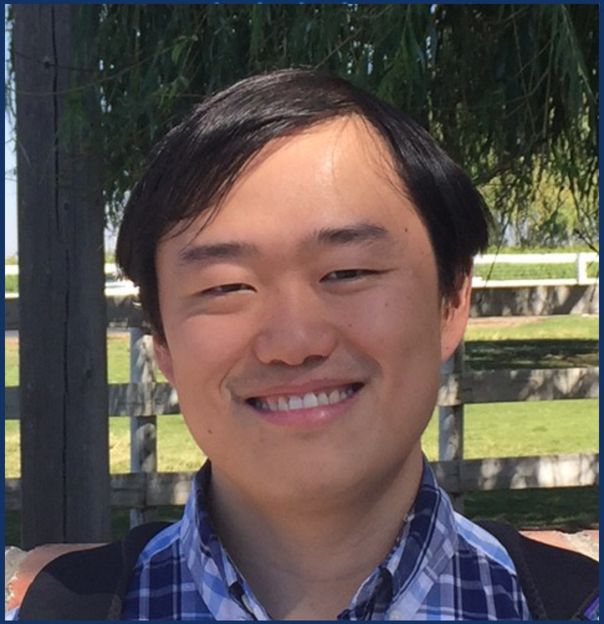
Provable Scaling Laws of Feature Emergence from Learning Dynamics of Grokking

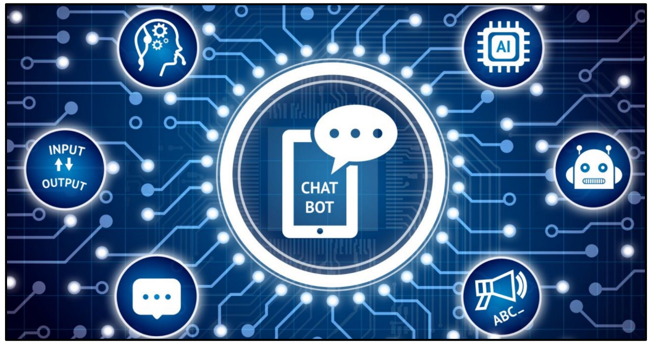
Yuandong Tian
(ex-)Research Scientist Director

Meta FAIR



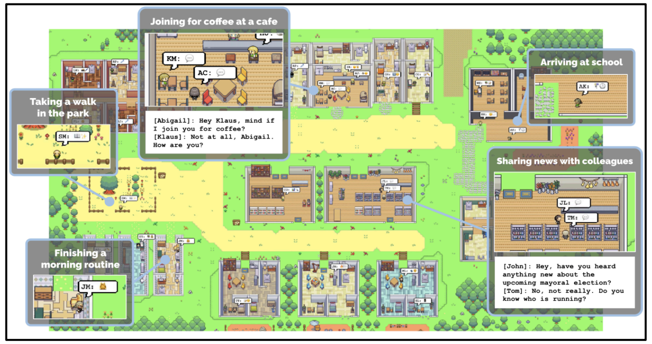
facebook Artificial Intelligence

Large Language Models (LLMs)

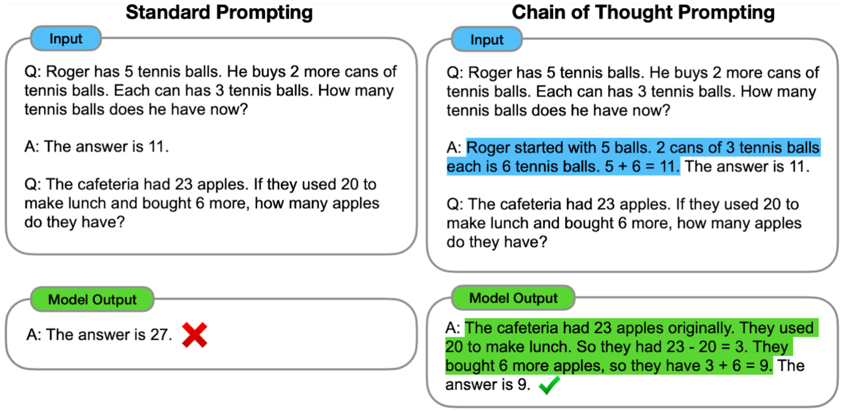


Conversational AI

Content Generation



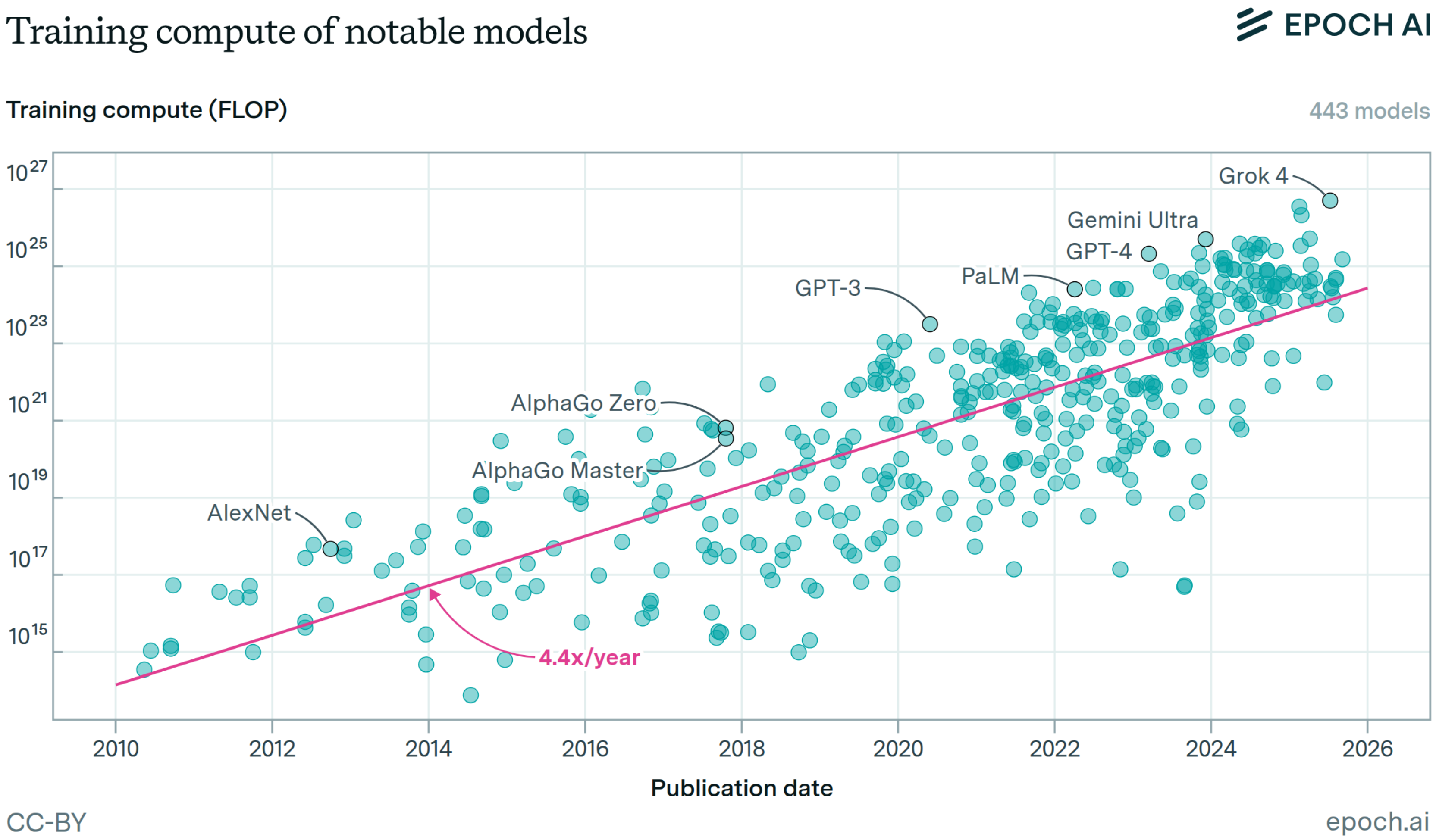
AI Agents



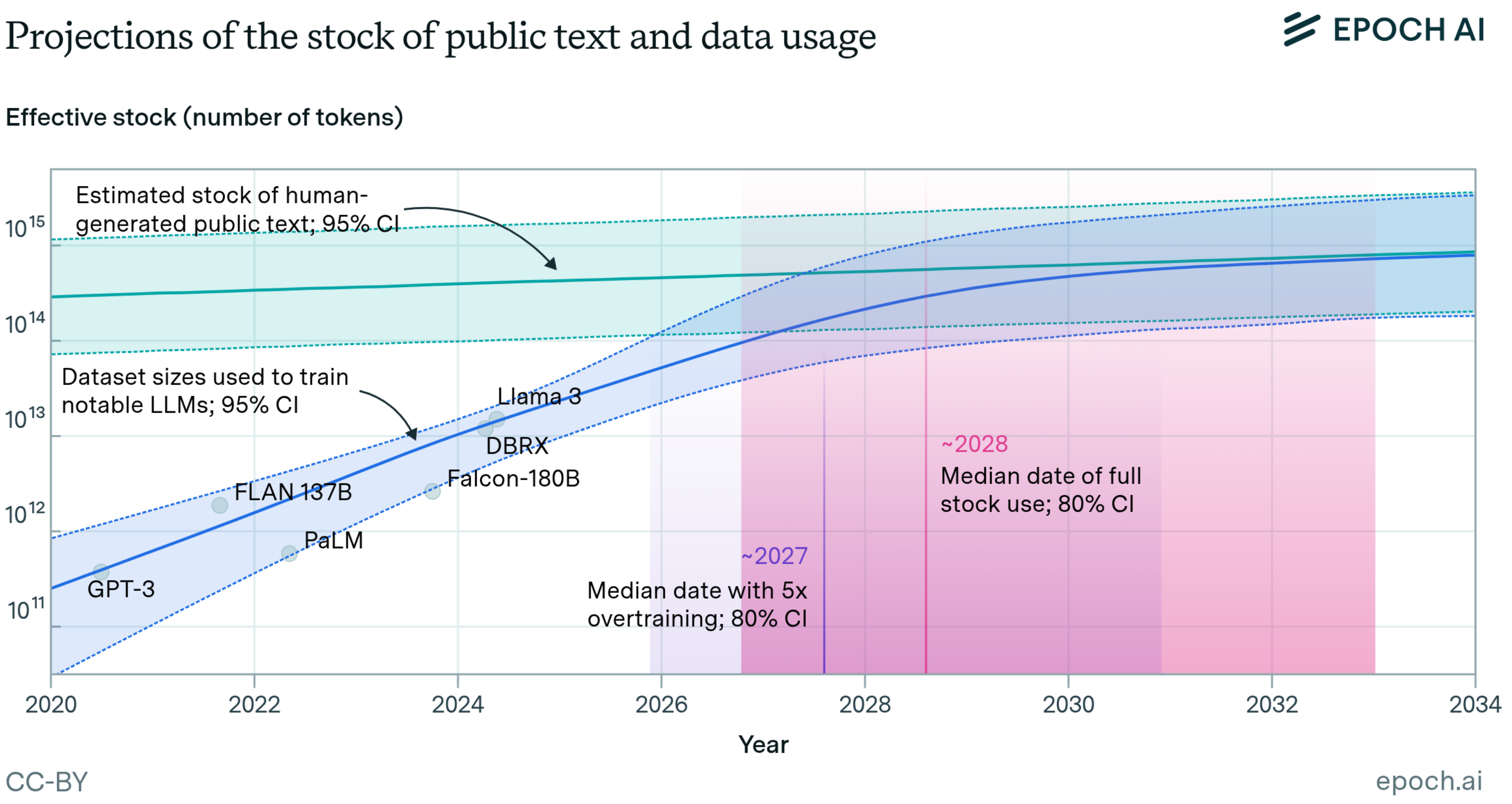
Reasoning

Planning

The Progress of Large Models



The Data Usage



Comparison between Human and SoTA LMs

Question: Is our AI as strong as humans yet?

	Training Data efficiency	Power Consumption	Adaptation to New Tasks	How to make decision?
Human Brain	< 10B text tokens, a lot of sensory inputs	Learning: ~20W Inference/Thinking: ~20W	Learn with a few examples	By casual relationships and deep understanding
Sota LMs	~10T-50T tokens	Learning: at least @ MWh Inference/Thinking: 1W-30W	Hundreds / Thousands of data points. May fail to generalize	Correlation & Pattern Matching

Estimated #tokens consumed by human in the life time: 70 years * 300 days / year * 12 hours / day * 3600 seconds / hours * 10 tokens / second = 9.1B

Andrej Karpathy
@karpathy

My pleasure to come on Dwarkesh last week, I thought the questions and conversation were really good.

I re-watched the pod just now too. First of all, yes I know, and I'm sorry that I speak so fast :). It's to my detriment because sometimes my speaking thread out-executes my thinking thread, so I think I botched a few explanations due to that, and sometimes I was also nervous that I'm going too much on a tangent or too deep into something relatively spurious. Anyway, a few notes/pointers:

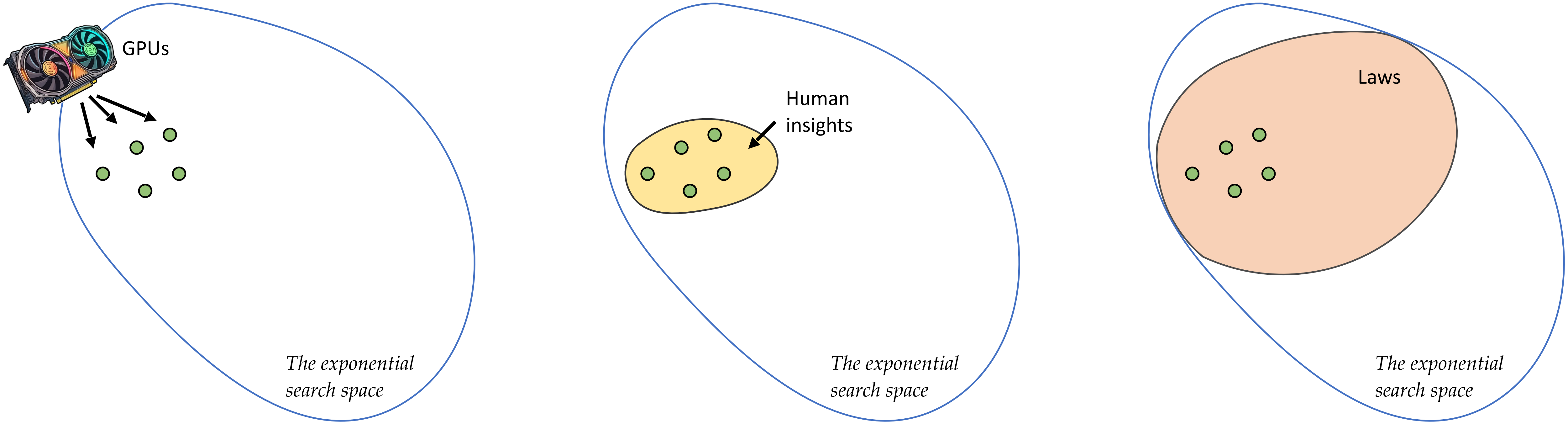
AGI timelines. My comments on AGI timelines looks to be the most trending part of the early response. This is the "decade of agents" is a reference to this earlier tweet x.com/karpathy/statu... Basically my AI timelines are about 5-10X pessimistic w.r.t. what you'll find in your neighborhood SF AI house party or on your twitter timeline, but still quite optimistic w.r.t. a rising tide of AI deniers and skeptics. The apparent conflict is not: imo we simultaneously 1) saw a huge amount of progress in recent years with LLMs while 2) there is still a lot of work remaining (grunt work, integration work, sensors and actuators to the physical world, societal work, safety and security work (jailbreaks, poisoning, etc.)) and also research to get done before we have an entity that you'd prefer to hire over a person for an arbitrary job in the world. I think that overall, 10 years should otherwise be a very bullish timeline for AGI, it's only in contrast to present hype that it doesn't feel that way.

How we should do our research from now on?

- The “Data Wall problem”
  - We may have used all the available data on the Internet.
  - How to deal with corner cases / personalization / private data?
  - Human is still much more efficient than current AI

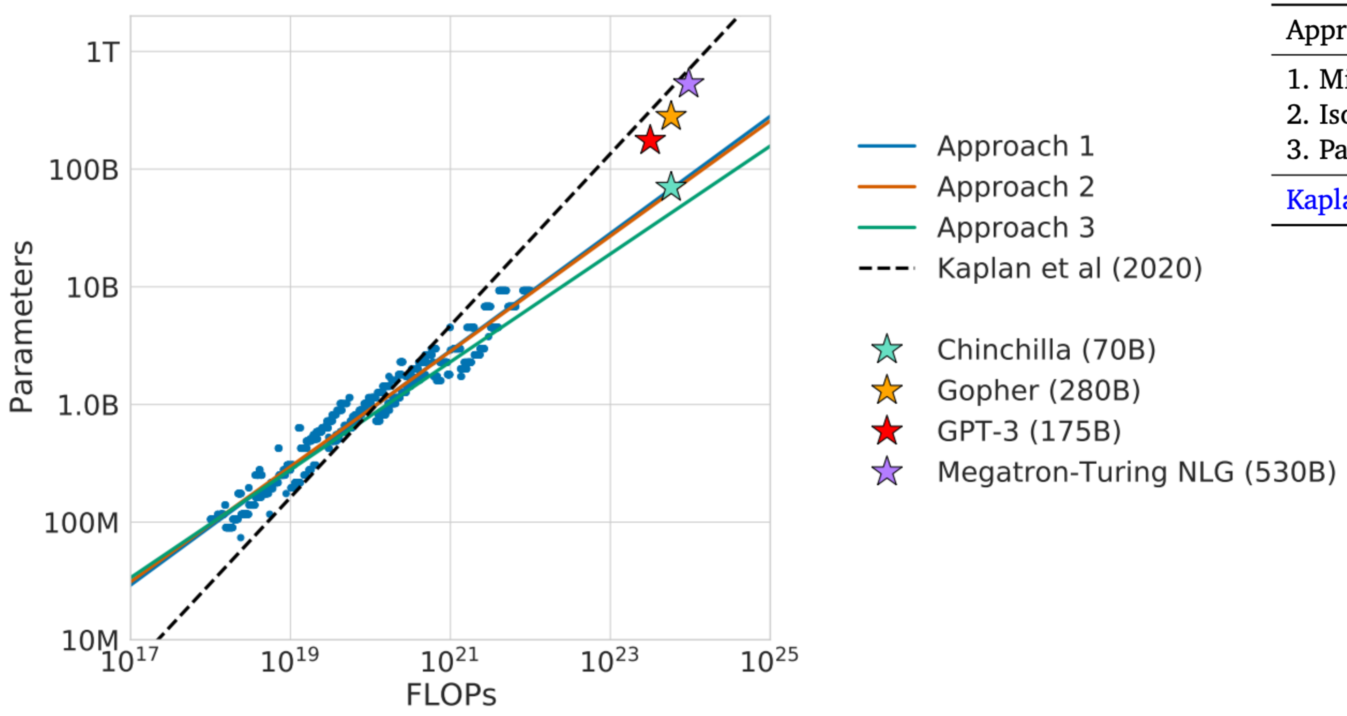
- Everyone is GPU poor
  - What are new axes to scale? GPUs are never enough.
  - Data itself cannot extrapolate, only human insights can.

The New (a.k.a. Old) Scaling Axis



Question: Can we **scale** the scaling laws?

How we get Scaling Laws?



Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Steps:

1. Collect the experiments
2. Form hypothesis (linear, power-law, etc)
3. Extrapolate

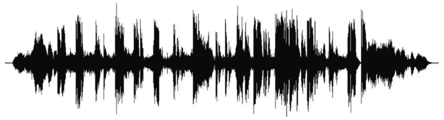
Still pure statistics and need exponential data.

(No leverage of the knowledge of architecture/data)

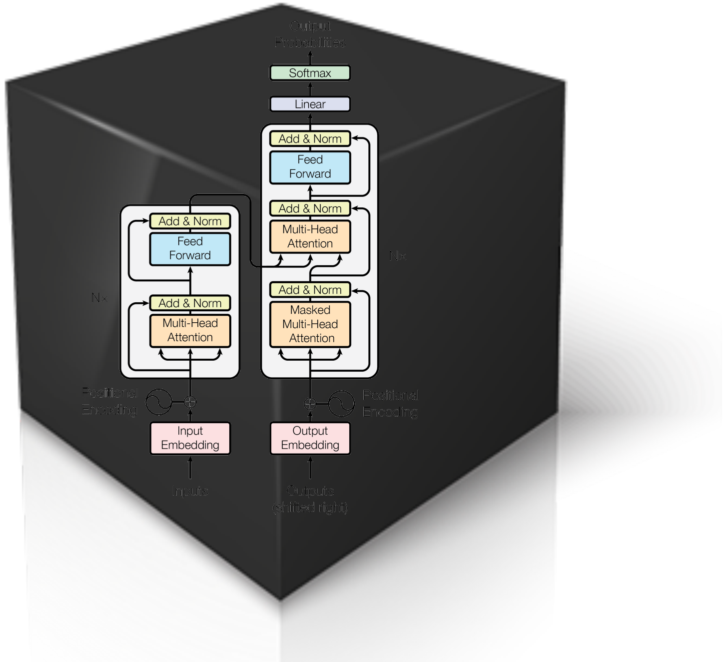
How does deep learning work?

Input

Output

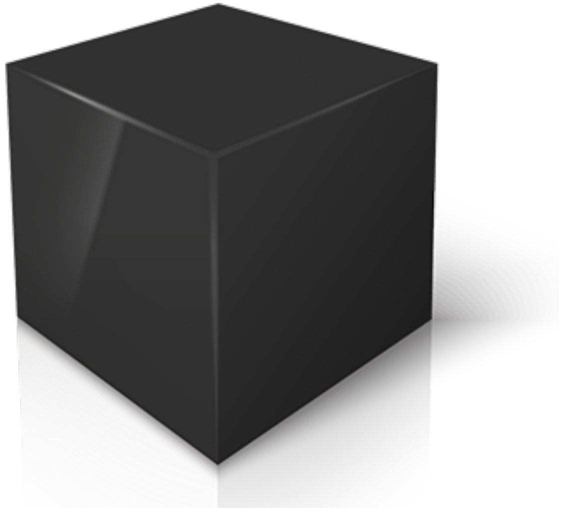


This is an apple

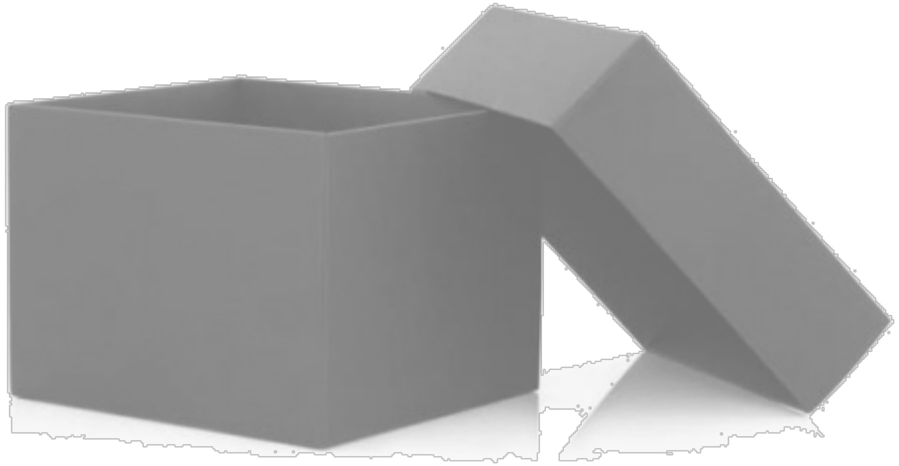


"Some Nonlinear Transformation"

Black-box versus White-box



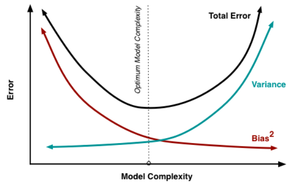
Black box



White box

What routes should we take?

Generalization



Architecture ✗

training dynamics ✗

Expressibility

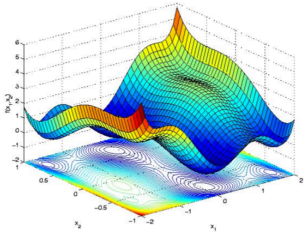
\+ −

− +

Architecture ✓

training dynamics ✗

Optimization



Architecture ✗

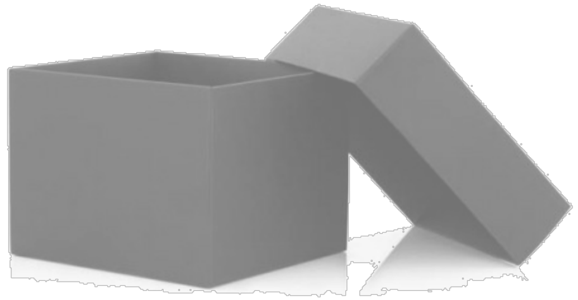
training dynamics ✓

How about

Architecture ✓

training dynamics ✓

Start From the First Principle



- Training follows Gradient and its variants (SGD, Adams, etc)

$$\dot{\boldsymbol{w}} := \frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} = -\nabla_{\boldsymbol{w}} J(\boldsymbol{w})$$

- **First principle** → Understand the behavior of the neural networks by checking the gradient **dynamics** induced by the neural **architectures**.

- Sounds complicated.. Is that possible? **Yes**

Architecture ✓

training dynamics ✓

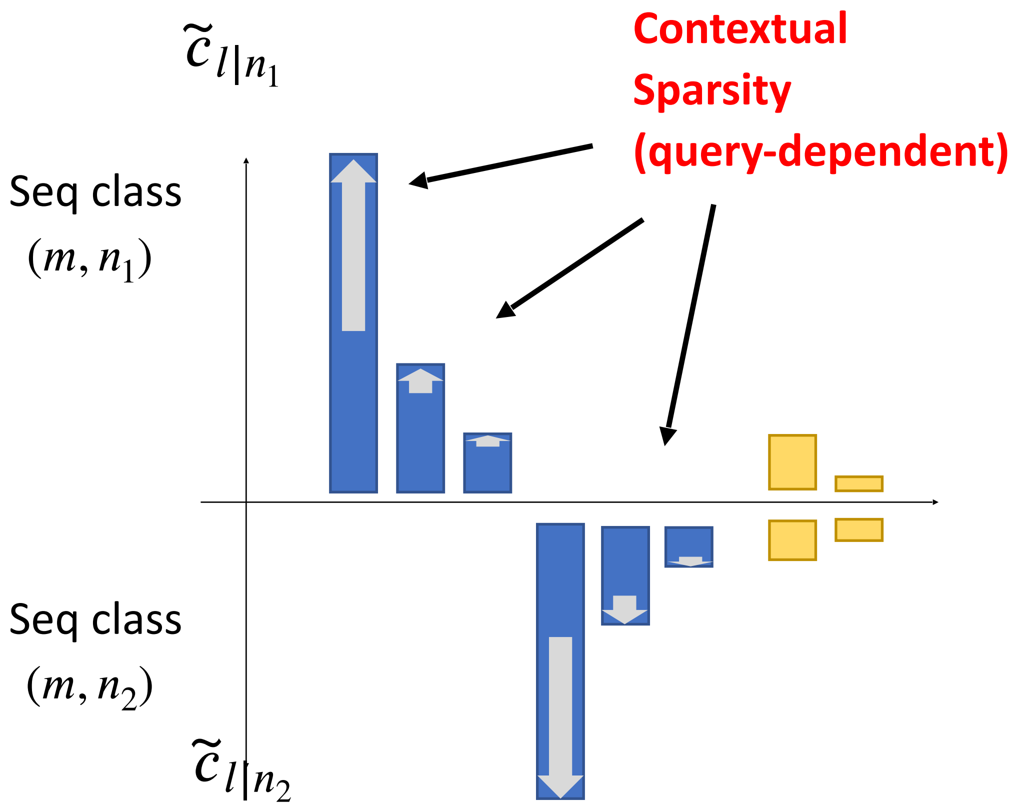
What Gradient Descent Analysis gives us?

Short-term:
Finding Simple Structures
(Low-rank, sparsity)

Long-term:
How the representation is learned
(Key to the success of deep models)

Leverage Them in Practical Algorithms

Finding Nice Structure: Attention Sparsity

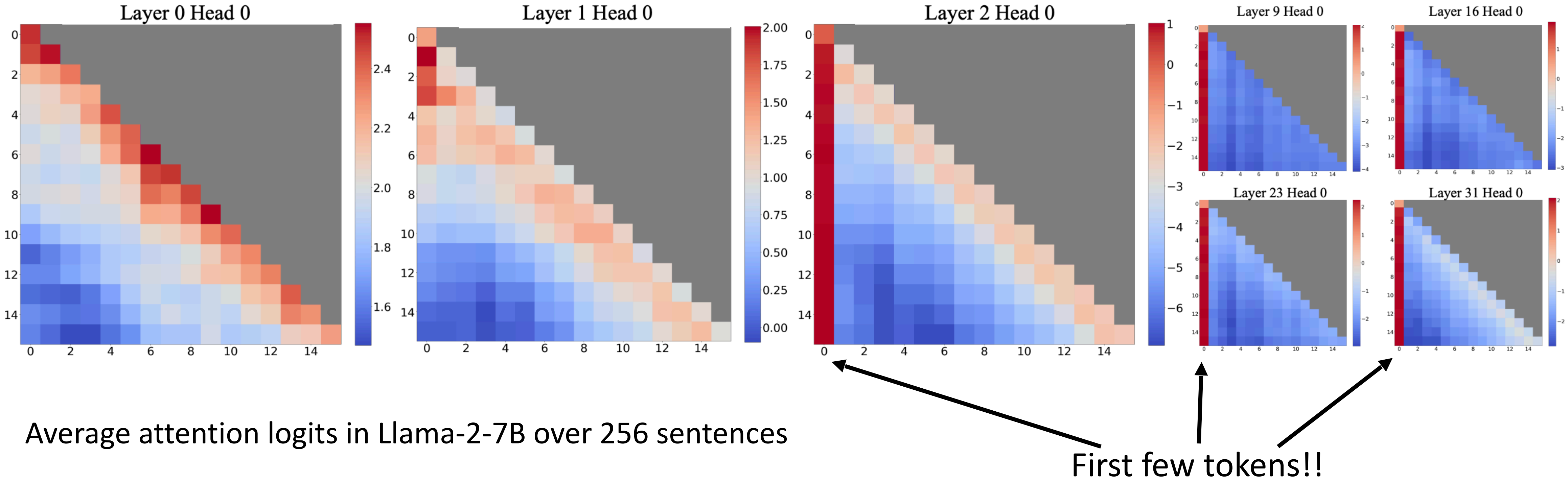


Attention = Learnable TF-IDF (Term Frequency, Inverse Document Frequency)

[*Y. Tian et al*, Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer, *NeurIPS'23]*

[*Y. Tian et al*, JoMA: Demystifying Multilayer Transformers via JOint Dynamics of MLP and Attention, *ICLR'24]*

Attention Sinks: Initial tokens draw strong attentions

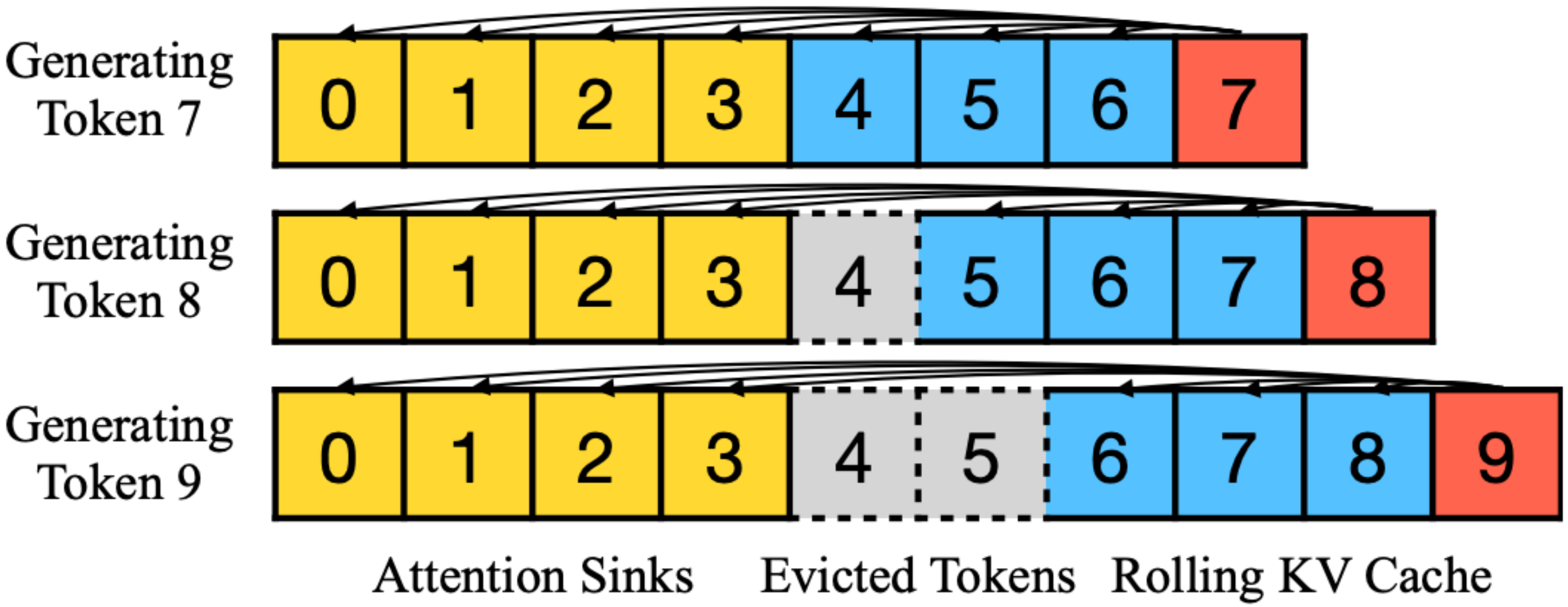


Average attention logits in Llama-2-7B over 256 sentences

- Observation: Initial tokens have large attention scores, even if they're not semantically significant.
- Attention Sink: Tokens that disproportionately attract attention irrespective of their relevance.

[G. Xiao et al, Efficient Streaming Language Models with Attention Sinks, ICLR'24]

StreamingLLM

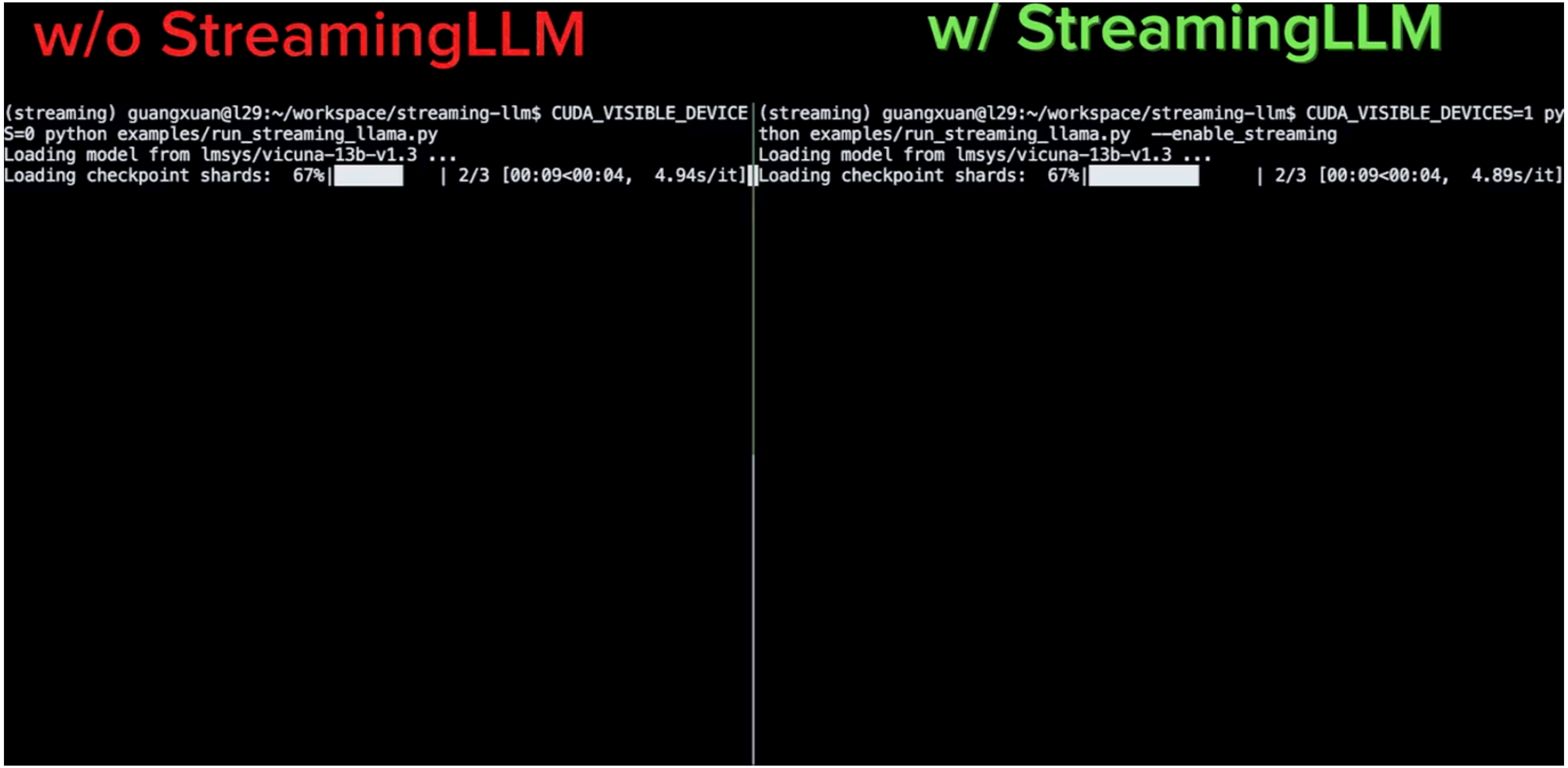


Key design: Position Rolling

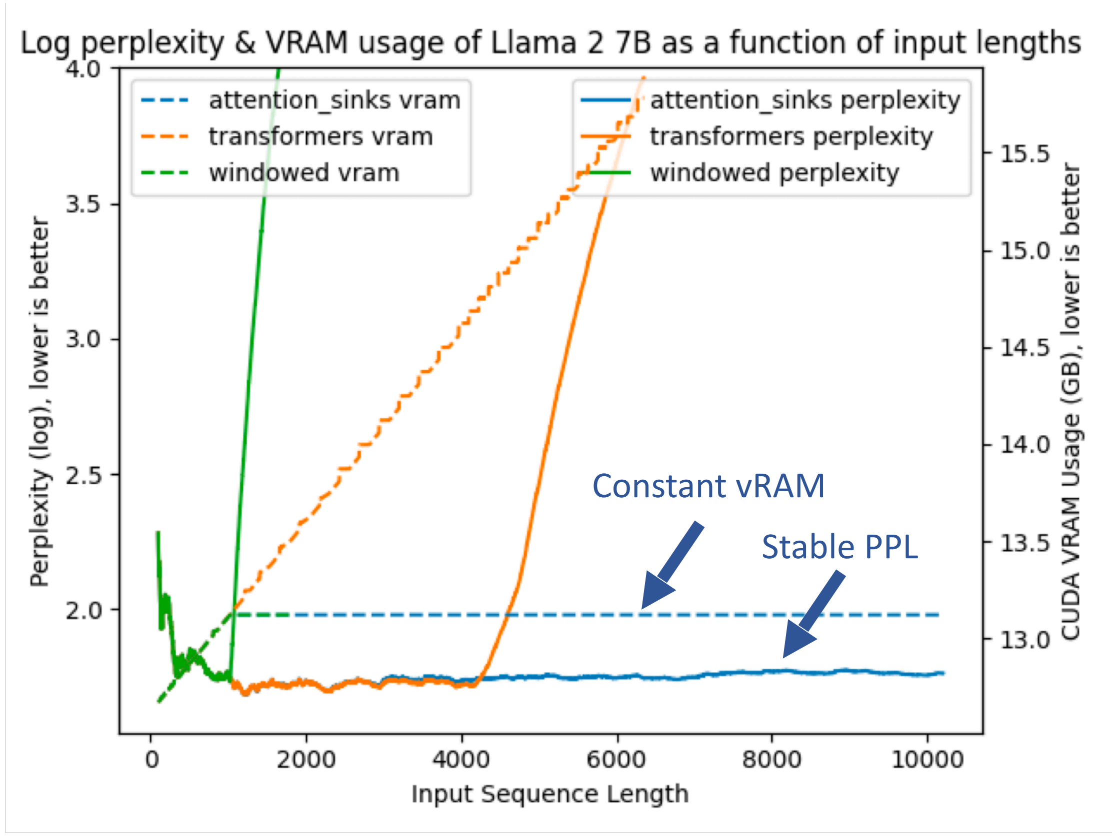
For all tokens, use their positions **within cache** to compute positional encoding!

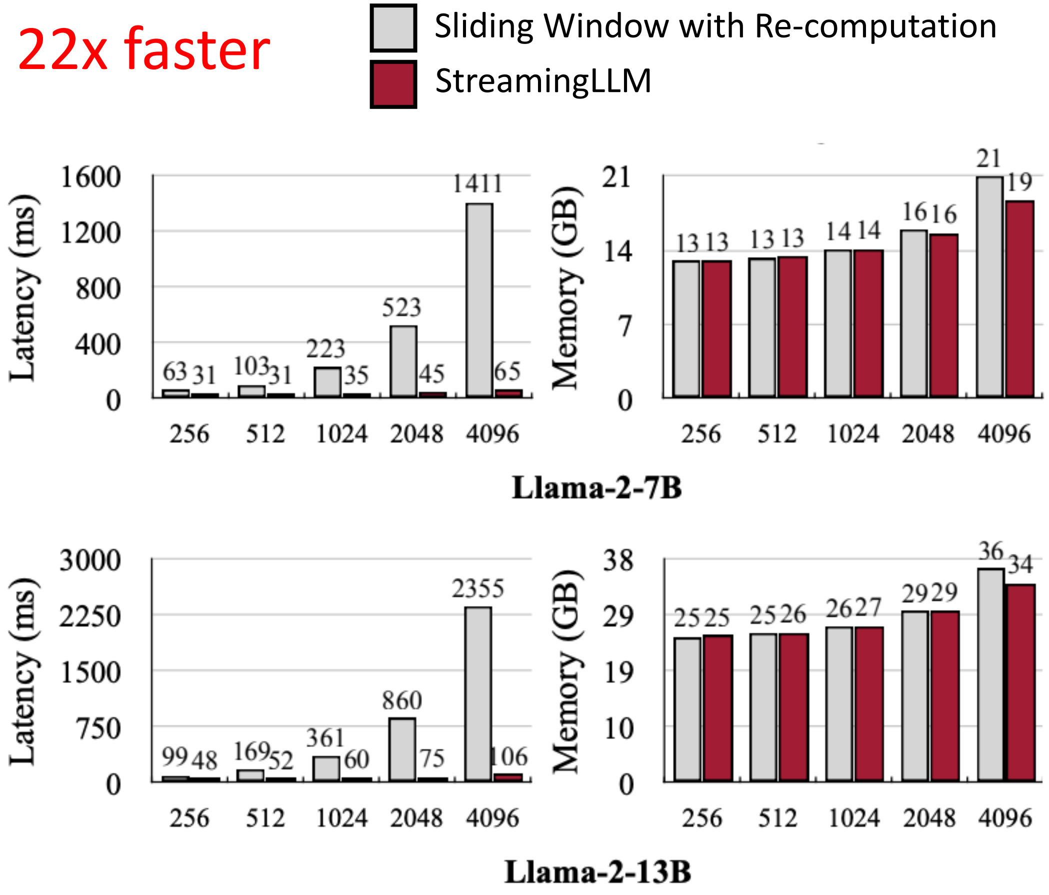
→ Token distance never exceeds pre-trained context window!

StreamingLLM



StreamingLLM: stable PPL, constant vRAM





Impact

Attention: Following GPT-3, attention blocks alternate between banded window and fully dense patterns [8][9], where the bandwidth is 128 tokens. Each layer has 64 query heads of dimension 64, and uses Grouped Query Attention (GQA [10][11]) with 8 key-value heads. We apply rotary position embeddings [12] and extend the context length of dense layers to 131,072 tokens using YaRN [13]. Each attention head has a learned bias in the denominator of the softmax, similar to off-by-one attention and attention sinks [14][15], which enables the attention mechanism to pay no attention to any tokens.

- 1k+ citations
- Used in GPT OSS models in pre-training

Long-term: How Network finds Representation

Type of Representations

Representation

(Traditional) Symbolic representation

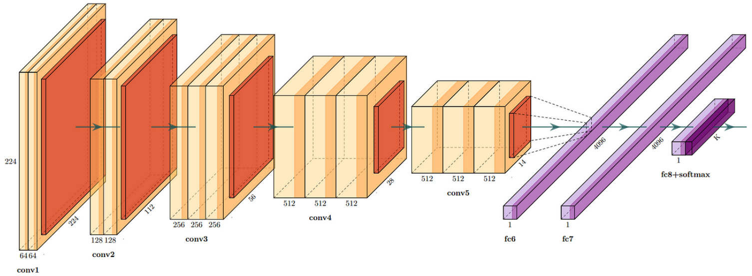
$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\varepsilon} \qquad \text{(Gauss' Law)}$$

$$\nabla \cdot \mathbf{H} = 0 \qquad \text{(Gauss' Law for Magnetism)}$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \qquad \text{(Faraday's Law)}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampere's Law)}$$

Neural Representation🤔



Why Neural Representation is so effective?

Representation

(Traditional) Symbolic representation

4 Conclusions

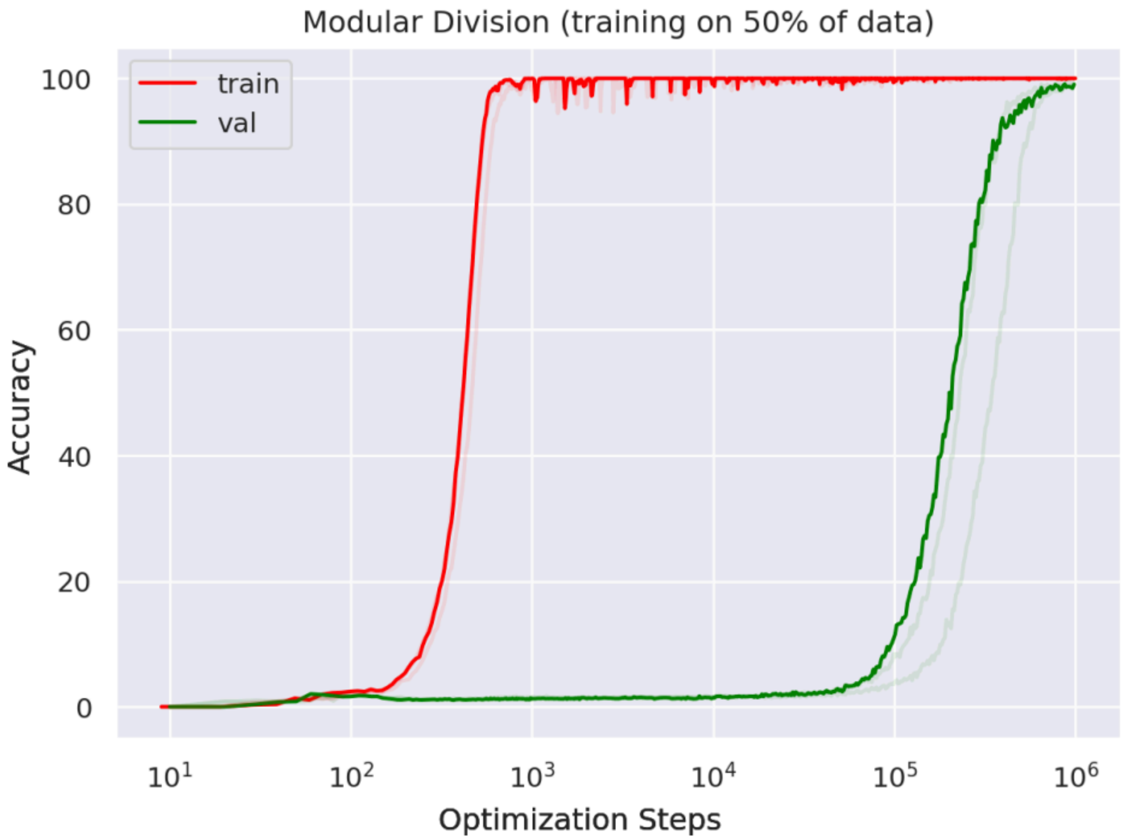
We have extended the GLU family of layers and proposed their use in Transformer. In a transfer-learning setup, the new variants seem to produce better perplexities for the de-noising objective used in pre-training, as well as better results on many downstream language-understanding tasks. These architectures are simple to implement, and have no apparent computational drawbacks. We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence.

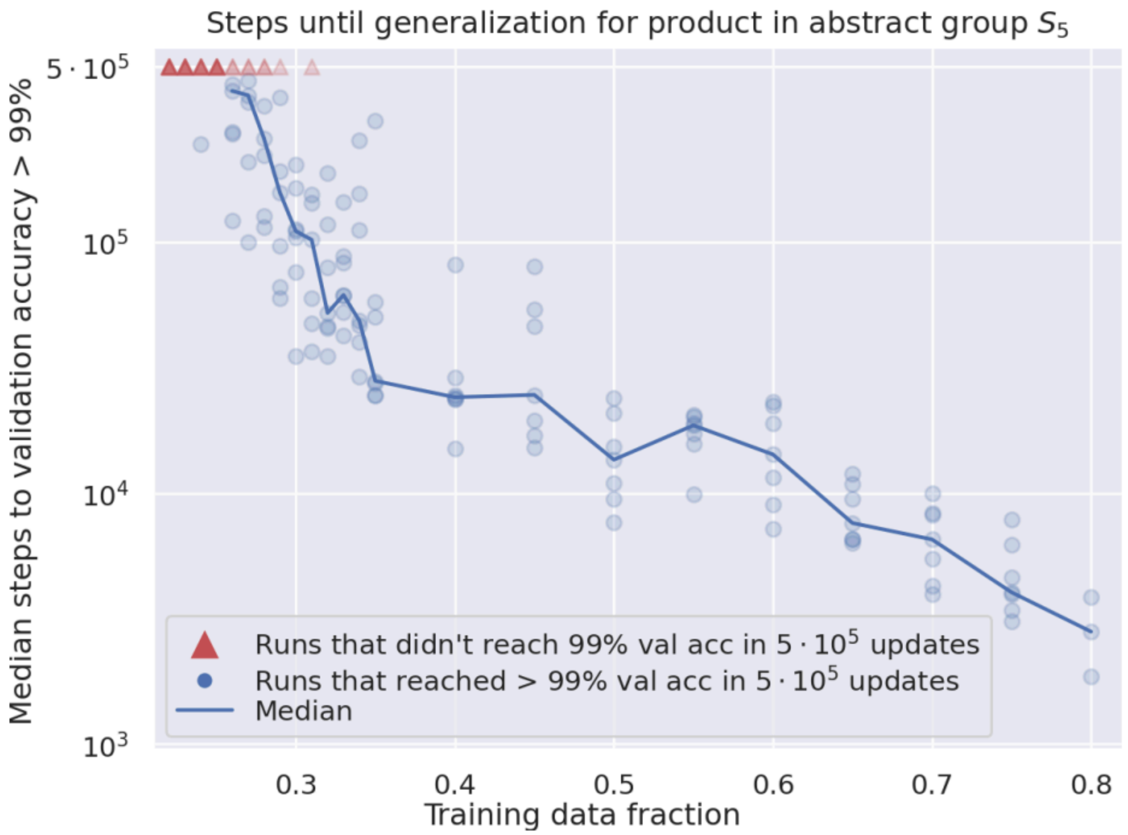
Neural Representation🤔

*Shall we just acknowledge that as "**divine benevolence**"?*

[Garfinkle, Allie (2024-08-02). "Character.AI's Noam Shazeer on what we know about AI—and what we don't". *Fortune*. Retrieved 2025-03-06.]
[N. Shazeer, *GLU Variants Improve Transformer*]

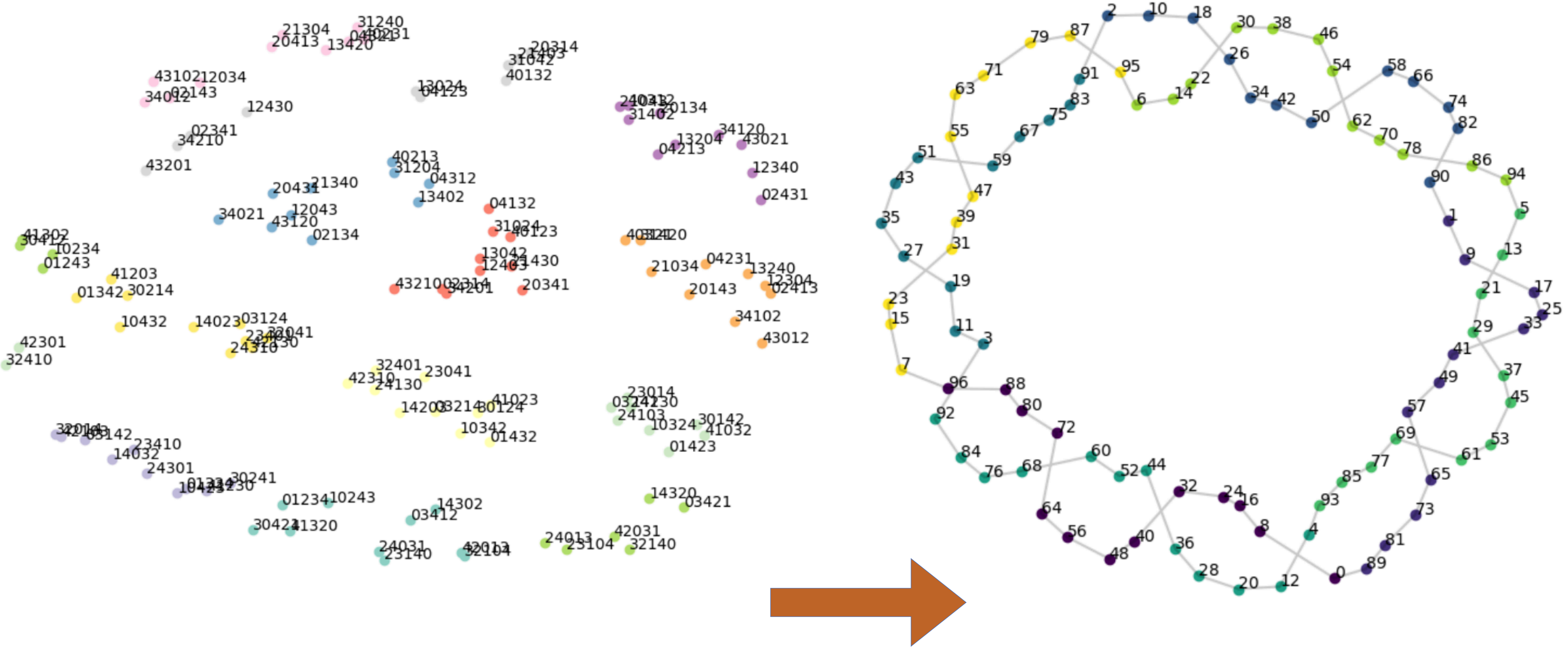
Why there is Grokking Behavior?





[A. Power et. al, *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*]

Feature Emergence through Grokking



🚨 Mechanism of Emergent Representation

Representation

(Traditional) Symbolic representation

Neural Representation🤔

Emerging Symbolic Structure

[Y. Tian, *Composing Global Solutions to Reasoning Tasks via Algebraic Objects in Neural Nets*, NeurIPS'25]
[Y. Tian, *Provable Scaling Laws of Feature Emergence from Learning Dynamics of Grokking*, arXiv'25]

Modular Addition

$$a + b = c \mod d$$

Does neural network have an *implicit table* to do retrieval?

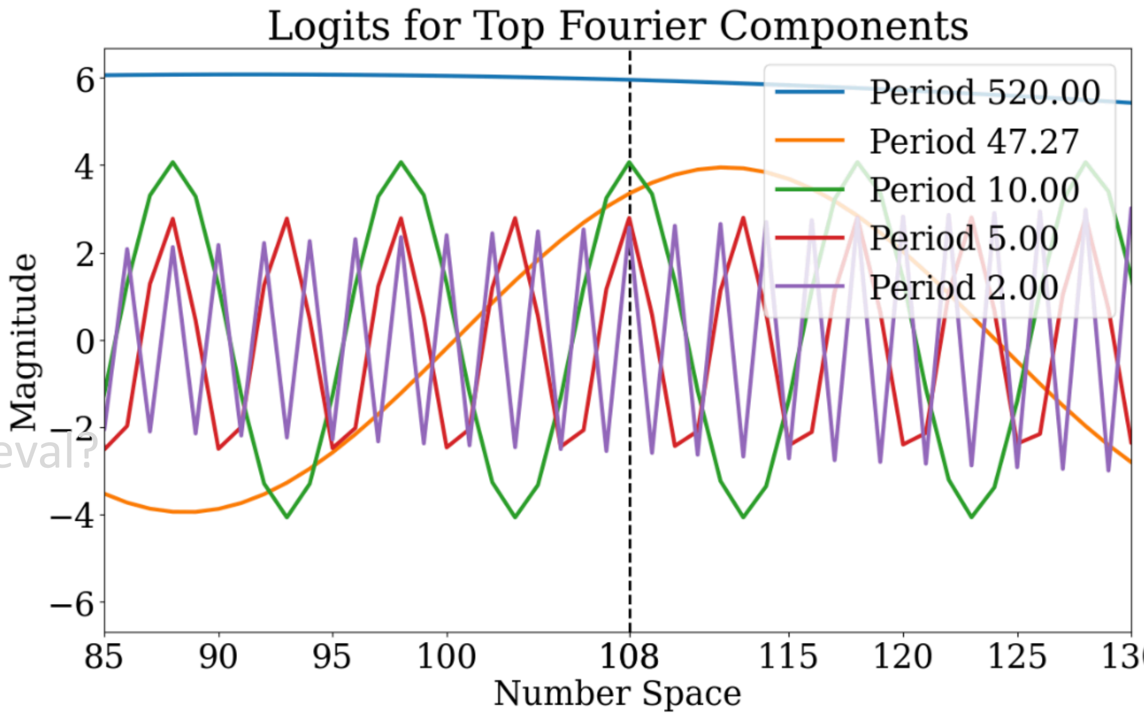
Modular Addition

$$a + b = c \bmod d$$

Does neural network have an *implicit table* to do retrieval?

Learned representation = Fourier basis 🤯

Why? 🤔

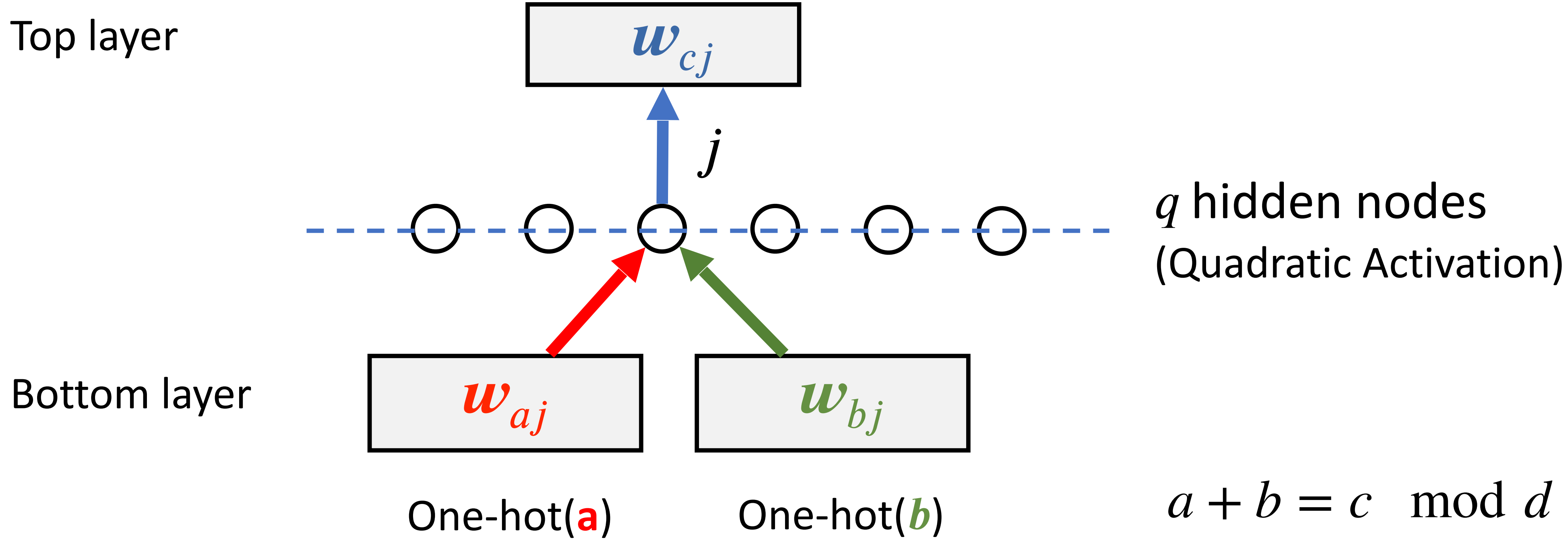


(a) Final logits for top Fourier components

[T. Zhou et al, *Pre-trained Large Language Models Use Fourier Features to Compute Addition*, NeurIPS'24]
[S. Kantamneni, *Language Models Use Trigonometry to Do Addition*, arXiv'25]

Minimal Problem Setup

MSE Loss: $Min \ \left\| \text{Output} - \text{one–hot}(\boldsymbol{c}) \right\|_2$

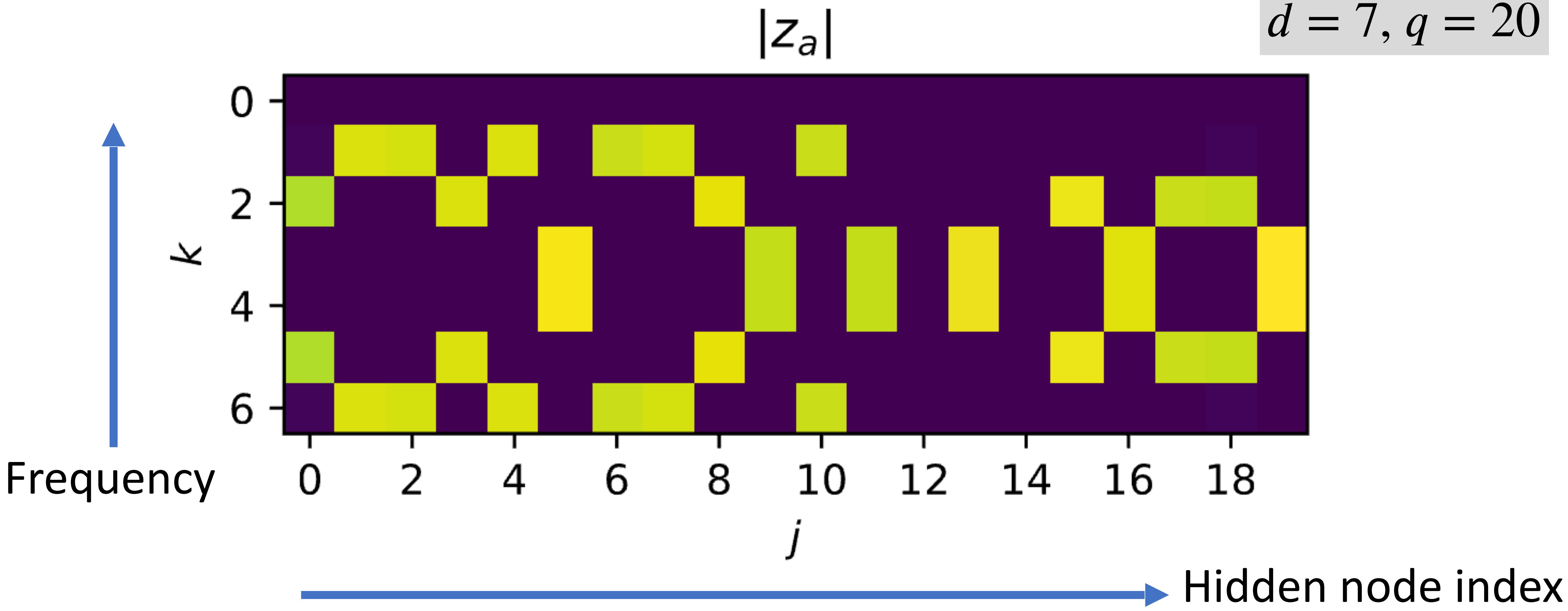


$$a + b = c \mod d$$

[Y. Tian, *Composing Global Solutions to Reasoning Tasks via Algebraic Objects in Neural Nets*, NeurIPS'25]

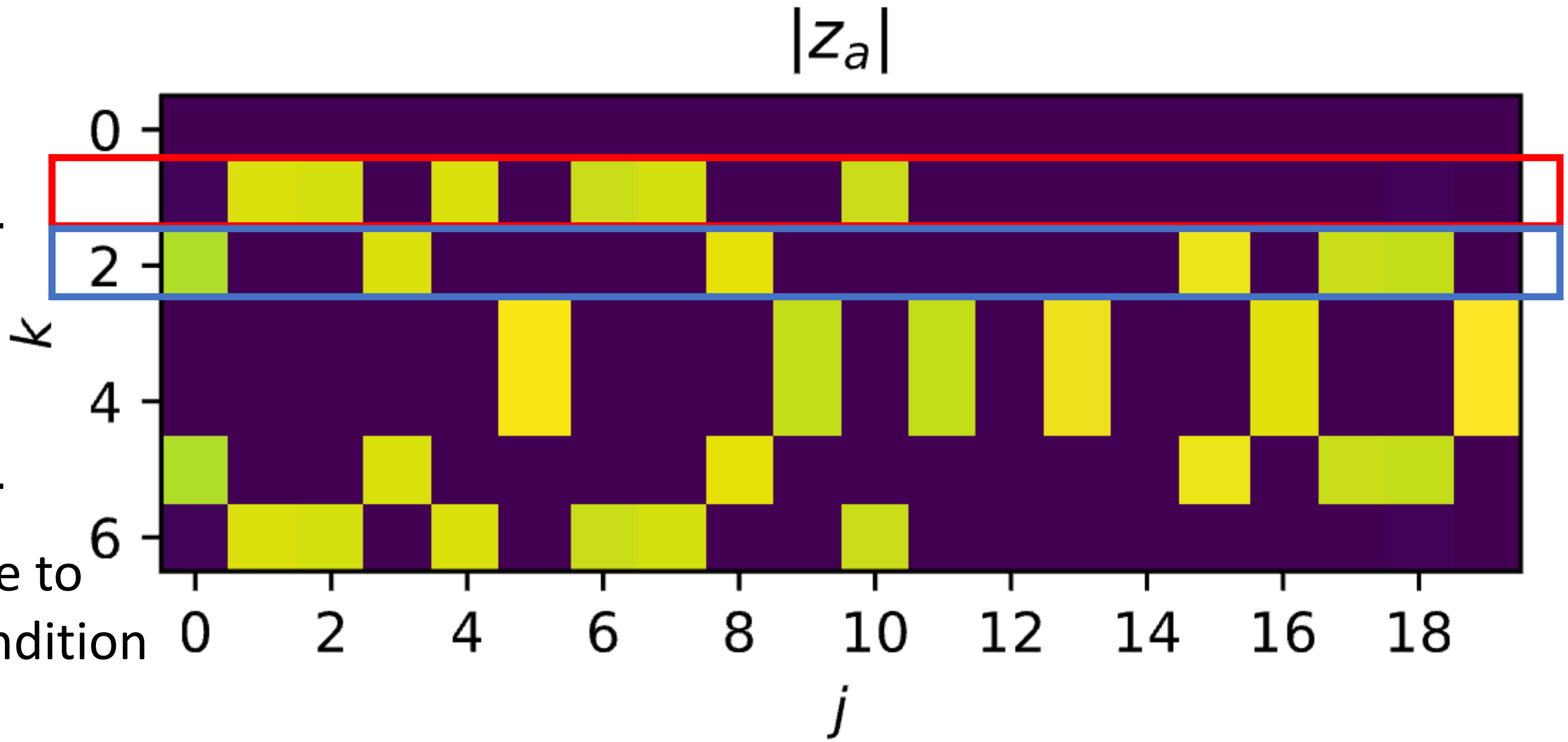
What a Gradient Descent Solution look like?

$d = 7, q = 20$



[Y. Tian, *Composing Global Solutions to Reasoning Tasks via Algebraic Objects in Neural Nets*, NeurIPS'25]

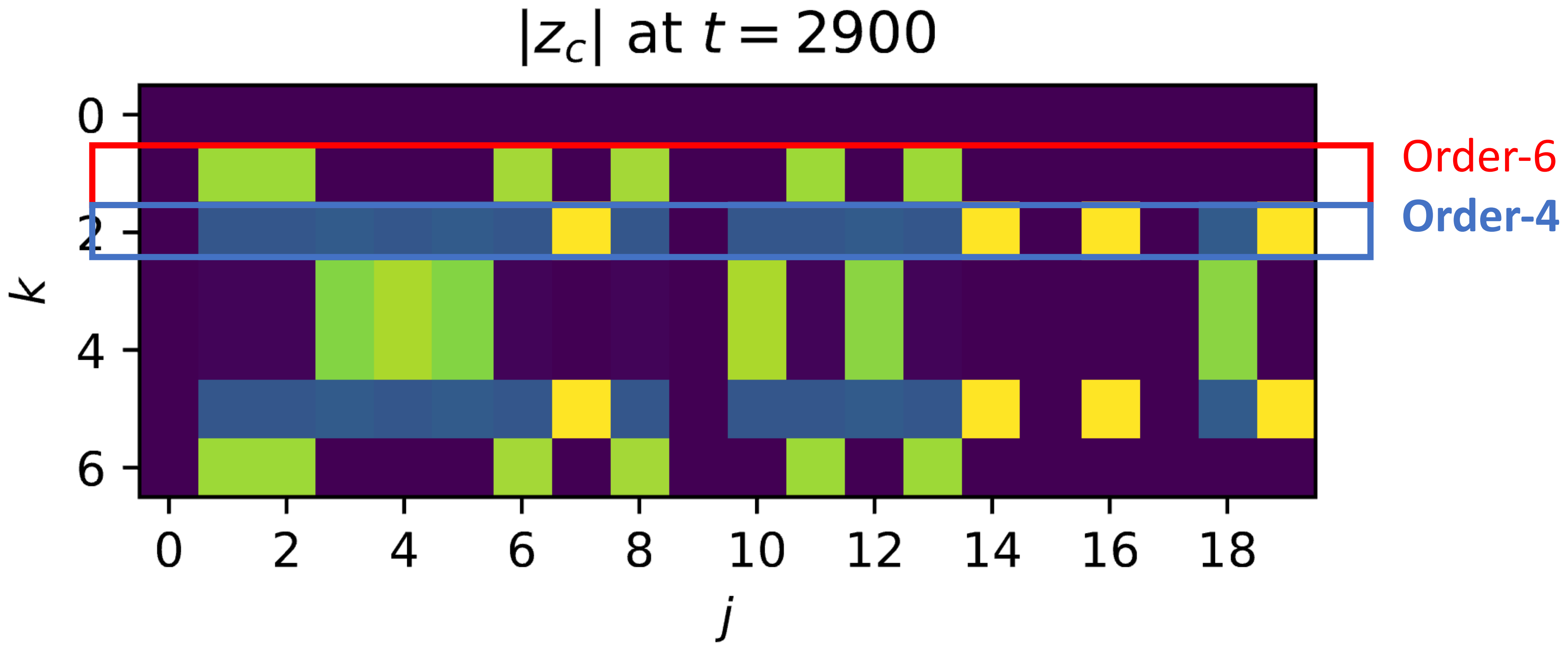
What a Gradient Descent Solution look like?



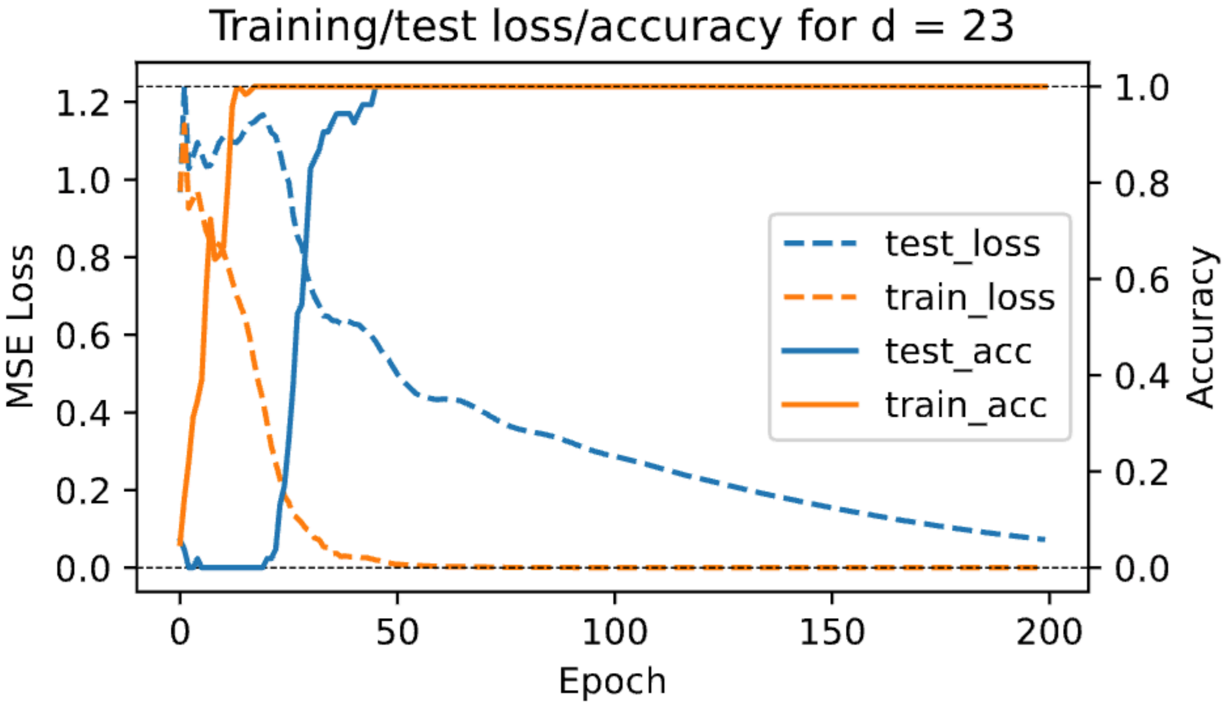
Order-6 solutions

Symmetry due to Hermitian condition

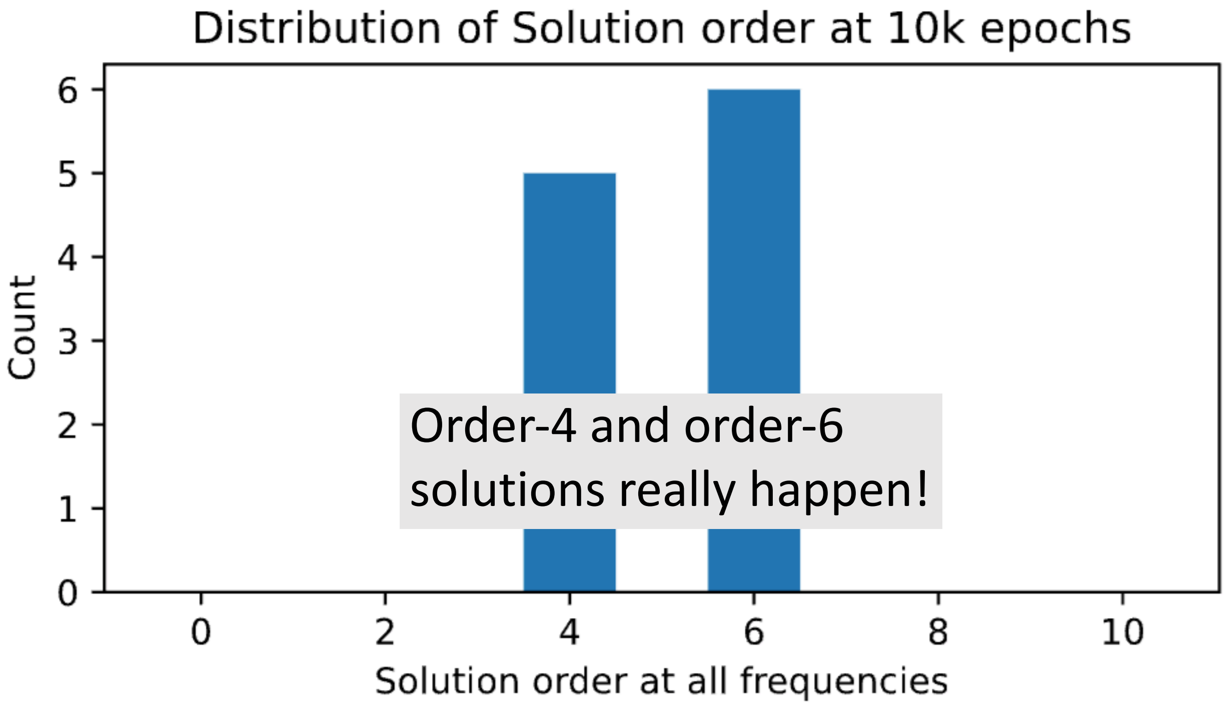
What a Gradient Descent Solution look like?



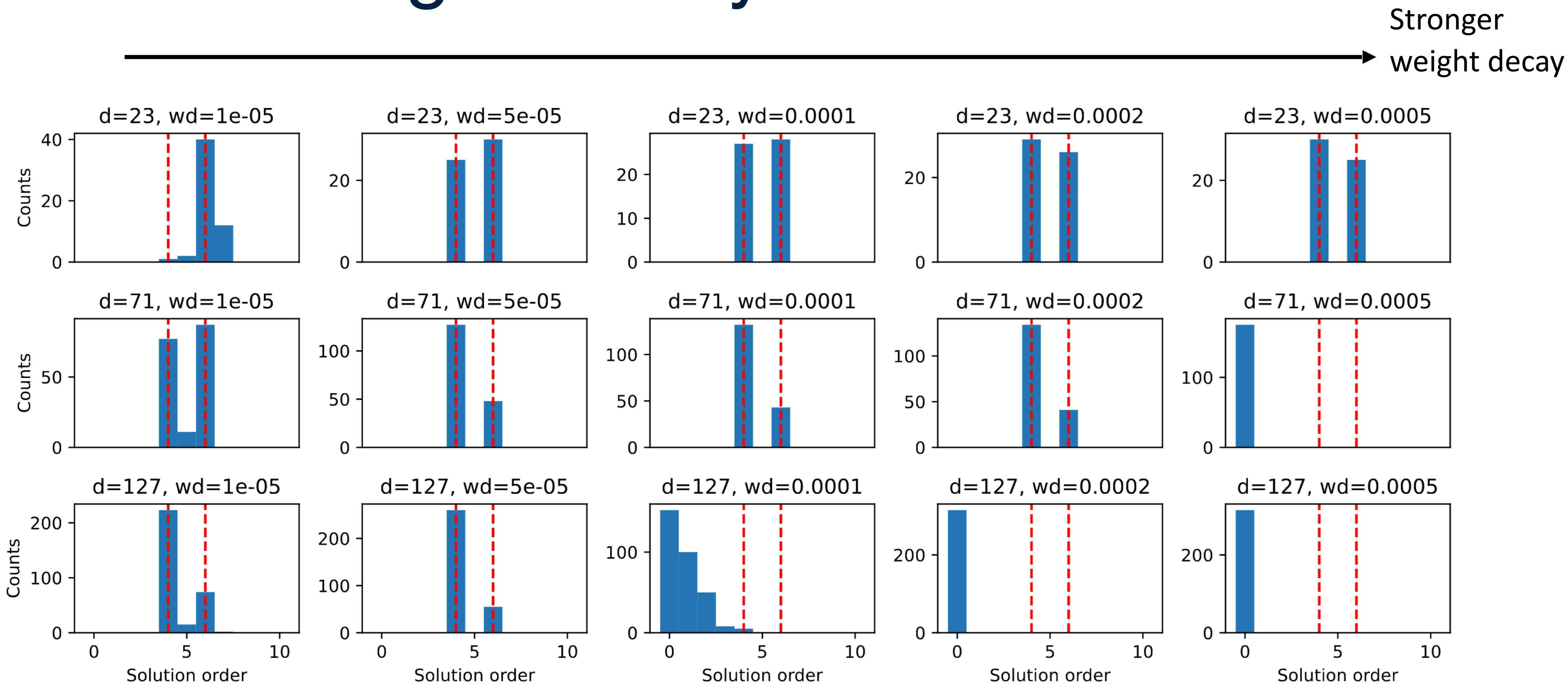
More Statistics on Gradient Descent Solutions



Grokking Behaviors!

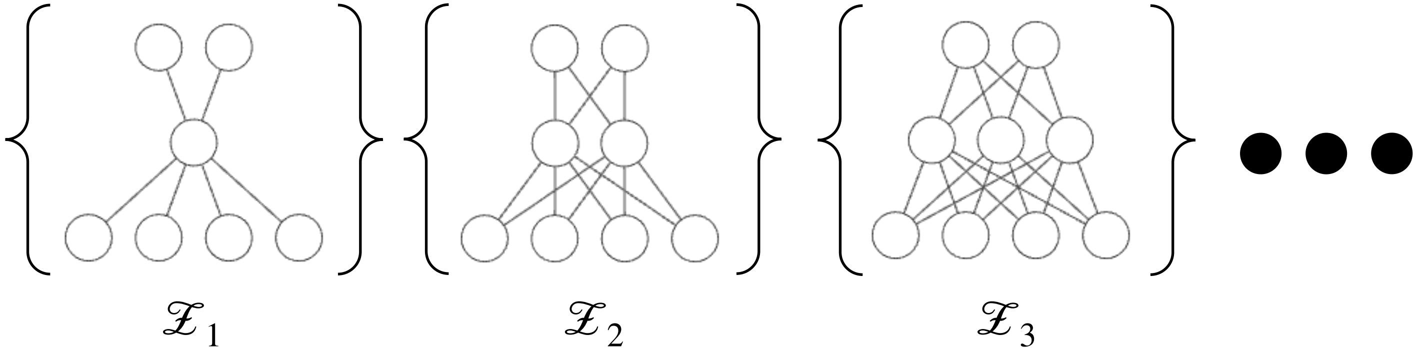


Effect of Weight Decay



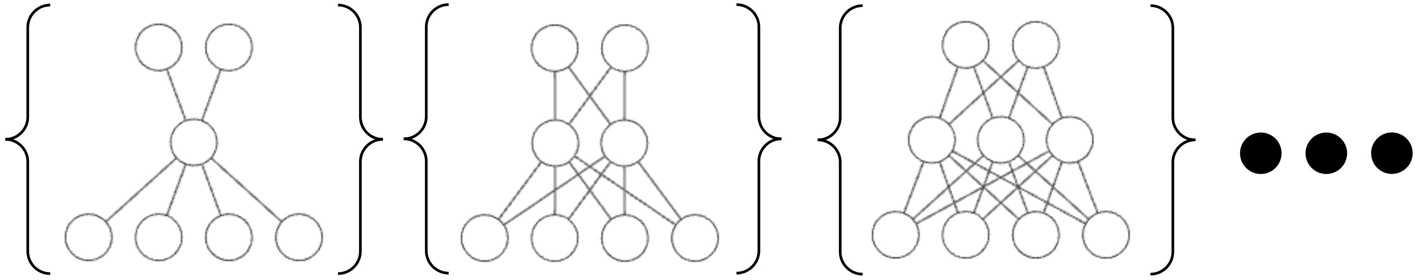
Why? 🤔

Nice *algebraic structures* exist for the solutions



$$\mathcal{Z} = \bigcup_{q \geq 0} \mathcal{Z}_q$$: All 2-layer networks with different number of hidden nodes

Nice *algebraic structures* exist for the solutions



$\mathcal{Z}_1$ $\mathcal{Z}_2$ $\mathcal{Z}_3$

$\mathcal{Z} = \bigcup_{q \geq 0} \mathcal{Z}_q$: All 2-layer networks with different number of hidden nodes

Ring addition $+$: Concatenate hidden nodes

Ring multiplication $*$: Kronecker production along the hidden dimensions

$\langle \mathcal{Z}, +, * \rangle$ is a ***semi-ring***

Ring Homomorphism

A function $r(\mathbf{z}) : \mathscr{Z} \mapsto \mathbb{C}$ is a *ring homomorphism*, if

- $r(\mathbf{1}) = 1$
- $r(\mathbf{z}_1 + \mathbf{z}_2) = r(\mathbf{z}_1) + r(\mathbf{z}_2)$
- $r(\mathbf{z}_1 * \mathbf{z}_2) = r(\mathbf{z}_1) r(\mathbf{z}_2)$

Ring Homomorphism

A function $r(\mathbf{z}) : \mathcal{Z} \mapsto \mathbb{C}$ is a *ring homomorphism*, if

- $r(\mathbf{1}) = 1$
- $r(\mathbf{z}_1 + \mathbf{z}_2) = r(\mathbf{z}_1) + r(\mathbf{z}_2)$
- $r(\mathbf{z}_1 * \mathbf{z}_2) = r(\mathbf{z}_1)r(\mathbf{z}_2)$

🎯 $r_{k_1k_2k}(z)$ and $r_{pk_1k_2k}(z)$ are ***ring homomorphisms***!

Ring Homomorphism

A function $r(\mathbf{z}) : \mathscr{Z} \mapsto \mathbb{C}$ is a *ring homomorphism*, if

- $r(\mathbf{1}) = 1$
- $r(\mathbf{z}_1 + \mathbf{z}_2) = r(\mathbf{z}_1) + r(\mathbf{z}_2)$
- $r(\mathbf{z}_1 * \mathbf{z}_2) = r(\mathbf{z}_1)r(\mathbf{z}_2)$

$r_{k_1k_2k}(\mathbf{z})$ and $r_{pk_1k_2k}(\mathbf{z})$ are ***ring homomorphisms***!

MSE Loss

$$\ell_k(\mathbf{z}) = -2r_{kkk} + \sum_{k_1k_2} \left|r_{k_1k_2k}\right|^2 + \frac{1}{4}\left|\sum_{p\in\{a,b\}}\sum_{k'} r_{p,k',-k',k}\right|^2 + \frac{1}{4}\sum_{m\neq0}\sum_{p\in\{a,b\}}\left|\sum_{k'} r_{p,k',m-k',k}\right|^2$$

Ring Homomorphism

A function $r(\mathbf{z}) : \mathscr{Z} \mapsto \mathbb{C}$ is a *ring homomorphism*, if

- $r(\mathbf{1}) = 1$
- $r(\mathbf{z}_1 + \mathbf{z}_2) = r(\mathbf{z}_1) + r(\mathbf{z}_2)$
- $r(\mathbf{z}_1 * \mathbf{z}_2) = r(\mathbf{z}_1) r(\mathbf{z}_2)$

🎯 $r_{k_1 k_2 k}(z)$ and $r_{pk_1 k_2 k}(z)$ are ***ring homomorphisms***!

MSE Loss

$$\ell_k(\mathbf{z}) = -2 r_{kkk} + \sum_{k_1 k_2} \left| r_{k_1 k_2 k} \right|^2 + \frac{1}{4} \left| \sum_{p \in \{a,b\}} \sum_{k'} r_{p,k',-k',k} \right|^2 + \frac{1}{4} \sum_{m \neq 0} \sum_{p \in \{a,b\}} \left| \sum_{k'} r_{p,k',m-k',k} \right|^2$$

Partial solution z_1 satisfies $r_{k_1 k_2 k}(z_1) = 0$

Partial solution z_2 satisfies $r_{pk',-k',k}(z_2) = 0$

Ring Homomorphism

A function $r(\mathbf{z}) : \mathscr{Z} \mapsto \mathbb{C}$ is a *ring homomorphism*, if

- $r(\mathbf{1}) = 1$
- $r(\mathbf{z}_1 + \mathbf{z}_2) = r(\mathbf{z}_1) + r(\mathbf{z}_2)$
- $r(\mathbf{z}_1 * \mathbf{z}_2) = r(\mathbf{z}_1)r(\mathbf{z}_2)$

🎯 $r_{k_1k_2k}(\mathbf{z})$ and $r_{pk_1k_2k}(\mathbf{z})$ are ***ring homomorphisms***!

MSE Loss

$$\ell_k(\mathbf{z}) = -2r_{kkk} + \sum_{k_1k_2}\left|r_{k_1k_2k}\right|^2 + \frac{1}{4}\left|\sum_{p\in\{a,b\}}\sum_{k'} r_{p,k',-k',k}\right|^2 + \frac{1}{4}\sum_{m\neq 0}\sum_{p\in\{a,b\}}\left|\sum_{k'} r_{p,k',m-k',k}\right|^2$$

Partial solution $\mathbf{z}_1$ satisfies $r_{k_1k_2k}(\mathbf{z}_1) = 0$

Partial solution $\mathbf{z}_2$ satisfies $r_{pk',-k',k}(\mathbf{z}_2) = 0$

$\mathbf{z} = \mathbf{z}_1 * \mathbf{z}_2$ satisfies both $r_{k_1k_2k}(\mathbf{z}) = r_{pk',-k',k}(\mathbf{z}) = 0$

Composing Global Solutions from Partial Ones

Partial Solution #1
$\mathbf{z}_{\text{syn}}^{(k)} \in R_{\text{c}} \cap R_{\text{n}}$ but $\mathbf{z}_{\text{syn}}^{(k)} \notin R_*$

Partial Solution #2
$\mathbf{z}_{\nu}^{(k)} \in R_*$

Composing Global Solutions from Partial Ones

Compositing
solutions using
ring multiplication $*$

Partial Solution #1

$\mathbf{z}_{\text{syn}}^{(k)} \in R_{\text{c}} \cap R_{\text{n}}$ but $\mathbf{z}_{\text{syn}}^{(k)} \notin R_*$

Better Solution

$\mathbf{z}_{\text{syn}}^{(k)} * \mathbf{z}_{\nu}^{(k)} \in R_{\text{c}} \cap R_{\text{n}} \cap R_*$

Partial Solution #2

$\mathbf{z}_{\nu}^{(k)} \in R_*$

[Y. Tian, *Composing Global Optimizers to Reasoning Tasks via Algebraic Objects in Neural Nets*, NeurIPS'25]

Composing Global Solutions from Partial Ones

Compositing solutions using ***ring multiplication*** $*$

Compositing solutions using ***ring addition*** $+$

Partial Solution #1

$\mathbf{z}_{\text{syn}}^{(k)} \in R_\text{c} \cap R_\text{n}$ but $\mathbf{z}_{\text{syn}}^{(k)} \notin R_*$

Partial Solution #2

$\mathbf{z}_{v}^{(k)} \in R_*$

Better Solution

$\mathbf{z}_{\text{syn}}^{(k)} * \mathbf{z}_{v}^{(k)} \in R_\text{c} \cap R_\text{n} \cap R_*$

Global Solutions to MSE loss!

$$\mathbf{z}_{F6} = \frac{1}{\sqrt[3]{6}} \sum_k \mathbf{z}_{\text{syn}}^{(k)} * \mathbf{z}_{v}^{(k)}$$

[Y. Tian, *Composing Global Optimizers to Reasoning Tasks via Algebraic Objects in Neural Nets*, NeurIPS'25]

Optimal solutions can be constructed

Order-6 z_{F6} (2*3)

$$\boldsymbol{z}_{F6} = \frac{1}{\sqrt[3]{6}} \sum_{k=1}^{(d-1)/2} \boldsymbol{z}_{\mathrm{syn}}^{(k)} * \boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{y}_k$$

Optimal solutions can be constructed

Order-6 z_{F6} (2*3)

$$\boldsymbol{z}_{F6} = \frac{1}{\sqrt[3]{6}} \sum_{k=1}^{(d-1)/2} \boldsymbol{z}_{\mathrm{syn}}^{(k)} * \boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{y}_k$$

Order-4 $z_{F4/6}$ (2*2)
(mixed with order-6)

$$\boldsymbol{z}_{F4/6} = \frac{1}{\sqrt[3]{6}} \hat{\boldsymbol{z}}_{F6}^{(k_0)} + \frac{1}{\sqrt[3]{4}} \sum_{k=1, k\neq k_0}^{(d-1)/2} \boldsymbol{z}_{F4}^{(k)}$$

Optimal solutions can be constructed

Order-6 z_{F6} (2*3)

$$\boldsymbol{z}_{F6} = \frac{1}{\sqrt[3]{6}} \sum_{k=1}^{(d-1)/2} \boldsymbol{z}_{\text{syn}}^{(k)} * \boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{y}_k$$

Order-4 $z_{F4/6}$ (2*2)
(mixed with order-6)

$$\boldsymbol{z}_{F4/6} = \frac{1}{\sqrt[3]{6}} \hat{\boldsymbol{z}}_{F6}^{(k_0)} + \frac{1}{\sqrt[3]{4}} \sum_{k=1, k\neq k_0}^{(d-1)/2} \boldsymbol{z}_{F4}^{(k)}$$

Perfect memorization
(order-d per frequency)

$$\boldsymbol{z}_a = \sum_{j=0}^{d-1} \boldsymbol{u}_a^j, \qquad \boldsymbol{z}_b = \sum_{j=0}^{d-1} \boldsymbol{u}_b^j$$

$$\boldsymbol{z}_M = d^{-2/3} \boldsymbol{z}_a * \boldsymbol{z}_b$$

Gradient Descent solutions matches with construction

d	%not	%non-factorable		error ($\times 10^{-2}$)		solution distribution (%) in factorable ones			
	order-4/6	order-4	order-6	order-4	order-6	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\xi}$	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\mathrm{syn},\alpha\beta}$	$\boldsymbol{z}^{(k)}_{\nu} * \boldsymbol{z}^{(k)}_{\mathrm{syn}}$	others
23	0.0 ± 0.0	0.00 ± 0.00	5.71 ± 5.71	0.05 ± 0.01	4.80 ± 0.96	47.07 ± 1.88	11.31 ± 1.76	39.80 ± 2.11	1.82 ± 1.82
71	0.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	5.02 ± 0.25	72.57 ± 0.70	4.00 ± 1.14	21.14 ± 2.14	2.29 ± 1.07
127	0.0 ± 0.0	1.50 ± 0.92	0.00 ± 0.00	0.26 ± 0.14	0.93 ± 0.18	82.96 ± 0.39	2.25 ± 0.64	14.13 ± 0.87	0.66 ± 0.66

$$q = 512,\ wd = 5 \cdot 10^{-5}$$

Gradient Descent solutions matches with construction

d	%not order-4/6	%non-factorable order-4	%non-factorable order-6	error ($\times 10^{-2}$) order-4	error ($\times 10^{-2}$) order-6	solution distribution (%) in factorable ones $\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\xi}^{(k)}$	$\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\mathrm{syn},\alpha\beta}^{(k)}$	$\boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{z}_{\mathrm{syn}}^{(k)}$	others
23	$0.0_{\pm 0.0}$	$0.00_{\pm 0.00}$	$5.71_{\pm 5.71}$	$0.05_{\pm 0.01}$	$4.80_{\pm 0.96}$	$47.07_{\pm 1.88}$	$11.31_{\pm 1.76}$	$39.80_{\pm 2.11}$	$1.82_{\pm 1.82}$
71	$0.0_{\pm 0.0}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.03_{\pm 0.00}$	$5.02_{\pm 0.25}$	$72.57_{\pm 0.70}$	$4.00_{\pm 1.14}$	$21.14_{\pm 2.14}$	$2.29_{\pm 1.07}$
127	$0.0_{\pm 0.0}$	$1.50_{\pm 0.92}$	$0.00_{\pm 0.00}$	$0.26_{\pm 0.14}$	$0.93_{\pm 0.18}$	$82.96_{\pm 0.39}$	$2.25_{\pm 0.64}$	$14.13_{\pm 0.87}$	$0.66_{\pm 0.66}$

100% of the per-freq solutions are order-4/6

Gradient Descent solutions matches with construction

d	%not	%non-factorable		error ($\times 10^{-2}$)		solution distribution (%) in factorable ones			
	order-4/6	order-4	order-6	order-4	order-6	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\xi}$	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\mathrm{syn},\alpha\beta}$	$\boldsymbol{z}^{(k)}_{\nu} * \boldsymbol{z}^{(k)}_{\mathrm{syn}}$	others
23	0.0 ± 0.0	0.00 ± 0.00	5.71 ± 5.71	0.05 ± 0.01	4.80 ± 0.96	47.07 ± 1.88	11.31 ± 1.76	39.80 ± 2.11	1.82 ± 1.82
71	0.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	5.02 ± 0.25	72.57 ± 0.70	4.00 ± 1.14	21.14 ± 2.14	2.29 ± 1.07
127	0.0 ± 0.0	1.50 ± 0.92	0.00 ± 0.00	0.26 ± 0.14	0.93 ± 0.18	82.96 ± 0.39	2.25 ± 0.64	14.13 ± 0.87	0.66 ± 0.66

95% of the solutions are
factorizable into “2*3” or “2*2”

Gradient Descent solutions matches with construction

d	%not	%non-factorable		error ($\times 10^{-2}$)		solution distribution (%) in factorable ones			
	order-4/6	order-4	order-6	order-4	order-6	$\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\xi}^{(k)}$	$\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\mathrm{syn},\alpha\beta}^{(k)}$	$\boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{z}_{\mathrm{syn}}^{(k)}$	others
23	0.0 ± 0.0	0.00 ± 0.00	5.71 ± 5.71	0.05 ± 0.01	4.80 ± 0.96	47.07 ± 1.88	11.31 ± 1.76	39.80 ± 2.11	1.82 ± 1.82
71	0.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	5.02 ± 0.25	72.57 ± 0.70	4.00 ± 1.14	21.14 ± 2.14	2.29 ± 1.07
127	0.0 ± 0.0	1.50 ± 0.92	0.00 ± 0.00	0.26 ± 0.14	0.93 ± 0.18	82.96 ± 0.39	2.25 ± 0.64	14.13 ± 0.87	0.66 ± 0.66

Factorization error is very small

Gradient Descent solutions matches with construction

d	%not	%non-factorable		error ($\times 10^{-2}$)		solution distribution (%) in factorable ones			
	order-4/6	order-4	order-6	order-4	order-6	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\xi}$	$\boldsymbol{z}^{(k)}_{\nu=\mathrm{i}} * \boldsymbol{z}^{(k)}_{\mathrm{syn},\alpha\beta}$	$\boldsymbol{z}^{(k)}_{\nu} * \boldsymbol{z}^{(k)}_{\mathrm{syn}}$	others
23	$0.0_{\pm 0.0}$	$0.00_{\pm 0.00}$	$5.71_{\pm 5.71}$	$0.05_{\pm 0.01}$	$4.80_{\pm 0.96}$	$47.07_{\pm 1.88}$	$11.31_{\pm 1.76}$	$39.80_{\pm 2.11}$	$1.82_{\pm 1.82}$
71	$0.0_{\pm 0.0}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.03_{\pm 0.00}$	$5.02_{\pm 0.25}$	$72.57_{\pm 0.70}$	$4.00_{\pm 1.14}$	$21.14_{\pm 2.14}$	$2.29_{\pm 1.07}$
127	$0.0_{\pm 0.0}$	$1.50_{\pm 0.92}$	$0.00_{\pm 0.00}$	$0.26_{\pm 0.14}$	$0.93_{\pm 0.18}$	$82.96_{\pm 0.39}$	$2.25_{\pm 0.64}$	$14.13_{\pm 0.87}$	$0.66_{\pm 0.66}$

98% of the solutions can be factorizable into the constructed forms

Gradient Descent solutions matches with construction

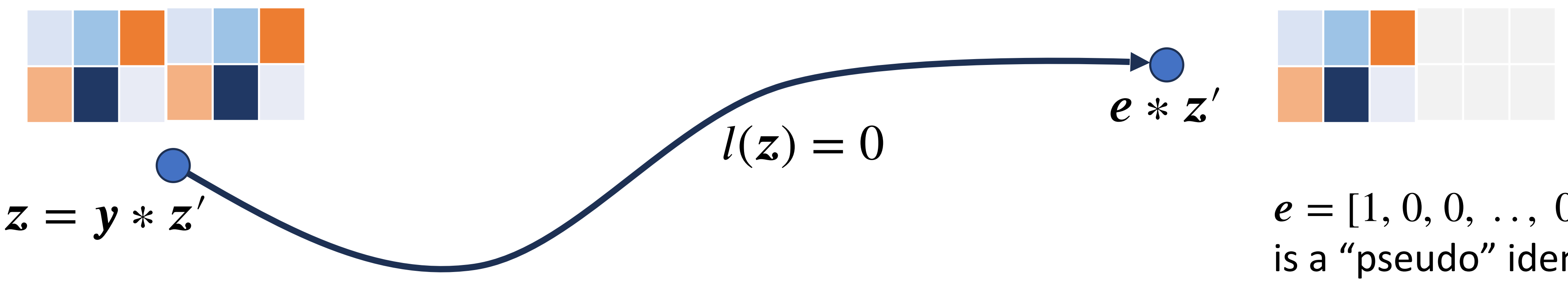
d	%not	%non-factorable		error ($\times 10^{-2}$)		solution distribution (%) in factorable ones			
	order-4/6	order-4	order-6	order-4	order-6	$\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\xi}^{(k)}$	$\boldsymbol{z}_{\nu=\mathrm{i}}^{(k)} * \boldsymbol{z}_{\mathrm{syn},\alpha\beta}^{(k)}$	$\boldsymbol{z}_{\nu}^{(k)} * \boldsymbol{z}_{\mathrm{syn}}^{(k)}$	others
23	[illegible]	[illegible]	[illegible]	[illegible]	± 0.96	47.07 ± 1.88	11.31 ± 1.76	39.80 ± 2.11	1.82 ± 1.82
71	[illegible]	[illegible]	[illegible]	[illegible]	± 0.25	72.57 ± 0.70	4.00 ± 1.14	21.14 ± 2.14	2.29 ± 1.07
127	[illegible]	[illegible]	[illegible]	[illegible]	± 0.18	82.96 ± 0.39	2.25 ± 0.64	14.13 ± 0.87	0.66 ± 0.66

98% of the solutions can be factorizable into the constructed forms

Emerging Symbolic structure from neural network training

How about Gradient Dynamics?

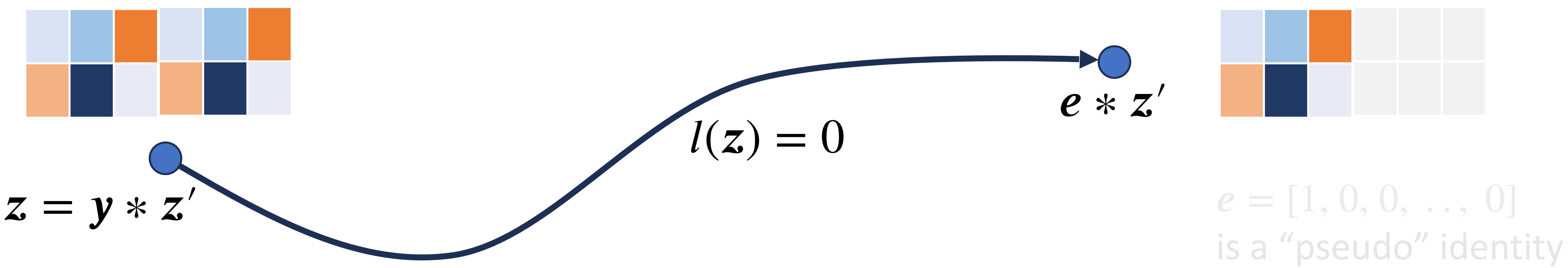
Theorem [**The Occam's Razer**] If $\boldsymbol{z} = \boldsymbol{y} * \boldsymbol{z}'$ and both $\boldsymbol{z}$ and $\boldsymbol{z}'$ are global optimal, then there exists a path of zero loss connecting $\boldsymbol{z}$ and $\boldsymbol{z}'$.



$\boldsymbol{e} = [1, 0, 0, \ldots, 0]$
is a "pseudo" identity

How about Gradient Dynamics?

Theorem [**The Occam's Razer**] If $\boldsymbol{z} = \boldsymbol{y} * \boldsymbol{z}'$ and both $\boldsymbol{z}$ and $\boldsymbol{z}'$ are global optimal, then there exists a path of zero loss connecting $\boldsymbol{z}$ and $\boldsymbol{z}'$.



L2 regularization will push the solution to $\boldsymbol{e} * \boldsymbol{z}'$ (simpler solutions), since $\|\boldsymbol{e} * \boldsymbol{z}'\|_2 \leq \|\boldsymbol{y} * \boldsymbol{z}'\|_2$

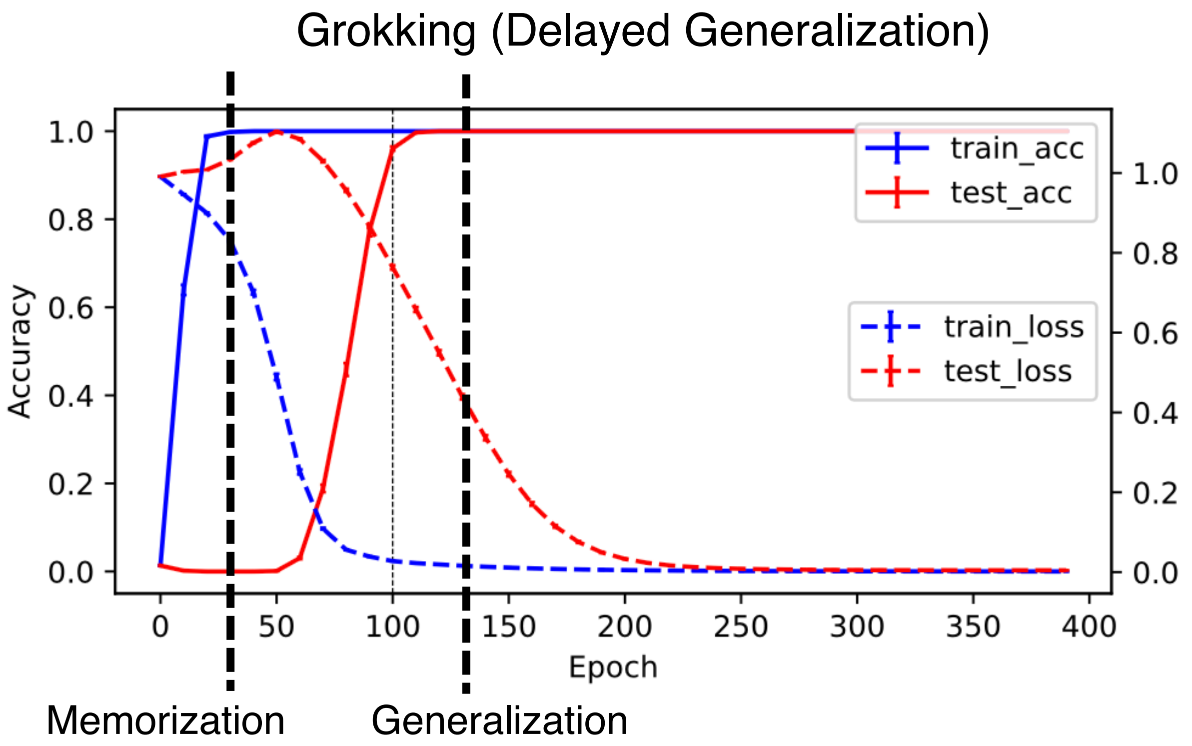
Limitation of the analysis

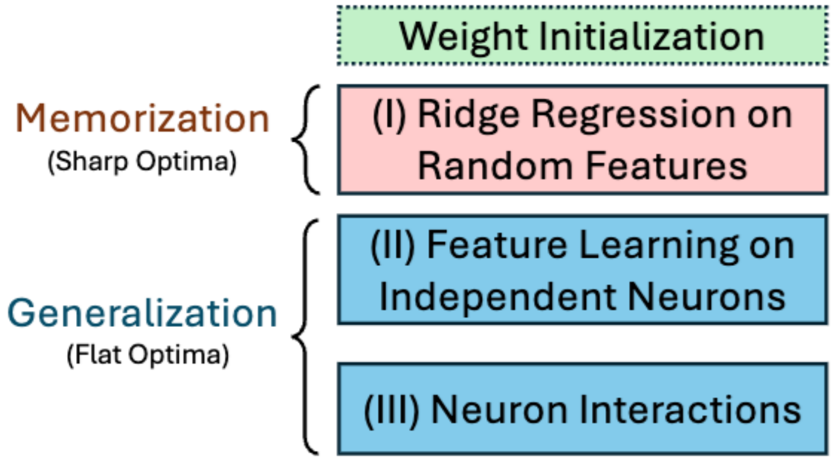
How such solutions are achieved — No analysis with gradient dynamics

Only apply to a combination of MSE loss + all data + quadratic activation

Next work is to **crack the grokking behaviors from gradient dynamics**

Understanding Grokking Behavior

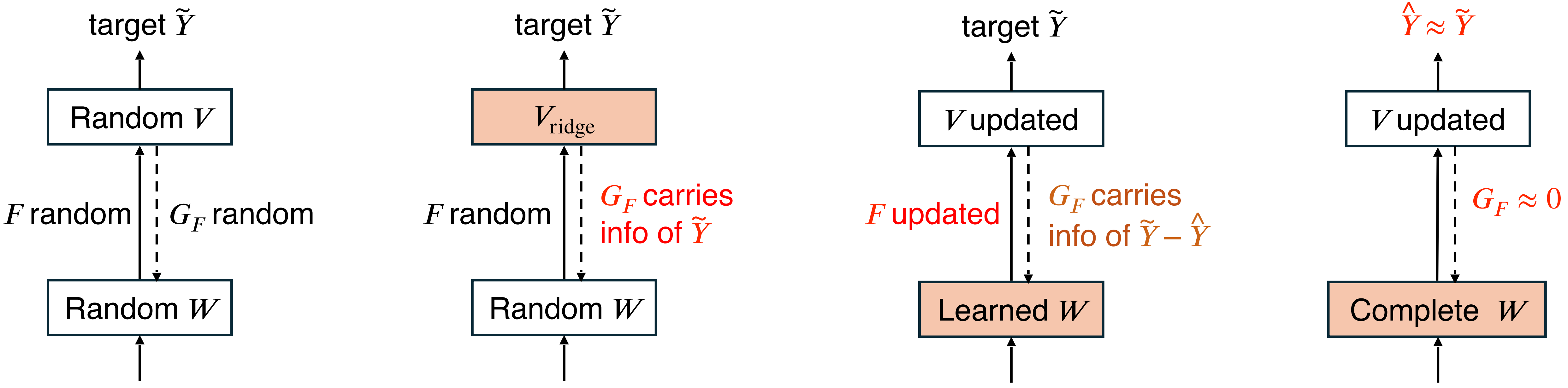




[Y. Tian, *Provable Scaling Laws of Feature Emergence from Learning Dynamics of Grokking*, arXiv'25]

Stages of Grokking Behaviors

$$\hat{Y} = \sigma(XW)V$$



(a) Initialization | *(b) After Lazy learning* | *(c) After Independent Feature Learning* | *(d) After Interactive Feature learning*

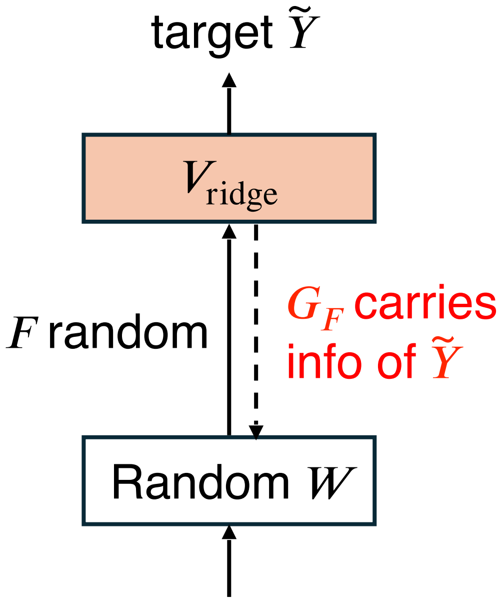
Stages I: NTK regime (Lazy Learning)

Objective function ($Y \in \mathbb{R}^{n\times M}, \quad X \in \mathbb{R}^{n\times d}$)

$$\min_{V,W} \frac{1}{2}\|P_1^{\perp}(Y - \sigma(XW)V)\|_F^2$$

Ridge Regression (with weight decay η)

$$V_{\text{ridge}} = (\tilde{F}^\top \tilde{F} + \eta I)^{-1}\tilde{F}^\top \tilde{Y}$$



$$\tilde{Y} = P_1^{\perp} Y$$

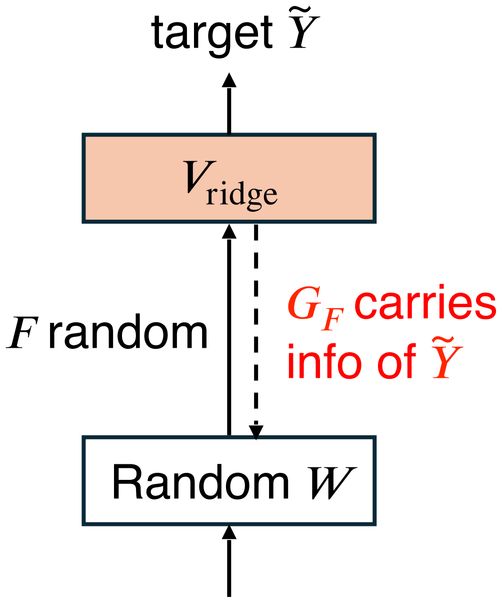
$$F = \sigma(XW) \quad \tilde{F} = P_1^{\perp} F$$

Here $P_1^{\perp} = I_n - \mathbf{1}\mathbf{1}^\top/n$ is the zero-mean projection

The backpropagated Gradient G_F

At Ridge regression solution V_{ridge}

$$G_F = \eta(\tilde{F}\tilde{F}^\top + \eta I)^{-1}\tilde{Y}\tilde{Y}^\top \tilde{F}(\tilde{F}^\top\tilde{F} + \eta I)^{-1}$$



Looks complicated... any interesting properties?

The backpropagated Gradient G_F

Lemma 1 (Structure of backpropagated gradient G_F). *Assume that (1) entries of W follow normal distribution, (2) $\|\mathbf{x}_i\|_2$ have the same norm, (3) $\mathbf{x}_i^\top \mathbf{x}_{i'} = \rho$ for all $i \neq i'$ and (4) large width K, then both $\tilde{F}^\top \tilde{F}$ and $\tilde{F}\tilde{F}^\top$ becomes a multiple of identity and Eqn. 5 becomes:*

$$G_F = \frac{\eta}{(Kc_1 + \eta)(nc_2 + \eta)} \tilde{Y}\tilde{Y}^\top F \tag{6}$$

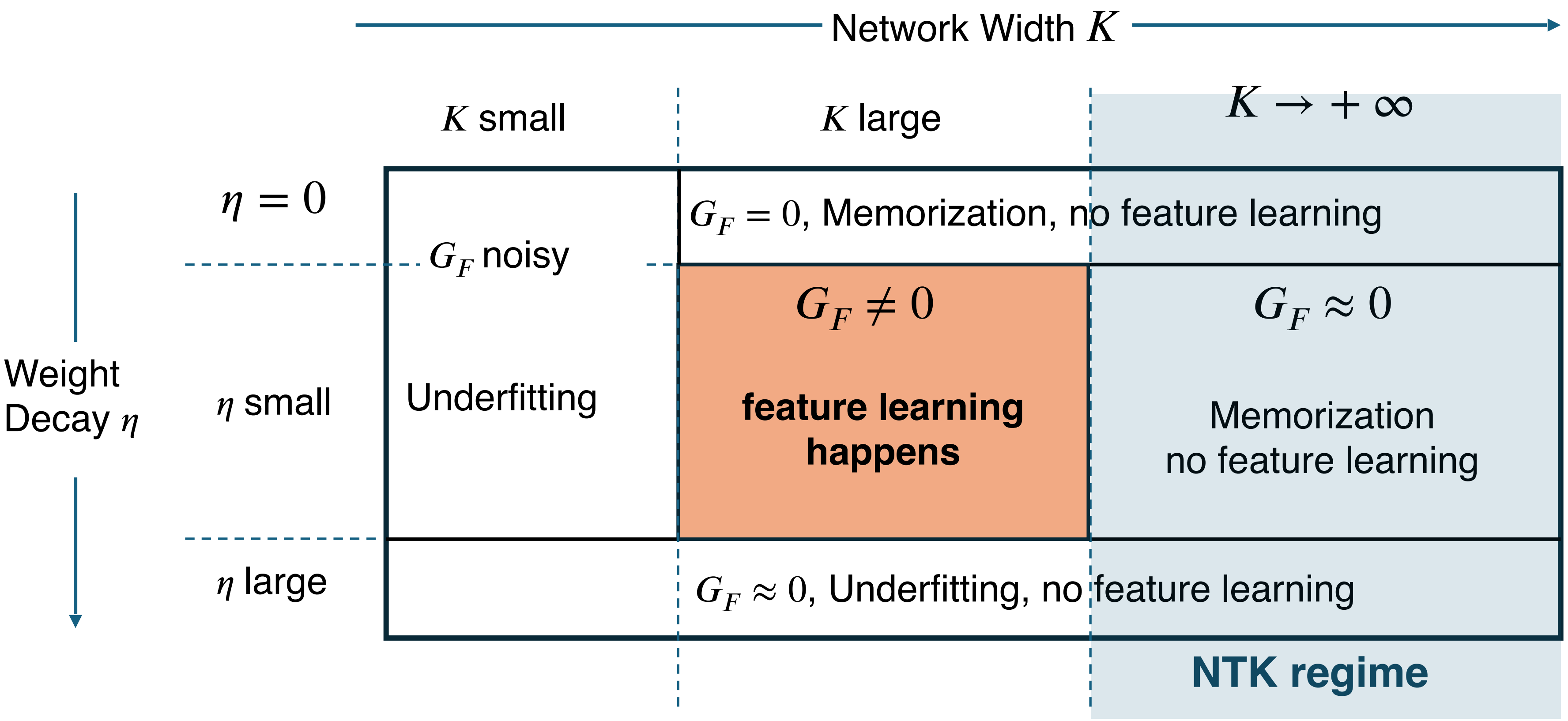
Key insights

If weight decay $\eta = 0$, then $G_F = 0$ (no feature learning)

If weight decay is large, then $G_F \to 0$

If number of hidden nodes $K \to +\infty$, then $G_F \to 0$ (NTK regime)

The regime that feature learning happens



Stage II: The Energy Function $\mathscr{E}(\mathbf{w})$

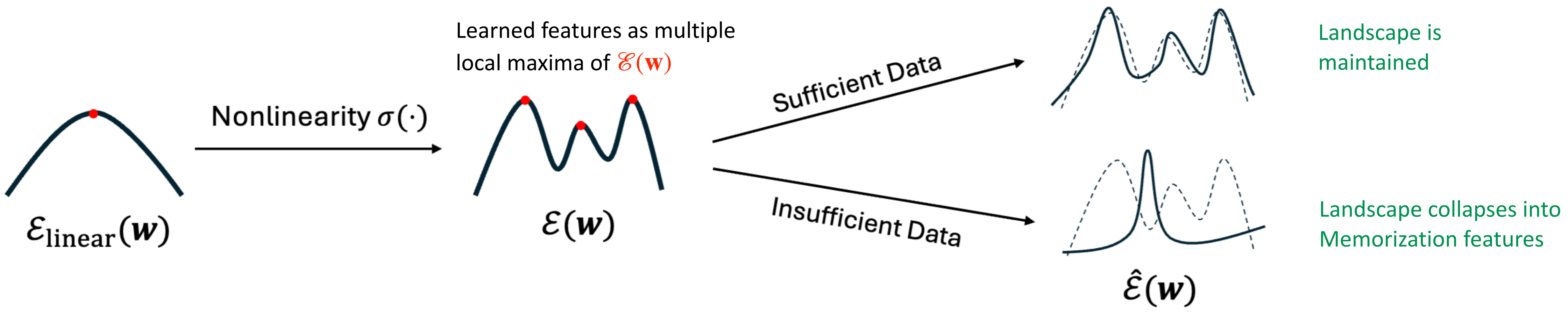
$$G_F \propto \eta \tilde{Y}\tilde{Y}^\top F \xrightarrow{\text{Component-wise dynamics}} \dot{\mathbf{w}}_j = X^\top D_j \mathbf{g}_j, \quad \mathbf{g}_j \propto \eta \tilde{Y}\tilde{Y}^\top \sigma(X\mathbf{w}_j)$$

Theorem 1 (The energy function $\mathcal{E}$ for independent feature learning). *The dynamics (Eqn. 7) of independent feature learning is exactly the gradient ascent dynamics of the energy function $\mathcal{E}$ w.r.t. $\mathbf{w}_j$, a nonlinear canonical-correlation analysis (CCA) between the input X and target $\tilde{Y}$:*

$$\mathcal{E}(\mathbf{w}_j) = \frac{1}{2}\|\tilde{Y}^\top \sigma(X\mathbf{w}_j)\|_2^2 \tag{8}$$

Connect Emerging Features with Data / Architecture

We discover that there exists an energy function $\mathcal{E}(\mathbf{w})$ that governs the feature learning process



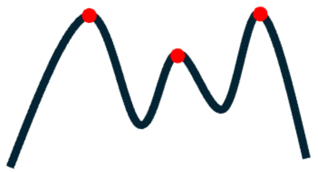
Group Representation Theory

The decomposition of group representation. The representation theory of finite group (Fulton & Harris, 2013; Steinberg, 2009) says that the regular representation R_h admits a decomposition into complex *irreducible* representations (or ***irreps***):

$$R_h = Q \left(\bigoplus_{k=0}^{\kappa(H)} \bigoplus_{r=1}^{m_k} C_k(h) \right) Q^* \tag{9}$$

where $\kappa(H)$ is the number of nontrivial irreps (i.e., not all h map to identity), $C_k(h) \in \mathbb{C}^{d_k \times d_k}$ is the k-th irrep block of R_h, Q is the unitary matrix (and Q^* is its conjugate transpose) and m_k is the multiplicity of the k-th irrep. This means that in the decomposition of R_h, there are m_k copies of d_k-dimensional irrep, and these copies are isomorphic to each other. So the k-th ***irrep subspace*** $\mathcal{H}_k$ has dimension $m_k d_k$.

Emerging Features are Symbolic!



$\mathcal{E}(\boldsymbol{w})$

$$\mathcal{E}(\mathbf{w}) = \frac{1}{2}\sum_h \langle \tilde{R}_h, S\rangle_F^2 = \frac{M}{2}\sum_{k\neq 0}\frac{1}{d_k}\left|\sum_r \mathrm{tr}(\hat{S}_{k,r})\right|^2$$

Theorem 2 (Local maxima of $\mathcal{E}$ for group input). *For group arithmetics tasks with $\sigma(x) = x^2$, $\mathcal{E}$ has multiple local maxima $\mathbf{w}^* = [\mathbf{u}; \pm P\mathbf{u}]$. Either it is in a real irrep of dimension d_k (with $\mathcal{E}^* = M/8d_k$ and $\mathbf{u} \in \mathcal{H}_k$), or in a pair of complex irrep of dimension d_k (with $\mathcal{E}^* = M/16d_k$ and $\mathbf{u} \in \mathcal{H}_k \oplus \mathcal{H}_{\bar{k}}$). These local maxima are not connected. No other local maxima exist.*

What is that specifically?

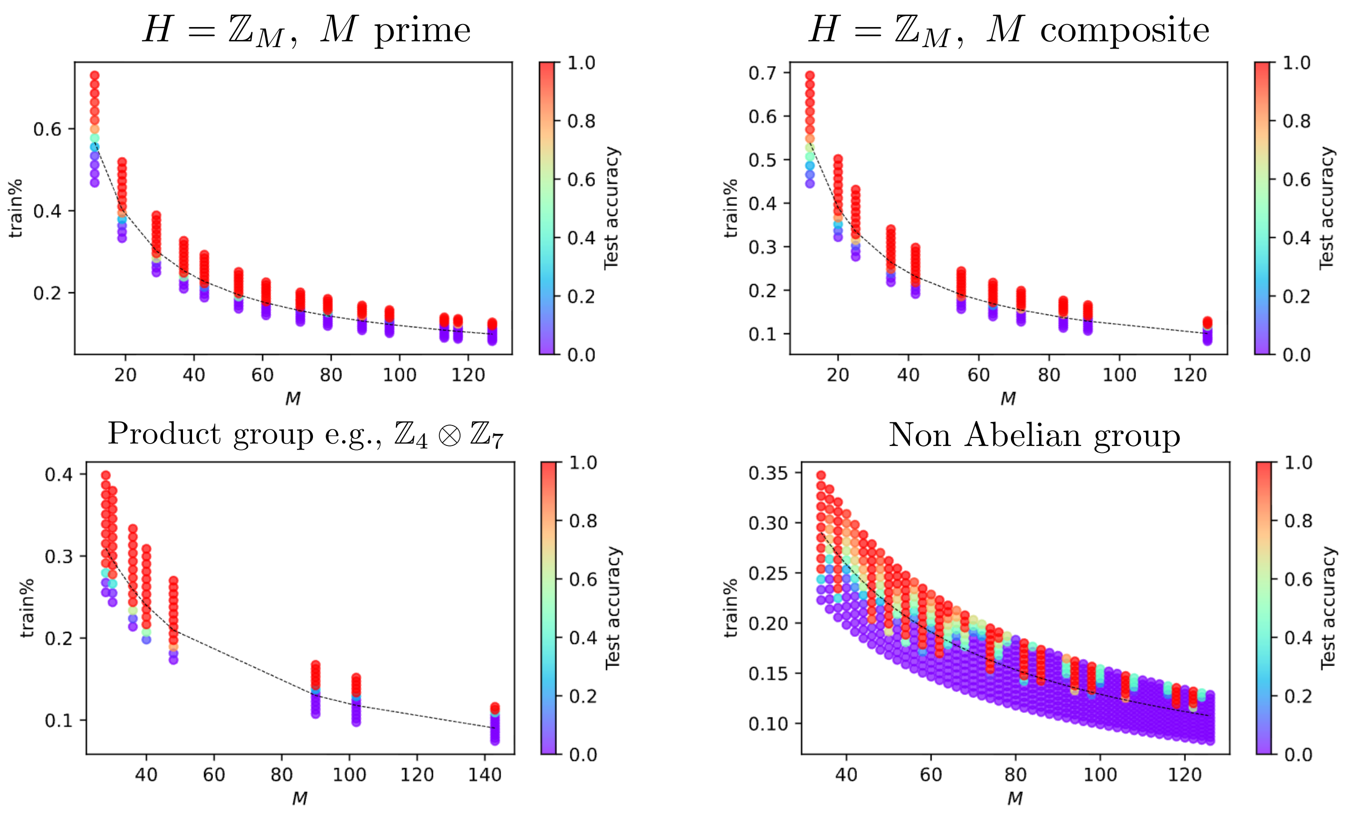
Corollary 2 (Modular addition). *For modular addition with odd M, all local maxima are single frequency $\mathbf{u}_k = a_k[\cos(km\omega)]_{m=0}^{M-1} + b_k[\sin(km\omega)]_{m=0}^{M-1}$ where $\omega := 2\pi/M$ with $\mathcal{E}^* = M/16$. For even M, $\mathbf{u}_{M/2} \propto [(-1)^m]_{m=0}^{M-1}$ has $\mathcal{E}^* = M/8$. Different local maxima are disconnected.*

Emerging Features are More Efficient!

Theorem 3 (Target Reconstruction). *Assume (1) $\mathcal{E}$ is optimized in complex domain $\mathbb{C}$, (2) for each irrep k, there are $m_k^2 d_k^2$ pairs of learned weights $\mathbf{w} = [\mathbf{u}; \pm P\mathbf{u}]$ whose associated rank-1 matrices $\{\mathbf{u}\mathbf{u}^*\}$ form a complete bases for $\mathcal{H}_k$ and (3) the top layer V also learns with $\eta = 0$, then $\hat{Y} = \tilde{Y}$.*

From the theorem, we know that $K = 2\sum_{k\neq 0} m_k^2 d_k^2 \leq 2\left[(M - \kappa(H))^2 + \kappa(H) - 1\right]$ suffice. In particular, for Abelian group, $\kappa(H) = M - 1$ and $\boxed{K = 2M - 2.}$ This is much more efficient than a pure memorization solution that would require M^2 nodes, i.e., each node memorizes a single pair $(h_1, h_2) \in H \times H$.

How much is **sufficient**? Provable Scaling Laws

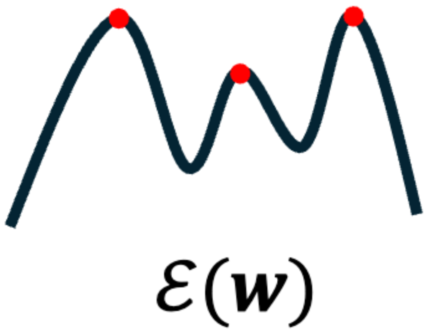


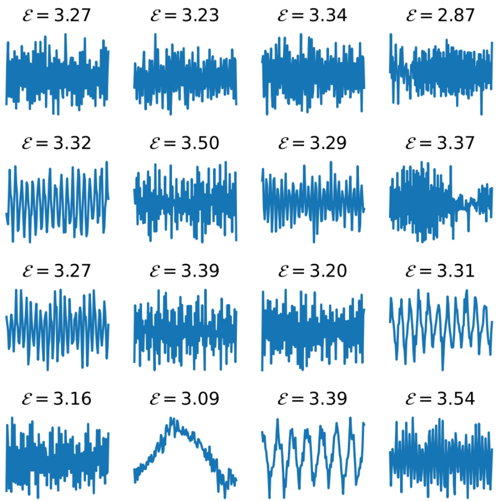
$O(M \log M)$

Theorem 4 (Amount of samples to maintain local optima). *If we select $n \gtrsim d_k^2 M \log(M/\delta)$ data sample from $H \times H$ uniformly at random, then with probability at least $1-\delta$, the empirical energy function $\hat{\mathcal{E}}$ keeps local maxima for d_k-dimensional irreps (Thm. 2).*

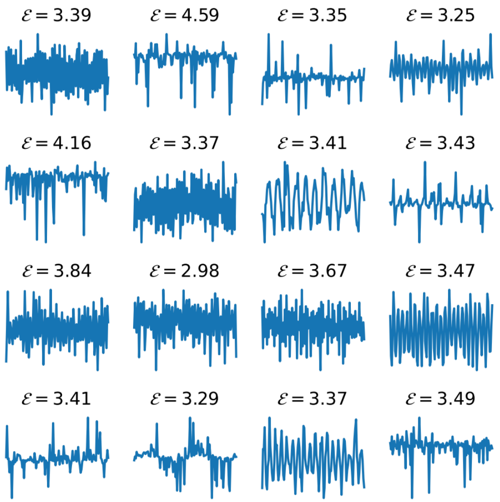
Boundary between Memorization and Generalization

At the boundary,
large Learning rate leads
to memorization (higher $\mathscr{E}$)

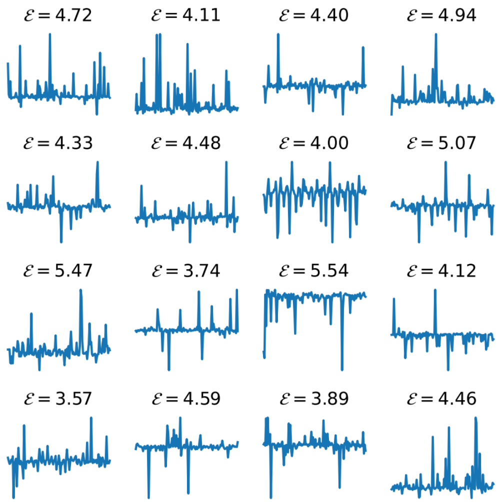




Small learning rate 0.001

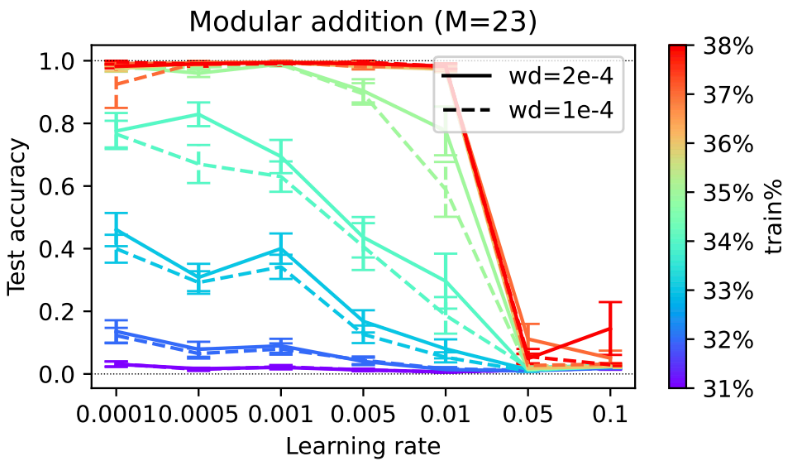


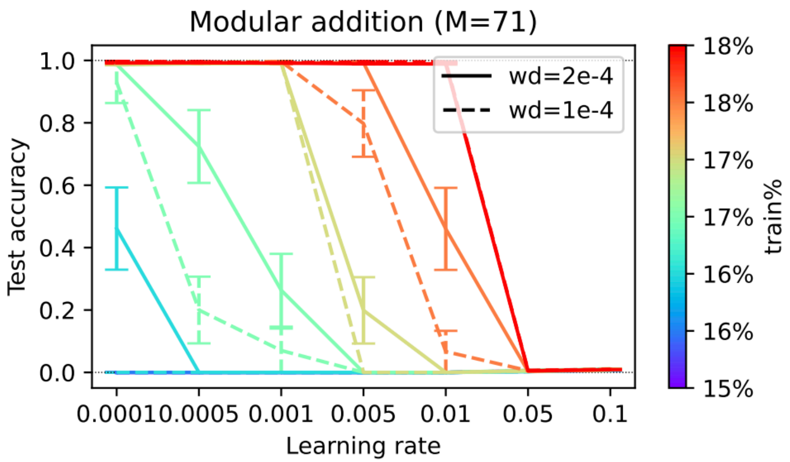
Mid learning rate 0.002

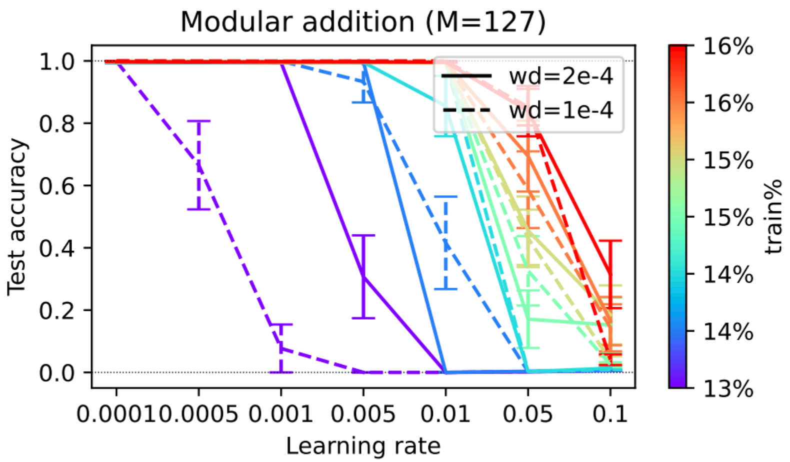


Large learning rate 0.005

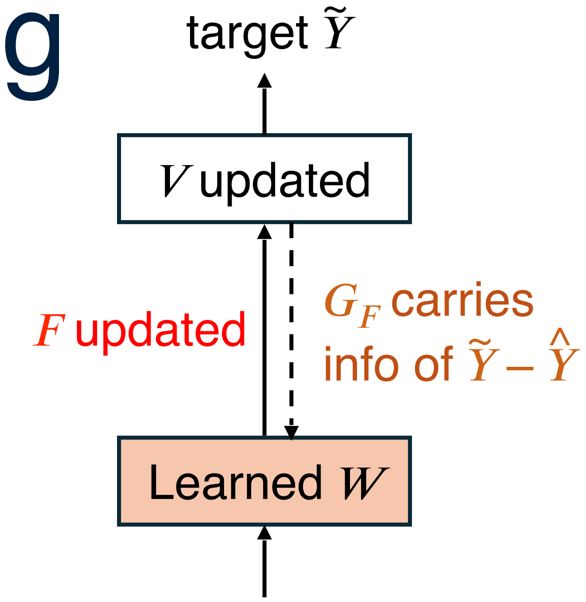
Boundary between Memorization and Generalization







Stage III: Interactive Feature Learning



Theorem 7 (Top-down Modulation). *For group arithmetic tasks with $\sigma(x) = x^2$, if the hidden layer learns only a subset $\mathcal{S}$ of irreps, then the backpropagated gradient $G_F \propto (\Phi_{\mathcal{S}} \otimes \mathbf{1}_M)(\Phi_{\mathcal{S}} \otimes \mathbf{1}_M)^* F$ (see proof for the definition of $\Phi_{\mathcal{S}}$), which yields a modified $\mathcal{E}_{\mathcal{S}}$ that only has local maxima on the missing irreps $k \notin \mathcal{S}$.*

Possible Implications

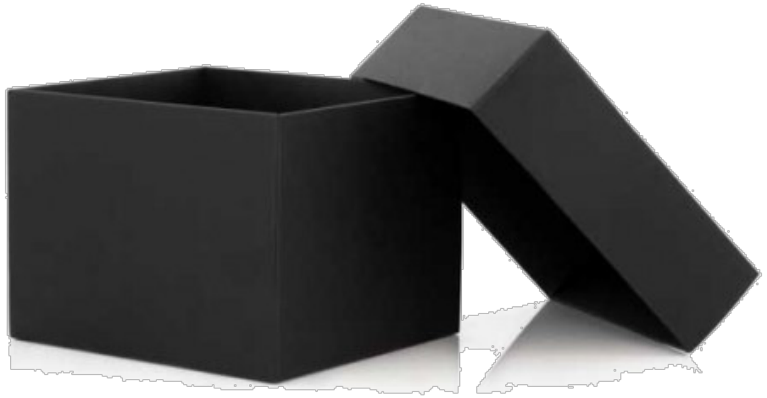
Do neural networks end up learning more efficient ***symbolic representations*** *that we don't know?*

Does gradient descent lead to a solution that can be reached by ***advanced algebraic operations****?*

Will gradient descent become ***obsolete****, eventually?*

Here this work is just a tiny step.

Next Step: Scale it to more complicated tasks and architectures



Thanks!