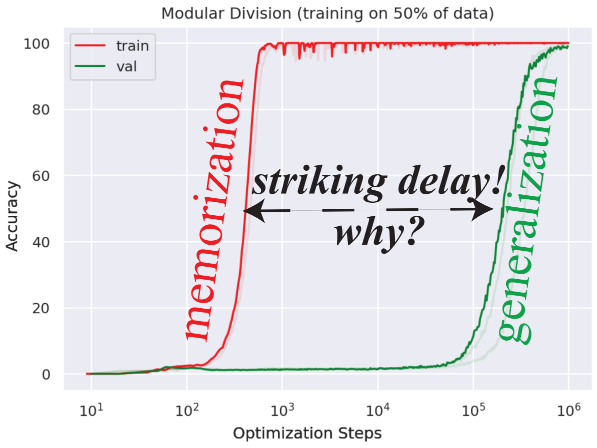
Competing Theories of Grokking

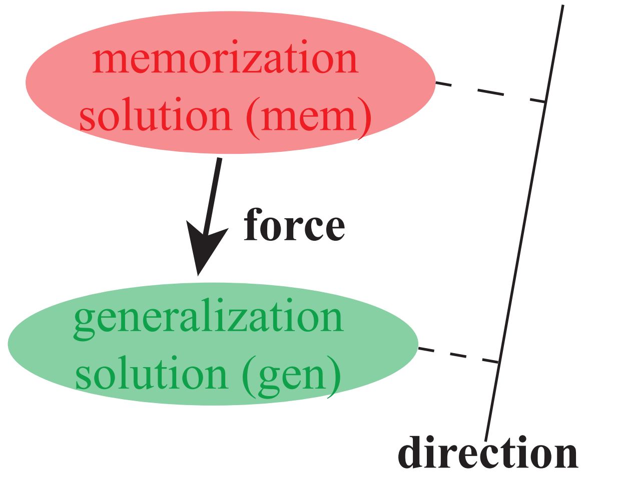
Ziming Liu, MIT & IAIFI, zmliu@mit.edu (Date: Aug 18, 2023)

Grokking Phenomenology [1]



Grokking Mechanism

All theories boil down to this mental picture:



The mental picture is:

F1: There are generlization (gen) and memorization (mem) solutions.

F2: Gen and mem solutions are separated in parameter space, along some **direction**.

F3: There is a **force** that drives the model from mem to gen.

All theories share the same mental picture above, but vary in details:

Q1: Why do generalization solutions exist at all?

All theories agree that represetation is key [1][2][3][4][5].

*Q2: What is the **direction** that separates gen and mem solutions?*

- Neuron activity [3]
- Sparsity [7]
- Fourier gap [8]
- weight norm of model parameters [6]
- time scales of pattern formation [9]
- last layer norm [10]

*Q3: What is the **force** that drives the model from mem to gen?*

- weight decay [6]
- Gradual process by optimization [3][7][8][9]
- Instability from Adam optimization [10]

Bonus Q: Universality and predictability

- Grokking can be avoided [2][6].
- Grokking can be predicted [13].
- Grokking can happen for non-algorithmic datasets [6].
- Grokking can occur for (analytically solvable) toy models [11] [12].

Reference

[1] Power et al. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets.

[2] Liu et al. Towards Understanding Grokking: An Effective Theory of Representation Learning.

[3] Nanda et al. Progressive Measures for Grokking via Mechanistic Interpretability.

[4] Chughtai et al. A Toy Model of Universality: Reverse Engineering how Networks Learn Group Operations.

[5] Pearce et al. (Google, blogpost): Do Machine Learning Models Memorize or Generalize?

[6] Liu et al. Omnigrok: Grokking Beyond Algorithmic Data.

[7] Merrill et al. A Tale of Two Circuits: Grokking as Competition of Sparse and Dense Subnetworks.

[8] Barak et al. Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit.

[9] Davies et al. Unifying Grokking and Double Descent.

[10] Thilak et al. The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the Grokking Phenomenon.

[11] Andrey Gromov. Grokking modular arithmetic.

[12] Zunkovic and Ilievski. Grokking phase transitions in learning local rules with gradient descent.

[13] Natsawo et al. Predicting Grokking Long Before it Happens: A look into the loss landscape of models which grok.