

Empirical Verification of Topological Phase Transitions During Grokking via Heavy-Tailed Spectral Analysis

1. Introduction

The phenomenon of "grokking" - where neural networks experience a delayed, sudden transition from memorization to generalization - suggests a fundamental topological reorganization of the network's weight matrices. Before exploring novel objective functions designed to induce or inhibit this phase change, it is necessary to establish a verified empirical baseline. This report presents a successful replication of the WeightWatcher (WW) framework, analyzing the spectral properties and circuit complexity of a depth-3, width-200 ReLU MLP trained on MNIST. By tracking multiple metrics across 10^5 optimization steps, we isolate the precise mathematical inflection point where the network abandons dense memorization for a sparse, generalizing circuit.

2. Multi-Metric Synthesis of the Phase Transition

Our baseline replication captures a distinct phase transition occurring roughly between 10^4 and 10^5 optimization steps. The synthesis of accuracy, weight norm, and spectral metrics reveals a multi-stage topological collapse.

- **Accuracy and Norm Divergence:** The network rapidly achieves 1.0 training accuracy by step 10^3 , entering the memorization basin. However, test accuracy continues a slow, sustained climb. Crucially, this late-stage generalization coincides with a dramatic exponential spike in the overall weight norm beginning at 2×10^4 steps. This indicates the network is traversing a flat region of the loss landscape to find a fundamentally different, generalizing basin.
- **Layer-wise Spectral Signature (α):** Analyzing the empirical spectral density (ESD) of the weight matrices provides deeper insight. The layer-wise WeightWatcher power-law exponent (α) exhibits high initial volatility but demonstrates a pronounced downward trend, particularly in Layer 0. A decreasing α (approaching the $2 < \alpha < 4$ range) signifies the weight matrices are developing heavier tails. This spectral signature strongly correlates with improved generalization capacity and the formation of correlated, structural sub-components rather than random, dense noise.
- **Topological Collapse and Circuit Sparsity:** The transition is most visible in the network's complexity metrics. Around 10^4 steps, coinciding with the test accuracy improvement, we observe a simultaneous explosion in Weights Entropy and a near-complete collapse in Greedy Circuit Complexity (dropping from ~ 5.0 to near 0.0). Paired with steadily increasing

Activation Sparsity, these metrics mathematically confirm that the network has pruned its dense memorization pathways, snapping into a highly efficient, modular algorithmic circuit.

3. Proposed Theoretical Framework: Objective Function Dependency

Having established the empirical baseline for the topological phase transition, we hypothesize that the emergence of this sparse modularity is not inherent to the architecture, but strictly governed by the local geometry induced by the objective function.

The standard grokking setup relies on the tension between an exponential-tail loss and an explicit complexity penalty:

$$L = L_{CE}(y, \hat{y}) + \lambda ||W||_2^2$$

We propose that altering this topological pressure will predictably shift or eliminate the phase transition. Future work will ablate the objective function under the following testable hypotheses:

- **Hypothesis 1: Removal of the L_2 Penalty.** Without weight decay, the network lacks the gradient pressure to escape the memorization basin. Predicted Metric Response: The Greedy Circuit Complexity will remain high, the WW alpha will stabilize at a high value (indicating random, Gaussian-like weight matrices lacking heavy tails), and the delayed spike in weight norm will not occur.
- **Hypothesis 2: Substitution with Mean Squared Error (MSE).** Unlike Cross-Entropy, MSE does not incentivize pushing logits to infinity for confident, correct classifications. Predicted Metric Response: The loss landscape geometry is fundamentally altered. We predict the network will not experience the characteristic "war of attrition" that forces the sudden structural snap. Sparsity will remain low, and the delayed test accuracy climb will be significantly suppressed.
- **Hypothesis 3: Label Noise Injection.** Introducing controlled stochasticity to the targets disrupts the smooth memorization basin. Predicted Metric Response: We expect the phase transition to be either significantly delayed (requiring more optimization steps to find the underlying algorithmic signal) or entirely prevented if the noise threshold exceeds the regularizer's capacity to force a sparse circuit.

4. Limitations & Future Work

Due to computational and time constraints during the hackathon sprint, empirical validation of the objective function ablations detailed in Section 3 remains future work. The current analysis is limited to identifying the baseline spectral and complexity signatures of the phase transition. Furthermore, while the WeightWatcher framework provides excellent post-hoc analysis of the weight matrix topology, it remains a proxy measurement. Future experiments should incorporate

direct, low-rank matrix factorization at the specific step thresholds identified (10^4) to visualize the exact structural circuits as they form under varying loss constraints.