Grokking and Generalization Collapse: Insights from HTSR theory

Hari K. Prakash HPRAKASH@UCSD.EDU
University of California San Diego, Data Science and Engineering

Charles H. Martin CHARLES@CALCULATIONCONSULTING.COM
Calculation Consulting, 8 Locksley Ave, 6B, San Francisco, CA 94122

Abstract

We study the well-known grokking phenomena in neural networks (NNs) using a 3-layer MLP trained on 1 k-sample subset of MNIST, with and without weight decay, and discover a novel third phase —*anti-grokking*–that occurs very late in training and resembles but is distinct from the familiar *pre-grokking* phases: test accuracy collapses while training accuracy stays perfect This late-stage collapse is distinct, however, from the known *pre-grokking* and *grokking* phases, and is not detected by other proposed grokking progress measures. Leveraging Heavy-Tailed Self-Regularization (HTSR) through the open-source `WeightWatcher` tool, we show that the HTSR layer quality metric α delineates *all three* phases. The anti-grokking is revealed by training for 10^7 and is invariably heralded by $\alpha < 2$ and the appearance of *Correlation Traps*—outlier singular values in the randomized layer weight matrices that make the layer weight matrix *atypical* and signal overfitting of the training set. Such traps are verified by visual inspection of the layer-wise empirical spectral densities, and using Kolmogorov–Smirnov tests on randomized spectra. Comparative metrics, including activation sparsity, absolute weight entropy, circuit complexity, and l^2 weight norms track pre-grokking and grokking but fail to distinguish grokking from anti-grokking. This discovery provides a way to measure overfitting and generalization collapse without direct access to the test data. These results strengthen the claim that the HTSR α provides universal layer-convergence target at $\alpha \approx 2$ and underscore the value of using the HTSR alpha (α) metric as a measure of generalization.

1. Introduction

Grokking is an intriguing phenomenon where a neural network achieves near-perfect training accuracy quickly, yet the test accuracy lags significantly, often near chance level, before abruptly surging towards high generalization [15]. Figure 1 illustrates this for a depth-3, width-200 ReLU MLP trained on a subset of MNIST.

To dissect this phenomenon and uncover deeper dynamics, our primary analytical lens is the recently developed theory of Heavy-Tailed Self-Regularization (HTSR), following Martin et.al.[10]. The HTSR theory examines the empirical spectral density (ESD) of individual layer weight matrices $(\mathbf{W})$, quantified by the heavy-tailed power law (PL) exponent α. We find α provides a sensitive measure of correlation structure within layers, tracking the transition into the grokking phase, and crucially, predicting a subsequent decrease in generalization.

For comparative context, we also investigate several other methodologies including Weight Norm Analysis [6] and Progress Measures [2] (Activation Sparsity, Absolute Weight Entropy, and Approximate Local Circuit Complexity). Detailed definitions for these comparative metrics are provided in Appendix 2. We observe that grokking occurs even without weight decay (leading to an increasing norm), confirming the weight-norm related findings by Golechha [2].

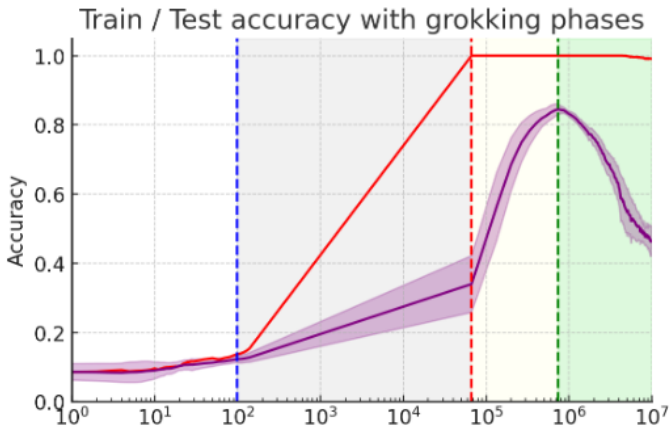

Figure 1: **The three phases of grokking.** Training curves for a depth-3, width-200 MLP on MNIST. The initial **pre-grokking** phase (grey): training accuracy (red line) surges at 10^2 steps, saturating between $10^4 - 10^5$ steps, while test accuracy (purple line) remains low; the **grokking** phase (yellow): with test accuracy rapidly increasing after $\sim 10^5$ steps, and reaching a maximum at 10^6 steps; and the newly revealed late-stage **anti-grokking** phase (green): test accuracy collapses (to 0.5).

Our Contributions: Our work makes several related contributions that helps explain the underlying mechanisms associated with the grokking phenomena:

1. By extending training significantly (up to 10^7 steps) under zero weight decay (`WD=0`), we identify and characterize **late-stage generalization collapse**: a substantial drop in test accuracy long after initial grokking, despite perfect training accuracy and a continually increasing l^2 weight norm. We call this **anti-grokking.**

2. We show that the `HTSR` layer quality metric α (the heavy-tailed power-law (PL) exponent), effectively tracks the grokking transition under both the traditional setting of weight decay (`WD>0`) and zero weight decay `WD=0`), outperforming the l^2 weight norm and the other progress metrics.

3. We identify the mechanism of the pre-grokking phase, where the training accuracy is perfect but the model does not generalize. This phase occurs because only a subset of the model layers are well trained (i.e. $\alpha \leq 4$), whereas at least one layer is underfit (i.e. $\alpha \geq 5$). Moreover, the layers can show great variability between training runs, indicating their instability. Importantly, the layer α's here are distinct from those in the anti-grokking phases, despite both phases having perfect training accuracy and low test accuracy. .

4. We demonstrate that an `HTSR` PL exponent $\alpha < 2$ correlates to the with the onset of the antigrokking phase. Also, in this phase, we observe the presence of anomalous rank-one (or greater) perturbations in one or more underlying layer weight matrices $\mathbf{W}$. We call these **correlation traps** and identify them by randomizing $\mathbf{W}$ elementwise, forming $\mathbf{W}^{rand}$, and looking for unusually large eigenvalues, $\lambda_{trap} \gg \lambda^+$ (where λ^+ is the right-most edge of the associated Marchenko-Pastur (MP) distribution [9]). Further details on correlation traps are in Appendix 7.

2. Related Work

Grokking [15], delayed generalization despite saturated training accuracy, has prompted mechanism research. Initial algorithmic task studies [12, 13, 16] often linked it to weight decay (WD) favoring lower-norm solutions [6]. Other inquiries cover mechanistic interpretability [13] and competing circuits [12, 16].

Varma et al. [16] defined 'ungrokking' as generalization loss from retraining with WD on *smaller* datasets ($D < D_{crit}$), attributing this to shifting circuit efficiencies. In contrast, we find **late-stage generalization collapse** ('anti-grokking') on the *original* dataset after prolonged WD-free training ($\sim 10^7$ steps). This distinct outcome is unpredicted by [16]'s WD-reliant model.

Grokking studies now cover real-world tasks [2, 4]. Golechha et al. [2], introducing progress measures, also saw grokking without WD (hence increasing l^2 norms), like our setup. Comparing with their metrics, our drastically longer training (to 10^7 steps) reveals an 'anti-grokking' collapse they did not report, despite similar `WD=0` conditions.

We employ Heavy-Tailed Self-Regularization (HTSR) [10, 11], tracking spectral exponent α. We show α uniquely predicts initial grokking, dips, and eventual 'anti-grokking' collapse, especially without WD. Our key contribution is identifying and characterizing this anti-grokking using α for long-term stability, extending prior work often constrained by WD or shorter training.

3. Measures and Metrics

3.1. Heavy–Tailed Self-Regularization (HTSR)

From weights to spectra. For each layer weight matrix $\mathbf{W} \in \mathbb{R}^{N\times M}$, we build the un-centred *correlation* (Gram) matrix

$$\mathbf{X} \;=\; \frac{1}{N}\,\mathbf{W}^{\top}\mathbf{W} \;\in\; \mathbb{R}^{M\times M}. \tag{1}$$

Let $\{\lambda_i\}_{i=1}^{M}$ be the eigenvalues of $\mathbf{X}$. Their empirical spectral density (ESD) is the discrete measure

$$\rho_{emp}(\lambda) \;=\; \frac{1}{M}\sum_{i=1}^{M}\delta\big(\lambda-\lambda_i\big). \tag{2}$$

Marchenko-Pastur baseline. If the entries of $\mathbf{W}$ are i.i.d. with zero mean and finite variance (and a finite $2+\epsilon$ moment), then, in the limit $N, M \to \infty$ with aspect ratio $Q = N/M \geq 1$ fixed, $\rho_{emp}(\lambda)$ (almost surely) converges to the Marchenko–Pastur (MP) density [7]. Randomizing $\mathbf{W} \to \mathbf{W}^{rand}$ elementwise should then yield an ESD fitting an MP distribution (Figure 2, Right).

Heavy–Tailed Self-Regularization (HTSR) Theory. Prior work[9–11] shows that the ESD of real-world DNN layers with learned correlations almost never sits entirely within the Marchenko–Pastur bulk; instead, the right edge flares into a power law (PL) tail. Formally,

$$\rho_{emp}(\lambda) \;\sim\; \lambda^{-\alpha}, \qquad \lambda_{\min} < \lambda < \lambda_{\max}, \tag{3}$$

with the exponent α quantifying the strength of the correlations. According to the HTSR framework [11], different ranges of α correspond to the different phases of training and different levels of convergence for each layer:

- $\alpha \gtrsim 5-6$: **Random-like or Bulk-plus-Spikes** — the spectrum is close to the Gaussian baseline; little task structure is present. This upper bound is somewhat looser as it can depend on the aspect ratio Q. See Martin et. al. [9, 10] for more details.

- $2 \lesssim \alpha \lesssim 5-6$: **Weak (WHT) to Moderate Heavy (Fat) Tailed (MHT)** — correlations build up; layers are well-conditioned and typically generalise better. The lower bound of $\alpha \approx 2$ for this Fat-Tailed phase is a relatively hard cutoff.
- $\alpha = 2$ **Ideal value:** Corresponds to fully optimized layers in models. Associated with layers in models that generalize best.
- $\alpha < 2$: **Very-Heavy-Tailed (VHT)** — extremely heavy tails indicate potentially over-fitting to the training data and often precede and/or are associated with decreases in the generalization / test accuracy.

Estimating α. The exponent α is estimated by fitting the tail of ρ_{emp} to a power law (Eq. 3), following [1, 11]. Figure 2 (Left) shows an example PL fit. Specifics of the fitting procedure, including $\lambda_{\min}$ selection and the use of the `WeightWatcher` tool [8], are detailed in Appendix 8.

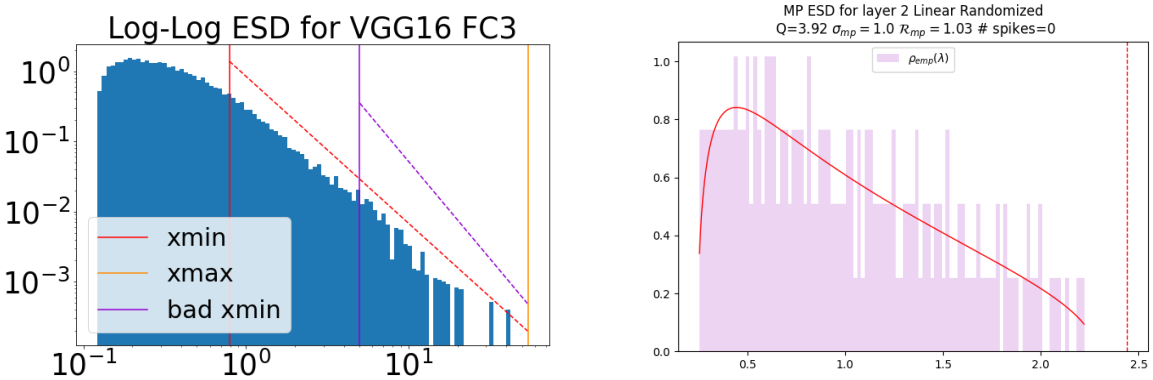

Figure 2: **Left**: Example of the ESD derived from a well-correlated $\mathbf{W}$ (blue) and the Power-Law fit to the tail (red), on a **Log-Log** plot. **Right**: Example of the ESD of $\mathbf{W}^{rand}$ (light purple) and the MP fit (red), on a **Log-Linear** plot.

The trajectory $\alpha(t)$ proves to be a highly sensitive indicator. Small fluctuations indicate layer convergence. Large drops toward $\alpha \approx 2$ coincide with grokking, while $\alpha < 2$ foreshadows anti-grokking.

3.2. Correlation Traps

Anomalous spectral features, termed **Correlation Traps** [9], can appear in the Empirical Spectral Density (ESD) of randomized layer weight matrices ($\mathbf{W}^{rand}$). These are identified as eigenvalues λ_{trap} significantly larger than the Marchenko-Pastur bulk edge (λ^{+}_{rand}). Their presence, often coinciding with $\alpha < 2$ for the original matrix $\mathbf{W}$, is indicative of overfitting and is associated with the anti-grokking phase. The `WeightWatcher` tool automatically detects these. A detailed explanation, along with examples (Figure 7) and data (Table 3), is provided in Appendix 7. Further statistical validation of their presence during anti-grokking is discussed in Appendix 5.

3.3. Comparative Metrics

We benchmarked our `HTSR`-based findings against l^2 weight norm analysis [6] and several other comparative progress measures, as proposed by Golechha et al. [2]: Activation Sparsity (A_s), Ab-

solute Weight Entropy ($H_{abs}(W)$), and Approximate Local Circuit Complexity (Λ_{LC}). Appendix 2 provides detailed definitions of these comparative metrics.

4. Results and Analysis

Experiments are performed with (`WD>0`) and without weight decay (`WD=0`). Below we focus on the dynamics under zero weight decay (`WD=0`). Appendix 1 explains the setup, and the (`WD=0.01`) case.For a detailed analysis into the comparitive measures and insight into the discovered correlation traps please refer to Appendix 3 and Appendix 7, respectively.

4.1. Layer Metrics for Tracking Grokking

HTSR layer quality metric α: The `HTSR` metric α reveals critical network dynamics (Figure 3). Initially high, α decreases as training fits data. A sharp drop towards $2 \lesssim \alpha \lesssim 5-6$ aligns with grokking's rapid test accuracy improvement ($\sim 10^4 - 10^5$ steps, Figure 1). Critically, with prolonged training ($\sim 10^6$ steps), α dips below 2 (VHT regime), notably in FC2. This drop, indicating potential over-correlation, precedes and coincides with the "anti-grokking" collapse (Figure 1). Table 4 summarizes α at key phases. The `HTSR` α uniquely identifies the grokking transition and signals subsequent instability and anti-grokking, with layer-wise analysis (Figure Figure 3) pointing to specific over-correlated layers ($\alpha < 2$) as potential sources of instability.

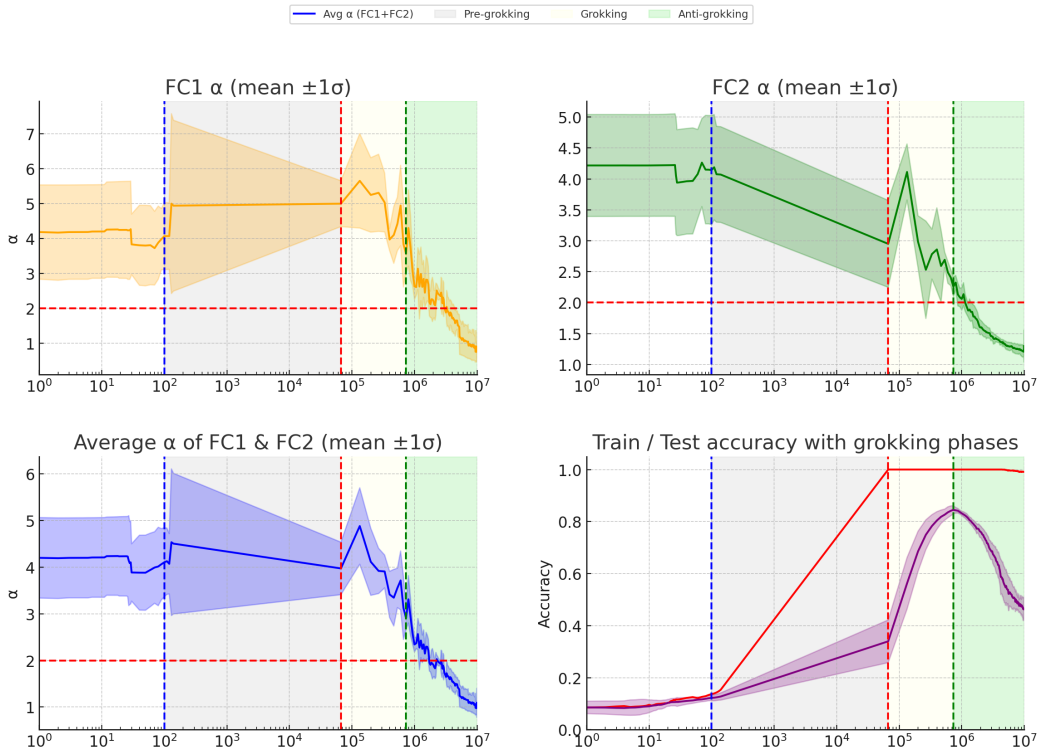

Figure 3: `HTSR` results vs. optimization steps (`WD=0`). Top: Average α across layers. Middle: α for the first fully connected layer (FC1). Bottom: α for the second fully connected layer (FC2). Note the significant dip below the critical threshold $\alpha = 2$, especially in FC2, coinciding with the "anti-grokking" performance drop seen in Fig. 1 after 1M steps.

5. Conclusion

This study demonstrates Heavy-Tailed Self Regularization's (HTSR) [10] metric α effectively tracks grokking across weight decay regimes (WD=0, WD>0), outperforming prior measures [2, 6]. Critically, under WD=0, α uniquely forewarns novel late-stage generalization collapse (anti-grokking)—a major test accuracy drop despite perfect training and large l^2 norms after extensive ($\sim 10^7$ steps) training—which other metrics missed. These findings highlight HTSR's utility for monitoring long-term generalization stability. Detailed mechanistic insights, limitations, and future work are discussed in Appendices 9, 6, and 10, respectively.

References

[1] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[2] Satvik Golechha. Progress measures for grokking on real-world tasks, 2024. URL https://arxiv.org/abs/2405.12755.

[3] Karim Huesmann, Luis Garcia Rodriguez, Lars Linsen, and Benjamin Risse. The impact of activation sparsity on overfitting in convolutional neural networks, 2021. URL https://arxiv.org/abs/2104.06153.

[4] Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why, 2024. URL https://arxiv.org/abs/2402.15555.

[5] Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, and Sanjiv Kumar. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TJ2nxciYCk-. arXiv:2210.06313.

[6] Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In Surbhi Koyejo, Sham Kakade (formerly Mohamed), Aarti Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34651–34663. Curran Associates, Inc., 2022.

[7] Vladimir A. Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 72(114)(4):507–536, 1967.

[8] Charles H. Martin. WeightWatcher: Analyze Deep Learning Models without Training or Data. https://github.com/CalculatedContent/WeightWatcher, 2018-2024. Version 0.7.5.5 used in this study. Accessed May 12, 2025.

[9] Charles H. Martin and Christopher Hinrichs. SETOL: A Semi-Empirical Theory of (Deep) Learning. https://github.com/CalculatedContent/setol_paper/blob/main/setol_draft.pdf, 2025. Preprint.

[10] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(1):165, January 2021. URL http://jmlr.org/papers/v22/20-410.html.

[11] Charles H. Martin, Tian 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*, 12:4122, jul 2021. doi: 10.1038/s41467-021-24025-8. URL https://doi.org/10.1038/s41467-021-24025-8.

[12] William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: grokking as competition of sparse and dense subnetworks, 2023. URL https://arxiv.org/abs/2303.11873.

[13] 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.

[14] Ze Peng, Lei Qi, Yinghuan Shi, and Yang Gao. Theoretical explanation of activation sparsity through flat minima and adversarial robustness, 2023. URL https://arxiv.org/abs/2309.03004.

[15] 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.

[16] Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023. URL https://arxiv.org/abs/2309.02390.

Appendices

1. Experimental Setup

We train a Multi-Layer Perceptron (MLP) on a subset of the MNIST dataset using the hyperparameters detailed in Table 1. The training subset is constructed by randomly selecting 100 samples from each of the 10 MNIST classes, ensuring a balanced dataset of 1,000 unique training points. This was run on an Nvidia Quadro P2000 and took approximately 11 hours. A considerable part of the time is due to the speed of saving the measures.

Table 1: Experimental hyperparameters used in the study.

Parameter	**Value**
Network Architecture	Fully Connected MLP
Depth	3 Linear layers (Input $\rightarrow$ Hidden1 $\rightarrow$ Hidden2 $\rightarrow$ Output)
Width	200 hidden units per hidden layer
Activation Function	ReLU (Rectified Linear Unit)
Input Layer Size	784 (Flattened MNIST image 28×28)
Output Layer Size	10 (MNIST digits 0-9)
Weight Initialization	Default PyTorch (Kaiming Uniform for weights), parameters scaled by 8.0
Bias Initialization	Default PyTorch (Uniform), then scaled by 8.0
Dataset	MNIST
Training Points	1,000 (100 per class, stratified random sampling)
Test Points	Standard MNIST test set (10,000 samples)
Batch Size	200
Loss Function	Mean Squared Error (MSE) with one-hot encoded targets
Optimizer	AdamW
Learning Rate (LR)	5×10^{-4}
Weight Decay (WD)	0.0 (for main results), 0.01 (for Appendix 4 comparison)
AdamW β_1	0.9 (PyTorch default)
AdamW β_2	0.999 (PyTorch default)
AdamW ϵ	10^{-8} (PyTorch default)
Optimization Steps	10^7
Data Type (PyTorch)	‘torch.float64‘
Random Seed	0 (for all libraries)
Software Framework	PyTorch
HTSR Tool	WeightWatcher v0.7.5.5 [8]

Note on Weight Decay: The primary results presented in this paper, particularly those demonstrating grokking followed by late-stage generalization collapse (Figure 1), were obtained with weight decay explicitly set to 0. This allows observation of the learning dynamics driven purely by the optimizer and the loss landscape while exhibiting both phenomena, whereas the other proposed measures fail to detect the grokking transition of increasing test accuracy. Runs with non-zero weight decay (e.g., WD=0.01, see Appendix 4) were also performed for comparison, showing different dynamics but confirming the general utility of HTSR.

2. Comparative Grokking Progress Metrics and Measures

Weight Norm Analysis Following observations that weight decay can influence grokking [6], we monitor the l^2 norm of the network's weights,

$$||\mathbf{W}||_2 = \sqrt{\sum_l ||\mathbf{W}_l||_F^2}, \tag{4}$$

throughout training. We specifically run experiments with weight decay disabled (`WD=0`) to isolate the effect of the optimization dynamics on the norm itself.

Activation Sparsity. For a given layer with activations $b_{i,j}$ (representing the activation of neuron j for input example i), the activation sparsity A_s is defined as:

$$A_s = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(b_{i,j} < \tau), \tag{5}$$

where T is the number of training examples, n is the number of neurons in the layer, τ is a chosen threshold, and $\mathbf{1}(\cdot)$ is the indicator function. This metric measures neuron inactivity. Prior studies have linked activation sparsity to generalization [5, 12, 14] and reported specific dynamics such as plateauing before grokking [2] or an increase preceding a rise in test loss [3].

Absolute Weight Entropy. For a weight matrix $W \in \mathbb{R}^{m\times n}$, the absolute weight entropy $H_{abs}(W)$ is given by:

$$H_{abs}(W) = -\sum_{i=1}^{m}\sum_{j=1}^{n}|w_{i,j}|\log|w_{i,j}|. \tag{6}$$

This entropy quantifies the spread of absolute weight magnitudes. Golechha et al. [2] suggested its sharp decrease signals generalization.

Approximate Local Circuit Complexity. Let $L^{(W)}(x)$ denote the output logits for input x using weights W, and let $L^{(W')}(x)$ denote the logits when 10% of the weights are set to zero (forming W'). The approximate local circuit complexity, denoted Λ_{LC}, is the summed KL divergence:

$$\Lambda_{LC} = \sum_{k=1}^{N_{data}}\sum_{j\in\mathcal{C}}\Pr\big(j|L^{(W)}(x_k)\big)\log\frac{\Pr\big(j|L^{(W)}(x_k)\big)}{\Pr\big(j|L^{(W')}(x_k)\big)}. \tag{7}$$

Here, N_{data} is the number of training examples x_k, $\mathcal{C}$ is the set of classes, and $\Pr(j|L(x))$ is the probability of class j derived from the logits $L(x)$ (e.g., via softmax). This measure captures output sensitivity to minor weight perturbations. Lower Λ_{LC} has been linked to stable, generalizable representations [2].

3. Detailed Analysis of Comparative Metrics' Performance (`WD=0`)

This section provides the detailed discussion of the comparative metrics' performance in our primary zero weight decay (`WD=0`) experiment, as referenced in Section 4. The evolution of these metrics is shown in Figure 4.

In our primary `WD=0` experiments, Activation Sparsity (A_s) generally increases throughout training (Figure 4), seemingly tracking the pre-grokking and grokking phases. However, it fails

as a negative control in the anti-grokking phase because it continues to increase in the same manner as in pre-grokking. Prior studies have linked activation sparsity to generalization [5, 12, 14] and reported specific dynamics such as plateauing before grokking [2] or an increase preceding a rise in test loss [3]. Specifically, we observe a subtle inflection or dip in A_s coinciding with the point of maximum test accuracy before a slight increase. While this feature appears to mark a shift around peak test accuracy, its specific predictive utility for subsequent generalization dynamics is questionable. In other words, without knowing a proper sparsity cutoff, it is impossible to determine if increasing A_s corresponds to pre-grokking or anti-grokking. In contrast, because the HTSR $\alpha = 2$ is a theoretically established universal cutoff, one can distinguish between the two phases correctly.

Additionally, in our WD=0.01 control experiment (Appendix 4), a similar inflection in A_s occurs where test accuracy, after a slight initial decrease from its peak, subsequently plateaus rather than undergoing a catastrophic collapse as seen in the WD=0 case. Therefore, observing this dip in A_s alone does not allow one to distinguish whether test accuracy will catastrophically decline or stabilize. This suggests it primarily indicates that some form of transitional change has occurred around the point of maximum generalization, rather than predicting the specific nature of the subsequent trajectory.

Our findings indicate limitations in the other two comparative metrics for tracking the anti-grokking phase. Absolute Weight Entropy ($H_{\text{abs}}(\mathbf{W})$), despite its suggested link to generalization [2], also decreases sharply during the collapse, thus not reliably distinguishing this anti-grokking phase from the initial grokking phase where it also decreases. Similarly, Approximate Local Circuit Complexity (Λ_{LC}) [2] remains low throughout the collapse, failing to reflect the performance degradation. We also confirm, consistent with [2], that grokking occurs robustly even with increasing weight norms and no weight decay.

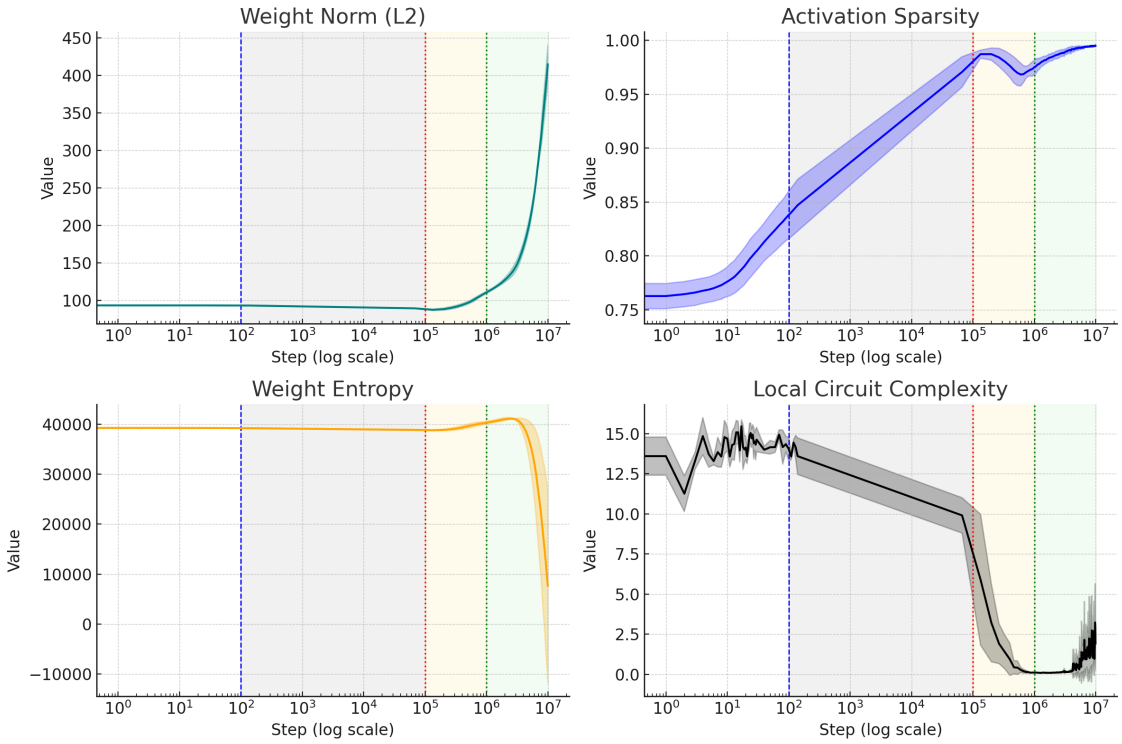

Figure 4: Alternative progress measures (Golechha [2]) vs. optimization steps (WD=0). Top: Activation Sparsity. Middle: Absolute Weight Entropy. Bottom: Approximate Local Circuit Complexity. While these metrics show changes during the initial training and grokking phases (Activation Sparsity for example), they do not exhibit clear signals predicting the magnitude of the late-stage "anti-grokking" performance dip observed after 10^6 steps.

In contrast, the other proposed comparative metrics capture initial training and grokking but fail to predict the late-stage generalization collapse, This is evident in Figure 4, Appendix 3). While showing trends during initial learning, their trajectories often lack distinct features for the "anti-grokking" drop (e.g., circuit complexity remains flat). Appendix 3 provides a detailed analysis of their performance.

4. Experiment with Weight Decay (WD=0.01)

To further understand the influence of weight decay on the observed generalization dynamics and the behavior of our tracked metrics, we conducted an experiment identical to our main study (WD=0) but with a small amount of weight decay (WD=0.01) applied. The training curves and metric evolutions for this WD=0.01 experiment are presented in Figures 5, and 6.

A key characteristic of training with weight decay is the tendency for the l^2 norm of the weights to decrease over time, or stabilize at a lower value, which is observed in this experiment (Figure 6). This contrasts with the continuously increasing l^2 weight norm seen in our primary WD=0 experiments.

In this WD=0.01 regime, the network still achieves a high level of test accuracy. Notably, after the initial grokking phase, the test accuracy slightly decreases and then enters a prolonged plateau,

maintaining near peak performance for a significant number of optimization steps (Figure 5, top left panel showing accuracies). Correspondingly, the average heavy-tail exponent, α, also exhibits the decrease and a distinct plateau around the critical value of $\alpha \approx 2$ during this period (Figure 5, top left panel for α).

The other progress measures considered—Activation Sparsity and Approximate Local Circuit Complexity—also tend to plateau or stabilize during this phase of peak test performance in the WD=0.01 setting (Figure 6). This contrasts with the WD=0 scenario where, despite eventual grokking, the system does not find such a stable long-term plateau and instead proceeds towards a late-stage generalization collapse. The observation that α (and other metrics) plateau in conjunction with peak, stable test accuracy under traditional weight decay settings aligns with some existing understanding of well-regularized training.

While HTSR and the α exponent provide valuable insights in both regimes, its unique capability to signal impending collapse in the absence of weight decay underscores its importance for understanding layer dynamics under various scenarios.

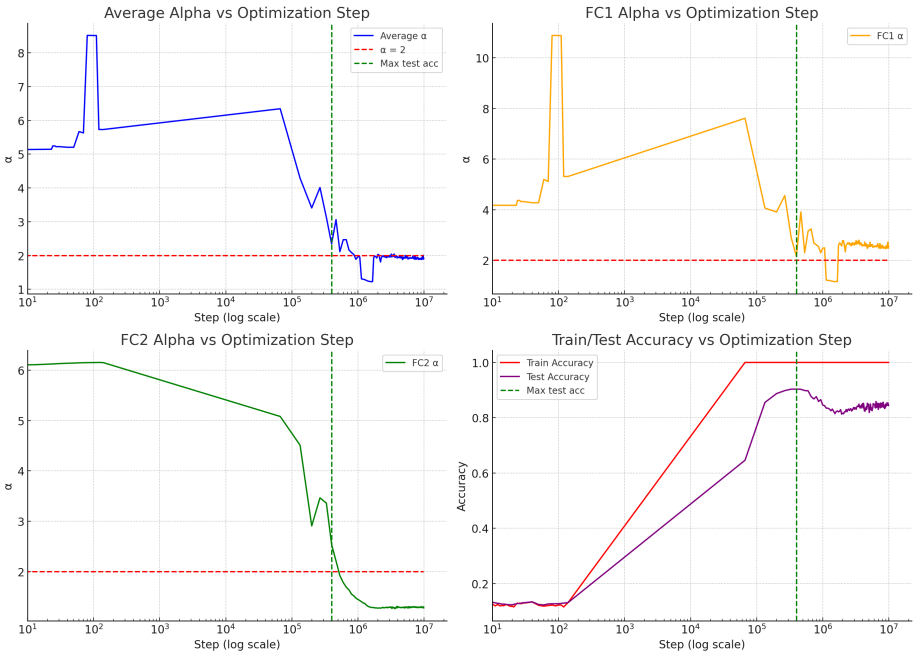

Figure 5: HTSR α exponent evolution for the MLP trained with WD=0.01. (Top-left panel also includes test/train accuracy for context).

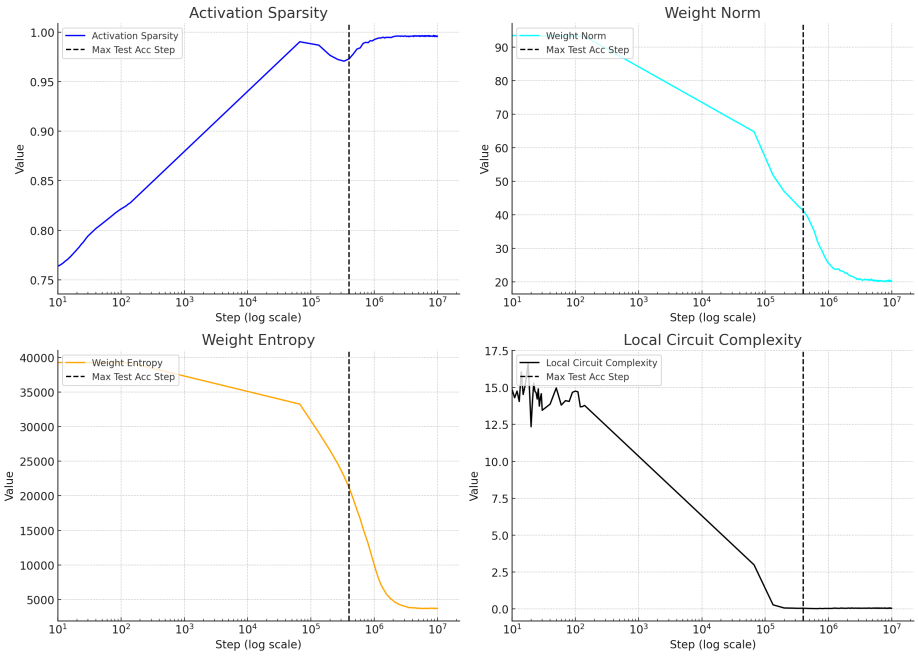

Figure 6: Progress measures (Activation Sparsity, Weight Entropy, Circuit Complexity) and l^2 Weight Norm for the MLP trained with `WD=0.01`.

5. Statistical Analysis and Validation of Correlation Traps

Here, to further validate the presence of correlation traps for the zero weight decay `WD=0` experiment , we report the results of statistical tests designed to determine if the randomized ESD of the $\mathbf{W}^{rand}$ fits an MP distribution or not. Briefly we fit the ESD to a MP distribution and report the fitted variance σ_{mp}, the Kolmogorov-Smirnov (KS) statistic of the fit, and the p-value for the MP fit as the null model. We also report the number of correlation traps, as determined using the open-source `WeightWatcher` tool[8]. Results for layer FC1 are presented in Table 2. Results for FC2 are similar (not shown). Additional details are provided in the supplementary material.

Table 2: **Statistical validation of correlation traps.** Selected results for layer FC1 at different training stages for zero weight decay (`WD=0`) experiment. MP Variance (σ_{MP}) Kolmogorov-Smirnov (KS) test statistic, p-value for MP fit, and number of detected correlation traps. Pre-grokking $\sim 10^5$ steps, Grokking $\sim 10^6$ steps, and Anti-grokking $\sim 10^7$ steps.

Model State	**MP variance** (σ_{mp})	**KS Statistic**	**p-value**	**# Traps**
Pre-Grokking	≈ 1.002	0.0120	≈ 1.0	0
Grokking (Max Test Accuracy)	≈ 0.999	0.0212	≈ 1.0	0
Anti-Grokking (Collapse)	≈ 0.949	0.3044	1.877×10^{-5}	9

Initial Layer State (Pre-Grokking `WD=0`): Immediately after initialization, the network weights are expected to be largely random, and their ESD should conform well to the MP distribution. Figure 2 (Right) shows an MP fit to an ESD from a representative layer $\mathbf{W}^{rand}$ of the newly initialized

model. A KS test comparing this empirical ESD to the fitted MP distribution (using $\sigma_{mp} \approx 1.0024$ as estimated by `WeightWatcher`) yielded a KS statistic of 0.0120 and a p-value ≈ 1.0. This high p-value indicates this ESD is statistically consistent with the MP distribution, as expected.

Best Layer State (Grokking phase `WD=0`): As the network learns and reaches its maximum test accuracy, significant structure develops in the elements of the weight matrices $W_{i,j}$. This can be seen by randomizing the layer weight matrix elementwise, $\mathbf{W} \rightarrow \mathbf{W}^{rand}$, and plotting ESD, and looking for deviations from the theoretical MP distribution. For our model at peak test accuracy, the KS test against a fitted MP model ($\sigma_{mp} \approx 0.999$) resulted in a KS statistic of 0.0212 and a p-value ≈ 1. Again, this indicates the bulk of $\mathbf{W}^{rand}$ is MP-like, even if $\mathbf{W}$ itself is highly structured. Traps are not yet dominant.

Final Layer State (Anti-Grokking phase `WD=0`): In the late-stage of training, as the model undergoes generalization collapse and enters an over-correlated state (characterized by $\alpha < 2$), the ESD of $\mathbf{W}^{rand}$ structure continues to reflect a non-random configuration, now with prominent traps. The KS test for the final model against an MP fit (with an estimated $\sigma_{mp} \approx 0.9492$ for FC1 as in Table 2) yielded a KS statistic of 0.3044 and a p-value of 1.877×10^{-5} (see Figure 7 Right for an example of a layer in this state). This result further confirms that the network's $\mathbf{W}^{rand}$ structure is significantly different from a random matrix baseline, consistent with the highly correlated state indicated by our `HTSR` analysis and the presence of traps.

These quantitative comparisons demonstrate a transition from an initially random-like state (consistent with MPD) to progressively more structured states. The inability of the MP distribution to describe $\mathbf{W}^{rand}$ in the anti-grokking phase (due to traps) highlights the pathological changes, while `HTSR` theory on $\mathbf{W}$ itself (the PL exponent α) characterizes the correlation structure leading to these phenomena.

6. Detailed Discussion on Limitations

Our study, while providing insights into generalization dynamics via Heavy-Tailed Self-Regularization (`HTSR`), has limitations that define important avenues for future research. The empirical findings are primarily derived from a specific three-layer MLP architecture trained on an MNIST subset. Consequently, the generalizability of the observed α trajectories and their specific predictive power for phenomena like grokking and late-stage generalization collapse warrants further validation across a wider range of model architectures (e.g., CNNs, Transformers), datasets, tasks, and diverse training configurations, including different optimizers and hyperparameter settings.

Furthermore, `HTSR` is an empirically-grounded, phenomenological framework, supported theoretically with a novel application of Random Matrix Theory (RMT). While its correlations between the heavy-tailed PL exponent α and network generalization states are compelling, the interpretation requires careful consideration of context. For instance, while well-generalized models often exhibit α values within the range (e.g., $2 \leq \alpha \leq 6$), and $\alpha \approx 2$ is frequently associated with optimal performance or critical transitions, this is not a strictly bidirectional implication. It is conceivable that layers or models might exhibit α values near or even below 2 (typically indicating over-correlation) yet display suboptimal generalization. Other very-well trained models may have layers fairly large alphas. This is not yet fully understood. This highlights that while α provides strong correlational insights into learning phases and stability, the precise mapping of specific α values to absolute performance levels can be context-dependent and is an area for ongoing refinement of the theory (see [9]). Our work contributes observations within specific phenomena, acknowledging that the broader applicability and predictive nuances of the `HTSR` theory will benefit from continued exploration.

These limitations underscore the importance of ongoing empirical and theoretical work to further refine, validate, and extend the understanding of HTSR theory in deep learning.

7. Details on Correlation Traps

Correlation Traps, as introduced by Martin et al. [9], refer to anomalously large eigenvalues observed in the Empirical Spectral Density (ESD) of a layer's weight matrix $\mathbf{W}$ after its elements have been randomized to form $\mathbf{W}^{rand}$. Specifically, if $\{\lambda_i^{rand}\}$ are the eigenvalues of the correlation matrix derived from $\mathbf{W}^{rand}$, and λ_{rand}^{+} is the right edge of the bulk of the Marchenko-Pastur (MP) distribution that best fits this ESD, a correlation trap λ_{trap} is an eigenvalue such that $\lambda_{trap} \gg \lambda_{rand}^{+}$.

The presence of these traps indicates that even after element-wise randomization (which should destroy local correlations if the original matrix elements were drawn from a simple distribution with finite moments), some strong, non-trivial global correlation structure persists or emerges, manifesting as isolated large eigenvalues detached from the main MP bulk. In the context of our work, these traps become prominent during the anti-grokking phase, particularly when the HTSR exponent α of the original (non-randomized) weight matrix $\mathbf{W}$ falls below 2. This suggests that the VHT regime ($\alpha < 2$) is associated with overfitting that creates such deeply embedded correlational structures that they survive (or are revealed by) randomization as these large outlying eigenvalues.

The `WeightWatcher` tool [8] is used to detect these correlation traps automatically. It randomizes $\mathbf{W}$, fits an MP distribution to the bulk of the ESD of $\mathbf{W}^{rand}$ to estimate σ_{MP}^2 and thus λ_{rand}^{+}, and then identifies eigenvalues significantly beyond this edge (considering Tracy-Widom fluctuations). Figure 7 provides visual examples of ESDs of $\mathbf{W}^{rand}$ where correlation traps are present in layers of models undergoing generalization collapse. The statistical significance of these deviations from a pure MP distribution can be quantified using tests like the Kolmogorov-Smirnov (KS) test, as detailed in Appendix 5.

Table 3: **Average number of detected correlation traps** in layers FC1 and FC2 at the right edge of of the three grokking phases. Results shown for experiments with zero weight decay (`WD=0`) and with weight decay (`WD>0`, see Appendix 4 for context).

Model, Layer	Pre-grokking	Grokking (Max Test Acc.)	Anti-grokking (Collapse)
$WD = 0$, FC1	0 ± 0	1 ± 0	7.5 ± 5.6
$WD = 0$, FC2	0 ± 0	1 ± 0	1 ± 0
$WD > 0$, FC1	0	0	2.0 ± 0.0
$WD > 0$, FC2	0	0	1.0 ± 0.0

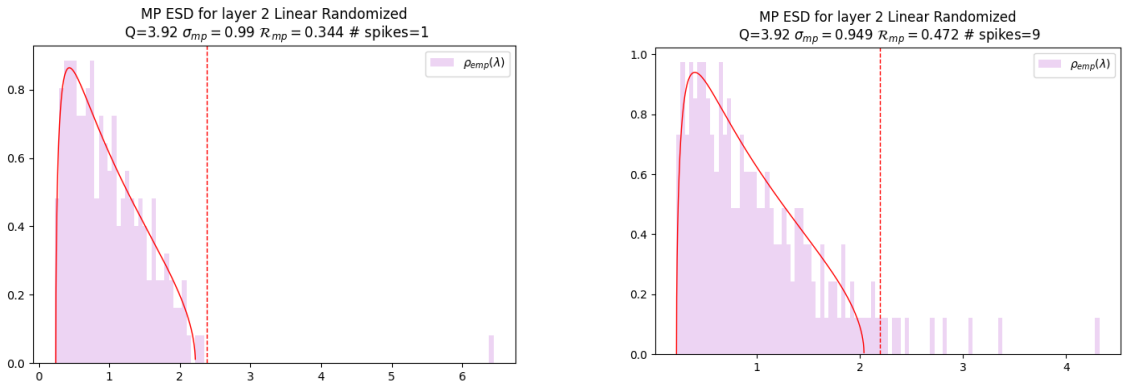

Figure 7: **Examples of Correlation Traps**. ESDs of ($\mathbf{W}^{rand}$) (light purple) of Layer 2 for the randomized weight matrix $\mathbf{W}^{rand}$ for different models, compared to an MP fit (red). Correlation traps λ_{trap} are depicted a small spikes to the right of the MP fit. (x-axis is log scale) **Left: Right Before Collapse** (i.e. at more than $\sim 10^6$ steps) ($\sigma_{mp} \approx 0.9879$). The KS test (P-value $\approx 4 \times 10^{-13}$) indicates a strong deviation from the MP model. A single, prominent correlation trap appears at $\lambda_{trap} \approx 10^{6.5}$. **Right: Final Generalization Collapse**. The KS test (P-value $\approx 1.877 \times 10^{-5}$) indicates a strong deviation from the MP model. Multiple correlation traps are observed, $\lambda_{trap} \in [10^{2.x}, 10^{6.5}]$.

Summary of Findings: Correlation Traps and Anti-Grokking (from Main Results) To better understand anti-grokking, we look for correlation traps in layer weight matrices. Data on the average number of detected correlation traps (Table 3) show that for the `WD=0` experiment, correlation traps appear in layers FC1 and FC2 only during the anti-grokking phase, coinciding with $\alpha < 2$ for these layers. This suggests overfitting reflected as specific spectral artifacts. Statistical validation details for these correlation traps are in Appendix 5.

8. Details of α Estimation

The power law (PL) fit to the tail of the empirical spectral density (ESD), ρ_{emp}, is performed using a maximum likelihood estimator (MLE), as described by Clauset et al. [1], following the approach in [11]. A crucial step in this process is determining the start of the PL tail, denoted $\lambda_{\min}$. This value is chosen automatically by a procedure that aims to minimize the Kolmogorov-Smirnov (KS) distance between the empirical data and the fitted power-law distribution over a range of possible $\lambda_{\min}$ values. This ensures that the PL model is fit to the region where it best describes the data.

All calculations related to `HTSR` analysis, including the singular value decomposition (SVD) to obtain singular values (whose squares are the eigenvalues λ_i), the PL fits (including the automated selection of $\lambda_{\min}$ and $\lambda_{\max}$ using KS goodness-of-fit tests), and the detection of correlation traps (discussed in Appendix 7), are performed using the open-source `WeightWatcher` tool, version 0.7.5.5 [8]. The careful and automated selection of $\lambda_{\min}$ is critical for an accurate estimation of the tail exponent α, as a poor choice can significantly bias the result.

9. Mechanistic Insights from HTSR Analysis

Our HTSR analysis suggests a mechanistic view of grokking's phases: pre-grokking involves partial layer convergence (some layers $\alpha \approx 4$, others $\alpha \approx 5$); true grokking occurs when all key layers optimize ($\alpha \approx 2$); and anti-grokking is marked by one or more layers over-correlating ($\alpha < 2$), often exhibiting correlation traps (detailed in Appendix 7). These traps, anomalous rank-one (or greater) perturbations in $\mathbf{W}$, cause a large mean-shift in underlying weight element distributions ($\mathbb{E}[W_{ij}] \rightarrow$ *large*), pushing the ESD into the VHT phase. Such atypical weight distributions are hypothesized to impair generalization. These results align with the established HTSR theory and the more recently developed SETOL theory [9].

Table 4: **Layer-wise and average HTSR α exponents.** At the right edge of each grokking phase: Pre-grokking $\sim 10^5$ steps, Grokking 10^6 steps, and Anti-grokking 10^7 steps. For the zero-weight-decay (WD=0) experiment; values are taken from Fig. 3. Various seeds are used and variability in initialization, optimizer trajectory may occur.

Layer, Metric	Pre-grokking	Grokking (Max Test Acc.)	Anti-grokking (Collapse)
FC1 α	5.0 ± 0.7	3.6 ± 0.5	0.9 ± 0.4
FC2 α	2.9 ± 0.7	2.3 ± 0.2	1.3 ± 0.3
average α	4.0 ± 0.6	2.9 ± 0.2	1.1 ± 0.3

10. Future Directions

This work opens several avenues for future research building upon the insights from Heavy-Tailed Self-Regularization (HTSR) theory:

- **Broader Empirical Validation:** The generalization of our findings regarding α's predictive power for grokking and anti-grokking should be validated across a more diverse range of model architectures (e.g., Convolutional Neural Networks, Transformers), datasets beyond MNIST subsets, varied tasks, and a wider spectrum of hyperparameter configurations and optimizers.
- **Refinement of HTSR Theory and its Application:**
  - Further investigation is needed into the precise mapping of α values to absolute performance levels, as this relationship can be context-dependent. This includes exploring cases where α might be low without suboptimal generalization, or high in well-performing models.
  - Continued empirical and theoretical work is crucial to refine, validate, and extend the understanding of HTSR theory, particularly its nuances in predicting generalization dynamics. See, in particular, Martin et. al.[9]
- **Practical Applications of HTSR Insights:**
  - The development of α-guided adaptive training strategies is a promising direction. This could involve dynamically adjusting learning rates, implementing novel early stopping criteria based on α trajectories, or performing layer-specific interventions during training to maintain optimal α ranges.

  - Designing and testing differentiable regularizers or loss terms based on the α exponent could enable training dynamics that explicitly encourage convergence towards values associated with stable generalization (e.g., $\alpha \approx 2$).

- **Deeper Mechanistic Understanding:**
  - A more in-depth study of the formation, characteristics, and precise impact of "correlation traps" is warranted, particularly their relationship with specific overfitting mechanisms and the VHT regime ($\alpha < 2$).
  - Further exploration of how layer-specific changes in α contribute to, or indicate, the onset of instability and how these dynamics propagate or originate within different network architectures would be valuable.

Addressing these areas will not only solidify the understanding of phenomena like grokking and anti-grokking but also enhance the practical utility of HTSR as a tool for developing more robust and reliable deep learning models.