacobian statistics should track training transitions that weight summaries miss; spectral polarization should depend on the learning regime; and directly changing the spectrum should change the transition.

Across three seeds each of SP, NTK, and μP, initial weight-matrix effective ranks are nearly identical, while Jacobian spectra differ substantially. Around the SP grokking transition, Jacobian effective rank falls by 46% over 1k steps as validation accuracy rises from 0.61 to 0.84; mean weight effective rank changes by only 1.5%. The same spectral compression is not required under μP, where feature learning begins earlier. Thus the Jacobian trajectory distinguishes how these regimes reach generalization more clearly than either weight norm or mean weight rank.

A controlled symmetry-orbit test. The trajectory comparison still changes both the parameters and the function. To isolate dependence on the parameter representative, we transform the six grokked ReLU checkpoints along exact parameter-symmetry orbits. We combine the $GL(d_{\text{head}})$ change-of-basis symmetries of query/key and value/output projections with positive per-neuron ReLU rescalings $W_1 \mapsto DW_1$, $W_2 \mapsto W_2 D^{-1}$, varying the condition number of the composite transformation up to $\kappa = 10^4$. These transformations preserve the represented function exactly.

The symmetry framing requires a precise qualification. If g is a smooth, invertible parameter symmetry, $f_{g(\theta)} = f_\theta$, and $T = Dg_\theta$, then

$$J_Z(g(\theta)) = J_Z(\theta)T^{-1}. \tag{4}$$

Consequently, $\operatorname{rank} J_Z$ and $\operatorname{im} J_Z$ are invariant, while its Euclidean singular values can change with the conditioning of T. We therefore do not claim that the full neural spectrum is invariant to parameter symmetry. The exact functional content is the rank and output-space tangent image; the spectrum gives a graded, coordinate-dependent description of sensitivity within that map.

At $\kappa = 10^4$, the global weight Frobenius and nuclear norms of this checkpoint change by 4.76× and 9.92×, and weight effective rank falls from 90.14 to 60.47, despite an unchanged function (Table 9). Repeating the transformation on all three SP and three μP checkpoints gives the same qualitative result: weight Frobenius changes by 3.59–20.1×, weight nuclear norm by 9.67–15.1×, and weight effective rank by $-33.2\% \pm 1.5\%$. Jacobian effective rank is not invariant, but changes by $-11.8\% \pm 7.6\%$, while the mean absolute change in polarization is 0.09. This is an empirical comparison for the tested transformation family, not a universal stability bound. Its purpose is to isolate the identification problem: weight statistics can be moved sharply with no functional change, whereas the Jacobian rank and output-space image are fixed and its normalized spectral shape is substantially less affected here.

Table 10: **Continuous interpolation from NTK to μP.** Mean $\pm$ standard deviation over three seeds. D_{shape} is the RMS distance from the mean-centered log-spectrum at initialization; rise@1k is the fraction of peak polarization completed by step 1k; align@1k is the readout log-alignment ratio of Everett et al. (2024), where 0.5 indicates no alignment.

θ	Groks	Memorize (k)	Final val.	D_{shape}	Δpol	Eff. rank (init $\to$ end)	Rise@1k	Grok (k)	Align@1k
0.00 (NTK)	0/3	10.7 ± 0.0	0.03 ± 0.01	0.06 ± 0.01	0.22 ± 0.03	$4788 \pm 80 \to 4474 \pm 17$	–	never	0.47 ± 0.00
0.17	0/3	3.8 ± 0.1	0.34 ± 0.08	0.08 ± 0.01	0.23 ± 0.03	$4300 \pm 78 \to 3952 \pm 27$	–	never	0.50 ± 0.01
0.33	0/3	1.1 ± 0.0	0.52 ± 0.01	0.06 ± 0.01	0.66 ± 0.18	$3685 \pm 75 \to 3440 \pm 740$	33%	never	0.53 ± 0.00
0.50 (SP)	3/3	0.4 ± 0.0	0.99 ± 0.00	0.17 ± 0.03	0.94 ± 0.15	$2964 \pm 63 \to 558 \pm 69$	15%	27.6 ± 5.6	0.55 ± 0.00
0.67	3/3	0.6 ± 0.0	0.99 ± 0.00	0.13 ± 0.01	0.87 ± 0.07	$1132 \pm 14 \to 1429 \pm 881$	35%	46.1 ± 3.1	0.55 ± 0.01
0.83	3/3	0.8 ± 0.1	1.00 ± 0.00	0.20 ± 0.02	0.83 ± 0.06	$456 \pm 7 \to 540 \pm 22$	56%	12.7 ± 9.5	0.60 ± 0.02
1.00 (μP)	3/3	1.0 ± 0.1	1.00 ± 0.00	0.30 ± 0.04	0.71 ± 0.12	$211 \pm 5 \to 973 \pm 268$	56%	14.5 ± 9.4	0.62 ± 0.01

Table 11: **Direct spectral intervention at weight decay** 0.01**.** Mean $\pm$ standard deviation over three seeds with shared initialization within each seed. Antipolarization preserves memorization but prevents grokking; gradnorm accelerates grokking without polarization preceding it.

Arm	Grok (k)	Speedup	Memorize (k)	Train acc.	Val. acc.	Δpol	t_{pol} (k)	Eff. rank (end)
Baseline	38.1 ± 2.6	1.0×	0.43 ± 0.00	1.000 ± 0.000	0.995 ± 0.004	0.64 ± 0.02	53.3 ± 15.9	858 ± 412
Polarize	7.3 ± 1.4	5.2×	0.43 ± 0.01	1.000 ± 0.000	0.999 ± 0.001	2.82 ± 0.26	2.0 ± 0.0	87 ± 66
Antipolarize	never (0/3)	–	0.42 ± 0.01	1.000 ± 0.000	0.413 ± 0.015	0.30 ± 0.00	never	8986 ± 38
Gradnorm	11.9 ± 1.4	3.2×	0.75 ± 0.02	1.000 ± 0.000	0.995 ± 0.004	0.80 ± 0.07	36.0 ± 3.3	787 ± 545

Test 2: does polarization specifically mark the lazy-to-rich transition? The discrete SP/NTK/μP comparison in Section 4.2 leaves open whether our explanation of their different timing is specific to three hand-selected endpoints. We therefore interpolate continuously from NTK to μP using the *abc* framework of Everett et al. (2024). The path $\theta \in [0, 1]$ log-linearly interpolates, for each layer class, the initialization scale, learning rate, weight decay, and attention temperature of our NTK ($\theta = 0$), SP ($\theta = 0.5$), and μP ($\theta = 1$) implementations. Equivalently, it is a linear path in the corresponding *abc* exponents. We verify the endpoints against the original implementations by copying weights; their outputs differ by less than 3×10^{-7} throughout training. We run three seeds at each of seven points with weight decay 0.01.

The sweep recovers a continuous lazy–rich picture (Table 10). On the NTK side, spectral shape remains close to initialization and no model groks within the budget. At SP, grokking first appears on the tested grid and coincides with the strongest effective-rank compression, from 2964 ± 63 to 558 ± 69. Toward μP, the initial spectrum is already concentrated and a larger fraction of its eventual restructuring occurs by step 1k. At the μP endpoint, effective rank begins at 211 ± 5 and later increases rather than undergoing an SP-like collapse. Thus, polarization is most phase-like when SP leaves an initially lazy regime. It is absent toward NTK because feature learning is suppressed and less temporally localized toward μP because the model is rich from initialization. This directly tests the explanation proposed for Section 4.2: polarization is a regime-dependent manifestation of spectral reorganization, not a universal requirement for generalization.

Test 3: does spectral restructuring contribute to the transition? The preceding experiments remain observational. To intervene directly, we add a differentiable estimate of the fraction of Jacobian spectral mass outside its leading eight directions. The estimate uses 64 random projections of the Jacobian on fresh 64-example training batches and requires one batched VJP per step. Minimizing it encourages polarization; reversing its sign encourages a diffuse spectrum. We also use a gradient-norm penalty as a nonspectral control for generic second-order gradient pressure. The sketch is used only for training: every reported spectrum is the exact SVD on the fixed 96-example probe. Across 228 checkpoints, the sketched head-mass statistic correlates with exact polarization at $r = 0.967$.

Each arm adds a single term $\lambda_{\text{reg}} P$ to the training loss. For the polarizing arm, P is the sketched fraction of spectral mass outside the leading eight directions; the antipolarizing arm uses $-P$; and the gradient-norm control uses $P = \|\nabla_\theta \mathcal{L}\|_2^2$. We use $\lambda_{\text{reg}} = 0.1$ for all runs.

Encouraging polarization advances its onset to step 2k and accelerates grokking 5.2×, without changing the memorization time (Table 11). Reversing the same estimator keeps effective

Table 12: **Weight-decay sweep under SP.** Mean ± standard deviation over three seeds. The ratio is computed per seed before averaging. Decay moves the timing of both grokking and polarization but does not monotonically determine the final Jacobian geometry.

WD	Grok (k)	t_{pol} (k)	Grok/t_{pol}	Δpol	Eff. rank (end)	Final val.
0.01	38.1 ± 2.6	41.0 ± 11.3	0.99 ± 0.24	0.73 ± 0.08	858 ± 412	1.00 ± 0.00
0.05	14.1 ± 1.1	20.3 ± 1.7	0.70 ± 0.03	1.34 ± 0.03	757 ± 607	1.00 ± 0.00
0.10	9.3 ± 0.7	16.3 ± 4.2	0.60 ± 0.13	1.02 ± 0.08	1335 ± 82	1.00 ± 0.00
0.25	4.1 ± 0.4	5.0 ± 0.0	0.81 ± 0.08	1.18 ± 0.02	921 ± 9	1.00 ± 0.00
0.50	2.1 ± 0.2	2.7 ± 0.5	0.84 ± 0.23	1.18 ± 0.02	693 ± 29	1.00 ± 0.00
1.00	1.1 ± 0.1	1.3 ± 0.5	0.92 ± 0.23	1.14 ± 0.04	533 ± 80	0.96 ± 0.06

Table 13: **Polarization intervention without weight decay.** Mean ± standard deviation over two seeds. Grok95/Grok99 are the first steps reaching 95%/99% validation accuracy.

Arm	Grok95 (k)	Grok99 (k)	Final val.	Δpol	t_{pol} (k)	Eff. rank
Baseline	54.7 ± 12.1	never	0.978 ± 0.003	0.46 ± 0.01	never	1672 ± 712
Polarize	8.7 ± 3.8	10.9 ± 5.3	0.9998 ± 0.0003	3.36 ± 0.01	3.0 ± 1.4	5 ± 1

rank near its maximum (8986 of 9504) and prevents grokking in all three seeds despite perfect training accuracy. Gradnorm also accelerates grokking, showing that polarization is not the only route to generalization, but its spectral onset occurs at 36.0k, well after grokking at 11.9k. The polarizing arm instead changes the spectrum before generalization. Directly advancing or suppressing the spectral change therefore alters the SP transition in the predicted direction. This is evidence for a causal role in this setting, not for uniqueness or optimality of the spectral regularizer.

Why test weight decay? The intervention above is initially performed on top of weight decay 0.01, and weight decay is already known to accelerate grokking and change parameter norms. This does not imply that weight decay causes polarization. It creates a specific confound: the observed timing and the intervention's benefit might both be inherited from the same decay-driven optimization path. We address the two parts separately. A decay sweep asks whether polarization continues to track the transition when decay changes its time scale; a zero-decay rerun asks whether the spectral intervention works when that mechanism is absent.

Increasing decay accelerates grokking from 38.1k to 1.1k and polarization onset from 41.0k to 1.3k (Table 12). Across the two-order-of-magnitude sweep, their mean per-seed timing ratio stays between 0.60 and 0.99, while final effective rank and peak polarization are not monotone in decay. Thus, decay controls how quickly SP reaches the transition, but it is not equivalent to the resulting spectral geometry.

With weight decay removed, both arms still memorize at step 0.4k, but the polarizing objective accelerates the 95% transition by $6.6 \pm 1.5\times$ when ratios are computed per seed (Table 13). In both runs, polarization precedes generalization, and only the polarizing arm reaches 99% validation accuracy. This removes weight decay as a necessary condition for the intervention's effect. We do not claim that the spectral loss is a better practical accelerator than tuned decay; its role is to test the relationship directly.

Checkpoint-scale feasibility. The spectrum is intended for checkpoint analysis rather than per-step monitoring. As a scale test, we stream a 2000×2000 Gram matrix of per-example loss gradients for a 150M-parameter, 12-layer GPT-2/NanoGPT-style model (Karpathy, 2022) on character-level Shakespeare. Computing gradients in chunks of 32, accumulating the Gram in fp64, and eigendecomposing it on CPU takes 30–90 minutes per snapshot on one A100 80GB. This empirical-Fisher variant is not the full output Jacobian used in the main experiments, but demonstrates the same pipeline two orders of magnitude beyond their parameter count.