Grokking vs Learning:
Features, Encodings, & Trajectories

Edward Hirst

Queen Mary, University of London
e.hirst@qmul.ac.uk

arXiv: 2502.01739

David Berman, Barry Bradlyn, Giuseppe De Tomasi,
Jacopo Gliozzi, Dmitry Manning-Coe, Alexander G. Stapleton

RTG, Inauguration Workshop
Universität Würzburg

18th March 2025

Personal Introduction

AI in `hep-th`

- Generate objects from theoretical physics in bulk, analyse with ML architectures, train to approximately solve problems.
- Objects related to Quiver Gauge Theories:
  Cluster Algebras, Amoebae, Brane Webs, Dessins d'Enfants

AI in Geometry

- Predict topological properties of compactification spaces to speed up searches.
- These spaces, and other geometric objects studied include:
  Calabi-Yau manifolds, G_2-manifolds, Polytopes, Hilbert Series
- Notable work: AI-approximation of Einstein Metrics on Spheres (`2502.13043`).

Physics for AI

- Recent works look at physics-inspired intuition for explainable AI.
- Largely related to Fisher-information, and some other work on weight matrix permutation symmetry breaking.

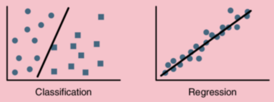
Overview

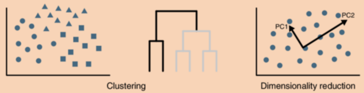
1. Machine Learning
   - Subfields
   - Neural Networks
   - Grokking
   - Fisher Information Metrics
2. Results
   - ML Problems
   - Features & Encodings
   - Trajectories
3. Summary

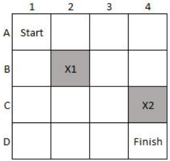
Machine Learning: Subfields

ML Subfields

1) Supervised learning: Function fitting $\{\text{inputs}\} \longmapsto \{\text{outputs}\}$
...classification (output a finite set) or regression (output continuous).
2) Unsupervised learning: Data analysis
...clustering (do datapoints self-classify) or dim-reduction (compression).
3) Reinforcement learning: Optimal solution search
...from a state space of solutions, and an action space of perturbations, learn sequence of actions to make solution optimal.
- Here we study a supervised architecture on classification problems.





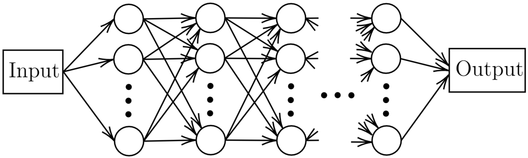


Machine Learning: Neural Networks

NN Structure

General highly-parameterised non-linear function.

$$f_{NN} := W_{i_L, i_{L-1}} \cdot \alpha(\; ... \; \alpha(W_{i_1, i_{input}} \cdot x_{input} + b_{i_1}) \; ... \;) + b_{i_L}$$



NN Training

Data is split into train and test, train data is batched and passed through the NN during training, repeating for many epochs. Test data is used to independently evaluate performance after training. Loss function $\mathcal{L}$, usually compares the outputs to expected values on known (input, output) pairs. Optimiser stochastically updates the parameters $\theta = (W, b)$ according to $-\frac{\partial \mathcal{L}}{\partial \theta}$ on batches of data.

Machine Learning: Grokking

Grokking: A Phenomena in Learning

- Grokking is a mode of training in which model generalisation emerges suddenly after a long period of overfitting.
- We contrast this to steady learning, attempting to shed light on two common questions in the literature:
  1) Does Grokking lead to a different representation?
  2) Does Grokking lead to a more efficient model?

...the first we asses by comparing learned features, the second by examining performance stability under pruning of the model.

Grokking Dynamics

- Considering the NN parameters (W, b) as coordinates in a model space, as a model trains and the parameters update the position in this model space changes and the training process traces out a trajectory.
- We examine the model space paths under Grokking and steady learning; introducing novel information-geometric measures.

Machine Learning: FIMs

Fisher Information Metrics

- NN classification model with N parameters $\{\theta_i\}$, has model space $\sim \mathbb{R}^N$.
- To define distances we require a metric (traditionally assumed to be δ_{ij}), first by defining a basis for the model space tangent space with score vectors $\ell_i = \frac{\partial}{\partial\theta_i}\ln(f_{NN})$, defining the FIM as: $g_{ij}^{FIM} := \mathbb{E}_x(\ell_i \ell_j)$,

...as an expectation over the data space x, approximated with a sample.

Connection to KL Divergence

- The FIM measures similarity between models, illustrated by its connection to D_{KL}, which measures relative entropy between probability distributions $p(x|\theta)$: $D_{KL}(\theta'|\theta) = \mathbb{E}_x\Big(\ln\Big(\frac{p(x|\theta')}{p(x|\theta)}\Big)\Big)$,

...which has a unique global minimum at $\theta' = \theta$, in the neighbourhood of that minimum it expands as:

$$D_{KL}(\theta + \delta\theta|\theta) = \frac{1}{2} g_{ij}^{FIM}(\theta)\delta\theta_i\delta\theta_j + \mathcal{O}(\delta\theta^3) \ .$$

Results: ML Problems

Ising Model Binary Classification

- NN model trained to identify phase of an Ising system.
- Inputs are 16×16 square grids of $\sigma_i = \pm 1$ spin values. Outputs are 0 or 1 for classification as either disordered or ordered.
- Data generated by running Monte-Carlo local-update simulation using temperatures either above or below the critical temperature.
- Energy $E = \sum_{\langle i,j \rangle} \sigma_i \sigma_j$, or Magnetisation $M = \sum_i \sigma_i$, would classify the phase in an infinite equilibrated system.

Modular Addition Classification

- NN model trained to solve modular addition of the form:

$$c = (a + b)\%P \,, \qquad \text{...for } P = 113.$$

- Inputs are one-hot encoded vectors representing (a, b), and output is a one-hot encoded vector representing c.

Results: Inducing Grokking

Formal Definitions

- *Grokking time*: the number of training epochs between the model being within 5% of its maximum test accuracy, and the model being within 5% of its maximum train accuracy: $t_{\text{grok}} := t_{test} - t_{train}$

...then a training run is Grokking if $t_{\text{grok}} > t_{\text{train}}$.

Varying Weight Multiplier

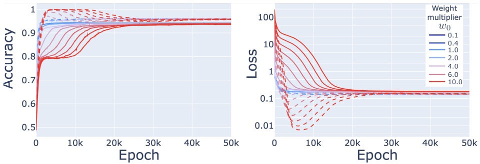
- To switch between Grokking and steady learning regimes, one needs to change hyperparameters of the training.
- To do this we increase the *weight multiplier*, which scales the parameters' initialisation values, to induce Grokking.

Example Visualisations

Ising measures, for varying weight multiplier w_0.

Dashed lines $\implies$ Train

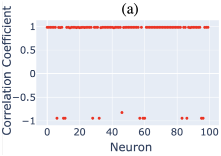
Solid lines $\implies$ Test

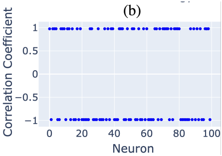


Results: Features & Encodings

Same Features

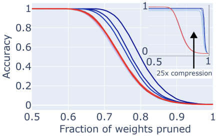
- Biasing the data such that the magnetisation doesn't correlate with phase well, NN learns the energy.
- Shown by perfect correlation across test data between final layer neuron activations and energy; for both Grokking (a) and steady learning (b).
- Similar behaviour is shown in the modular addition task learning Fourier coefficients.





Different Encoding Compressibility

- Ising task shows no difference in compressibility.
- Modular addition task shows hyperparameter phases where Grokking regimes lead to *less* compressible final models. Designed some measures based on $a(p) = \int a(p)dp$, $p \Rightarrow$ proportion of weights pruned.



Results: Trajectories

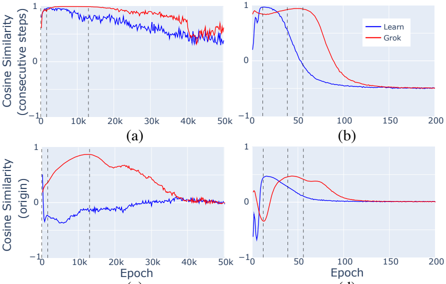
Novel Information-Geometric Measures

- For model space position θ^e at epoch e, and training step $s^e := \theta^{e+f} - \theta^e$ at a frequency of f epochs, we define novel measures for trajectory *speed* and *direction*:

$$|s|_{FIM} := \sqrt{s_i \cdot g_{ij}^{FIM} \cdot s_j} \quad \in [0, \infty) \ , \qquad S_{C-FIM}(s, s') := \frac{s_i \cdot g_{ij}^{FIM} \cdot s_j'}{|s|_{FIM} |s'|_{FIM}} \quad \in [-1, 1] \ .$$

Fisher Cosine Similarities

- $S_{C-FIM}(s^e, s^{e+1})$ shows the Grokking phase is a *straight line* in information space.
- $S_{C-FIM}(s^e, s^{OT})$ shows that in the Ising task Grokking behaviour is dominated by weight decay, but is *not* in the modular addition task.



(a) (b)
(c) (d)

Results

- Grokking is a learning phenomena of delayed generalisation, it can be induced with varying weight multiplier across tasks.
- Grokking regimes learn the same features as steady learning regimes, but can express differences in compressibility.
- Novel information-geometric measures show the Grokking trajectory in information-space is especially unique, following a straight line.

Outlook

- Refining the novel information-geometric measures, new perspectives can be provided on learning dynamics for other phenomena.
- Ongoing work has designed a Fisher equivalent of K-Means clustering to measure similarity of ensembles of trained ML models.