Grokking Under Differential Privacy

Santhosh Janakiraman (sj1230) & Harsha Rajendra (hr458)

CS 596: Learning Theory Under Information Constraints Project Proposal (Fall 2025)

1 Background and Setup

Understanding how neural networks learn to generalize is still a major challenge for learning theory. The phenomenon of delayed generalization, or **grokking**, basically comes down to a significant challenge in optimization stability within huge, overparameterized models. We are proposing a focused test to investigate why learning sometimes fails completely, using differential privacy to intentionally restrict information available to the network. We need to determine if privacy helps stabilize learning, or if it destroys the gradient signal needed for generalization. This project tests whether privacy constraints affect the transition from memorization to generalization.

Grokking (Delayed Generalization)

Grokking, first described by Power et al. (2022), challenges **traditional early stopping criteria**. These criteria are simple rules for when to stop training. The standard approach is to stop when validation error starts rising, indicating overfitting. Grokking shows that on specific tasks, networks first achieve near-perfect training accuracy (memorization) but exhibit performance on test data consistent with **random chance**, indicating a total lack of generalization. Only after subsequent, prolonged training does the test accuracy sharply increase (generalization). This outcome proves that a simpler, superior generalizing solution exists, but the optimizer requires extensive time or strong bias to finally locate it.

Numerical Instability and Softmax Collapse

The reason generalization is delayed and often requires intervention is tied to numerical stability in the final output layer. As training progresses on cross-entropy loss, the network enters a state of **Naïve Loss Minimization (NLM)**. NLM is the pathological optimization regime where the network continually scales up the magnitude of its logits to infinitesimally reduce the cross-entropy loss without altering the model's prediction. This uncontrolled logit growth runs into floating-point limitations, causing smaller, useful gradient terms to be absorbed by larger ones within the Softmax function. This results in **Softmax Collapse (SC)**, where the essential generalization gradient signal vanishes, freezing the model in its memorization state. This mechanical failure is why **traditional regularization** (explicit penalties like weight decay added to the loss function) is required: it stops NLM, preventing SC.

The Differential Privacy Constraint

Differential Privacy (DP) provides an explicit, quantifiable constraint on the information learned from any single data point. The Differential Privacy-Stochastic Gradient Descent **(DP-SGD)** optimization algorithm enforces this using two core mechanisms:

1. **Gradient Clipping:** This bounds the ℓ_2 **norm** (the Euclidean magnitude) of per-sample gradients to a positive constant, C. Clipping ensures that no single data point can cause an excessively large update.

2. **Noise Injection:** This adds Gaussian noise, specified by variance σ^2, scaled proportionally to the clipping bound C at each step. The overall privacy guarantee is measured by the parameter ϵ.

We test whether DP can enable grokking without weight decay, using the privacy budget (ϵ) as our control parameter.

2 Problem Statement and Goal

Our project is designed to test the consequences of introducing Differential Privacy (DP) into the already numerically unstable learning dynamic of grokking. The objective is to establish the precise effect of the DP constraint (ϵ) on the **initiation** and **latency** of the grokking phase transition. Our primary goal is to determine the critical trade-off: does the Gradient Clipping mechanism implicitly solve the numerical instability problem identified by Prieto et al. (2025), and if so, characterize the critical privacy budget (ϵ_0) required to maintain stable generalization. We first want to confirm that generalization still happens under DP. In other words, does strong privacy (low ϵ) destroy the gradient signal completely and prevent any learning. We will then measure the latency (Time to Grok) to see if clipping suppresses NLM and prevents Softmax Collapse, which might actually speed up generalization compared to no regularization. Finally, we hypothesize there might be an optimal privacy level where generalization happens fastest, somewhere between very high privacy (which prevents learning) and very low privacy (which might not provide enough regularization).

3 Best Known Results

Power et al. (2022): Existence and Dependency

Question Answered: Can networks generalize long after overfitting, and what conditions enable it?
Result: Grokking exists, demonstrated robustly on algorithmic tasks such as modular multiplication ($a \times b \pmod{97}$). Crucially, Power et al. (2022) found that a small but non-zero amount of regularization (specifically weight decay) was necessary to overcome the long delay and successfully generalize. This established the prerequisite for all subsequent research.

Prieto et al. (2025): The Mechanism of Delay

Hypothesis Tested: Is the required regularization simply an architectural preference, or does it correct an intrinsic training defect?
Result: Prieto et al. (2025) formalized that regularization's success stems from curing a numerical defect. They proved that regularization is effective because it suppresses **Naïve Loss Minimization (NLM)**, thereby preventing the model's descent into **Softmax Collapse (SC)**. This is the foundation of our work: we replace weight decay with DP-SGD to test if an information constraint can perform the same mechanical stabilization.

4 Proposed Approach

Our methodology centers on isolating and quantifying the mechanical competition between DP's two components: noise and clipping.

1. **Dataset & Architecture:** We will utilize the modular multiplication task ($a \times b \pmod{97}$) with a standard 2-layer Transformer/MLP architecture. The data will be generated and split 50/50 for training and testing.

2. **Intervention:** We replace conventional regularization with a comprehensive sweep over the DP parameter: $\epsilon \in \{0.1, 0.5, 1.0, 5.0, 10.0\}$. This range spans from high stability/high noise to low stability/low noise. We will utilize the `Opacus` library for DP-SGD implementation.

3. **Analysis & Deliverable:** Our primary deliverable is a **Phase Diagram** (Heatmap) charting the relationship between the privacy budget ϵ and the time required for generalization. This diagram will empirically confirm the cost of privacy (the delay) while testing the hypothesis that clipping provides a numerical benefit by avoiding the pathological condition of Softmax Collapse.

References

[1] Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets.* arXiv preprint arXiv:2201.02177.

[2] Prieto, L., Barsbey, M., Mediano, P. A. M., & Birdal, T. (2025). *Grokking at the Edge of Numerical Stability.* Published as a conference paper at ICLR 2025.

[3] Notsawo, P. J. T., Dumas, G., & Rabusseau, G. (2025). *Grokking Beyond the Euclidean Norm of Model Parameters.* Proceedings of the 42nd International Conference on Machine Learning.

[4] Tian, Y. (2025). *Provable Scaling Laws of Feature Emergence from Learning Dynamics of Grokking.* arXiv preprint arXiv:2509.21519.