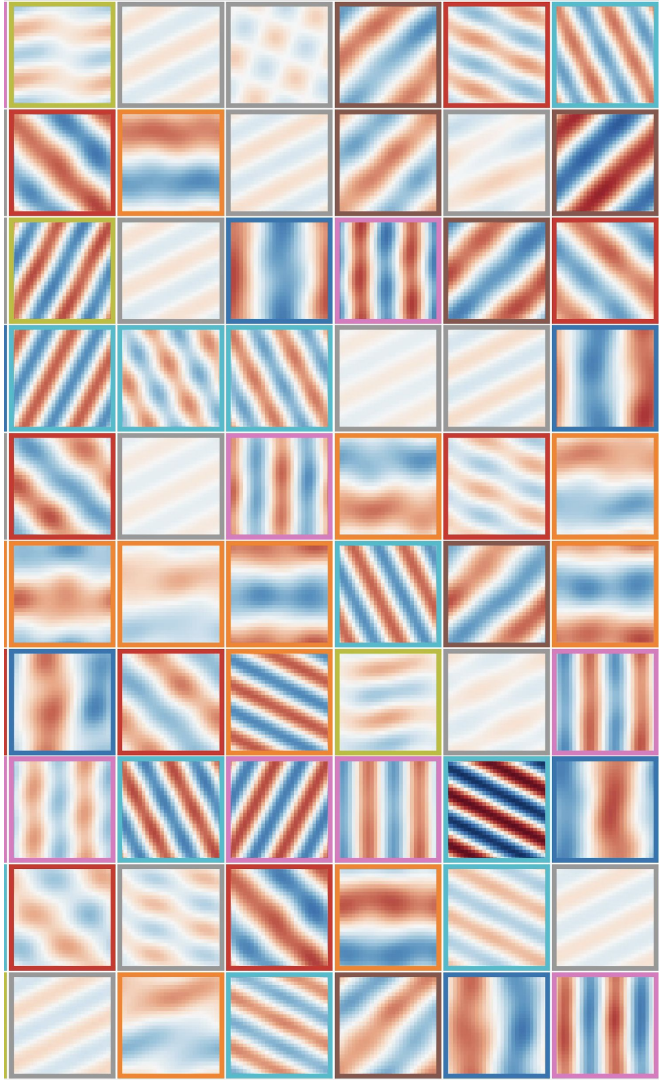
Sequential Group Composition:

A Window into the Mechanics of Deep Learning



Giovanni Luca Marchetti*

Daniel Kunin*

Adele Myers

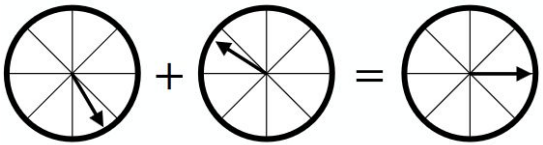
Francisco Acosta

Nina Miolane

*Equal contribution

Overview

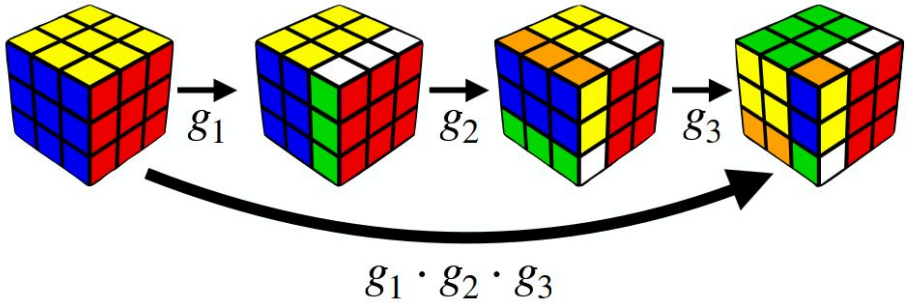
Modular addition



$a + b \mod p$

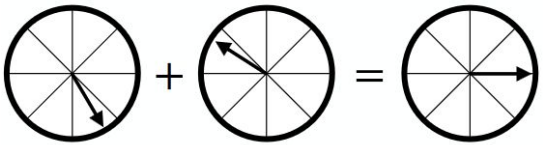
$\subset$

Group composition



Overview

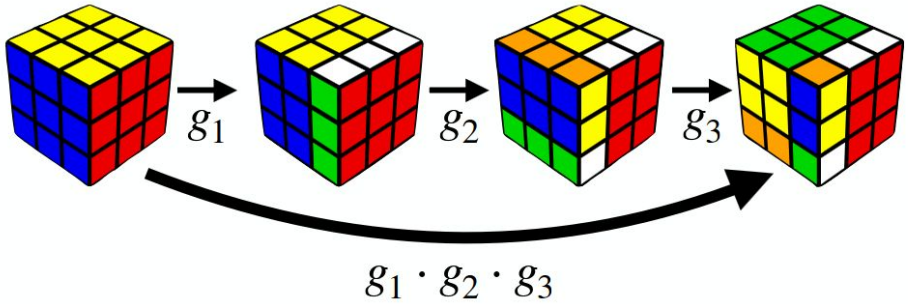
Modular addition



$a + b \mod p$

$\subset$

Group composition



Q: How do neural networks learn to compose a sequence of elements from an arbitrary group?

A: Via abstract Fourier analysis over groups!

Sequential Group Composition

Task

Given a finite group, map a sequence of group elements to their product

$$(g_1, \ldots, g_k) \in G^k \mapsto \prod_{i=1}^{k} g_i \in G$$

Data

Group elements are encoded by permuting a fixed “encoding” vector

Architecture

Shallow MLPs w/ polynomial activation, plus results for deep MLPs and RNNs

Learning Rule

MSE loss optimized via gradient flow, with initialization tending to zero

Main Result for Shallow Networks

We rely on an previous theory of feature learning in the vanishing initialization regime [1].

[1] *Kunin et al., Alternating Gradient Flows: A Theory of Feature Learning in Two-layer Neural Networks, NeurIPS 2025*

Main Result for Shallow Networks

We rely on an previous theory of feature learning in the vanishing initialization regime [1].

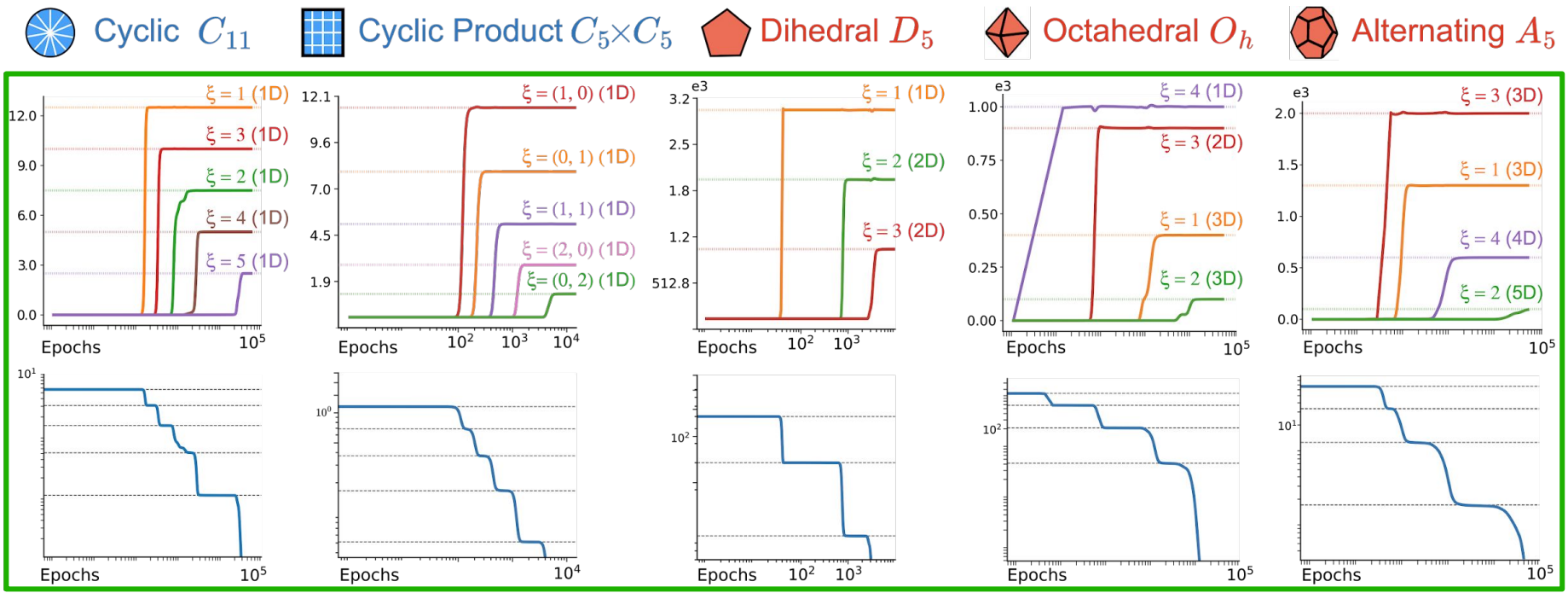
Informal Theorem

The task is learned in phases of "grokking". At each phase:

- *Neurons independently specialize to a new irreducible representation ("harmonic") of the group.*
- *Specialized neurons collectively minimize the loss.*

[1] *Kunin et al., Alternating Gradient Flows: A Theory of Feature Learning in Two-layer Neural Networks, NeurIPS 2025*

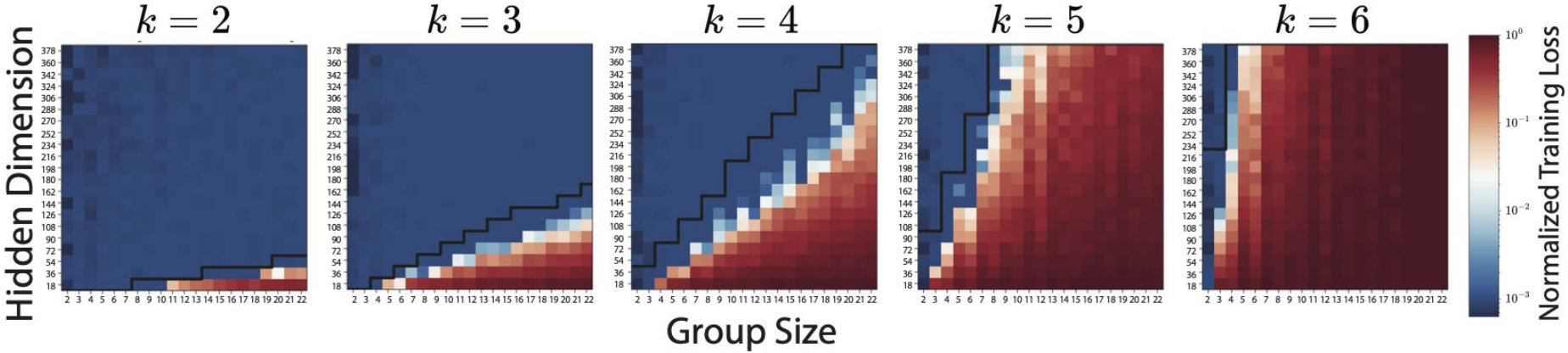
Empirics


Cyclic C_{11}
Cyclic Product $C_5 \times C_5$
Dihedral D_5
Octahedral O_h
Alternating A_5
Power Spectrum
Training Loss
ξ = 1 (1D)
ξ = 3 (1D)
ξ = 2 (1D)
ξ = 4 (1D)
ξ = 5 (1D)
ξ = (1, 0) (1D)
ξ = (0, 1) (1D)
ξ = (1, 1) (1D)
ξ = (2, 0) (1D)
ξ= (0, 2) (1D)
ξ = 1 (1D)
ξ = 2 (2D)
ξ = 3 (2D)
ξ = 4 (1D)
ξ = 3 (2D)
ξ = 1 (3D)
ξ = 2 (3D)
ξ = 3 (3D)
ξ = 1 (3D)
ξ = 4 (4D)
ξ = 2 (5D)
Epochs
e3

Shallow Networks need an Exponential Width

Two-layer networks require a hidden width exponential *in sequence length*

$$H = (k+1)2^{k-1}\left\lfloor \frac{|G|}{2} \right\rfloor$$



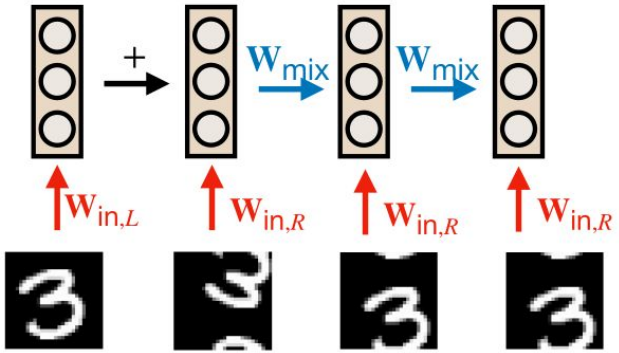
Q: Can deeper networks learn more efficient solutions?

Deeper Architectures

A: Deeper MLPs and RNNs can learn more efficient solutions by leveraging associativity of group composition.

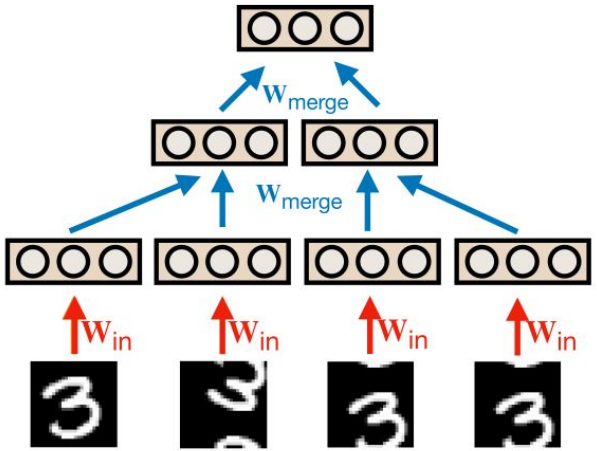
RNN

$$g_1 \cdots g_k = (((((g_1 \cdot g_2) \cdot g_3) \cdot g_4) \cdots g_{k-1}) \cdot g_k)$$



MLP

$$g_1 \cdots g_k = \big((g_1 \cdot g_2) \cdot (g_3 \cdot g_4)\big) \cdots (g_{k-1} \cdot g_k)$$



Thank you!

paper

code