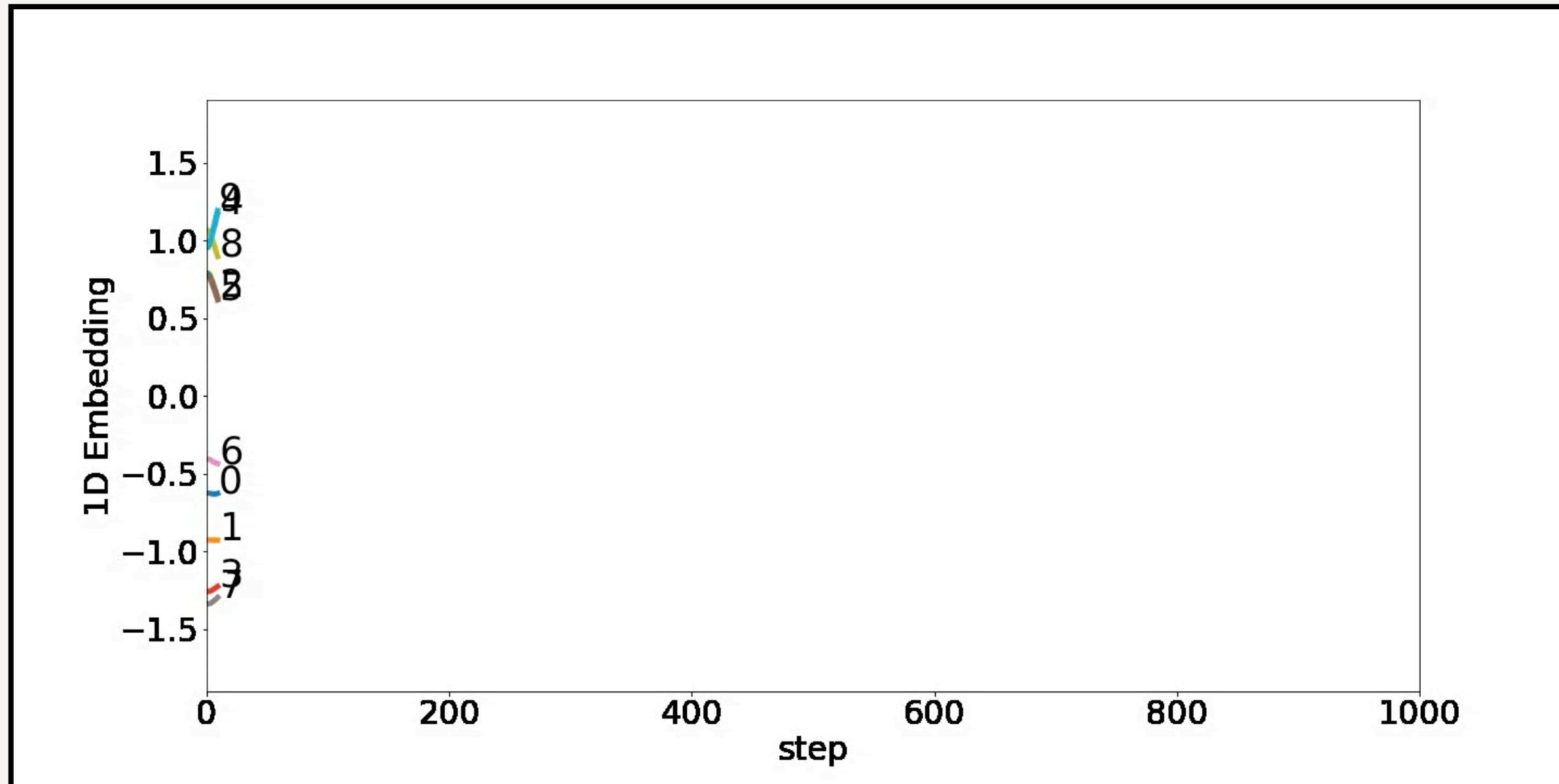


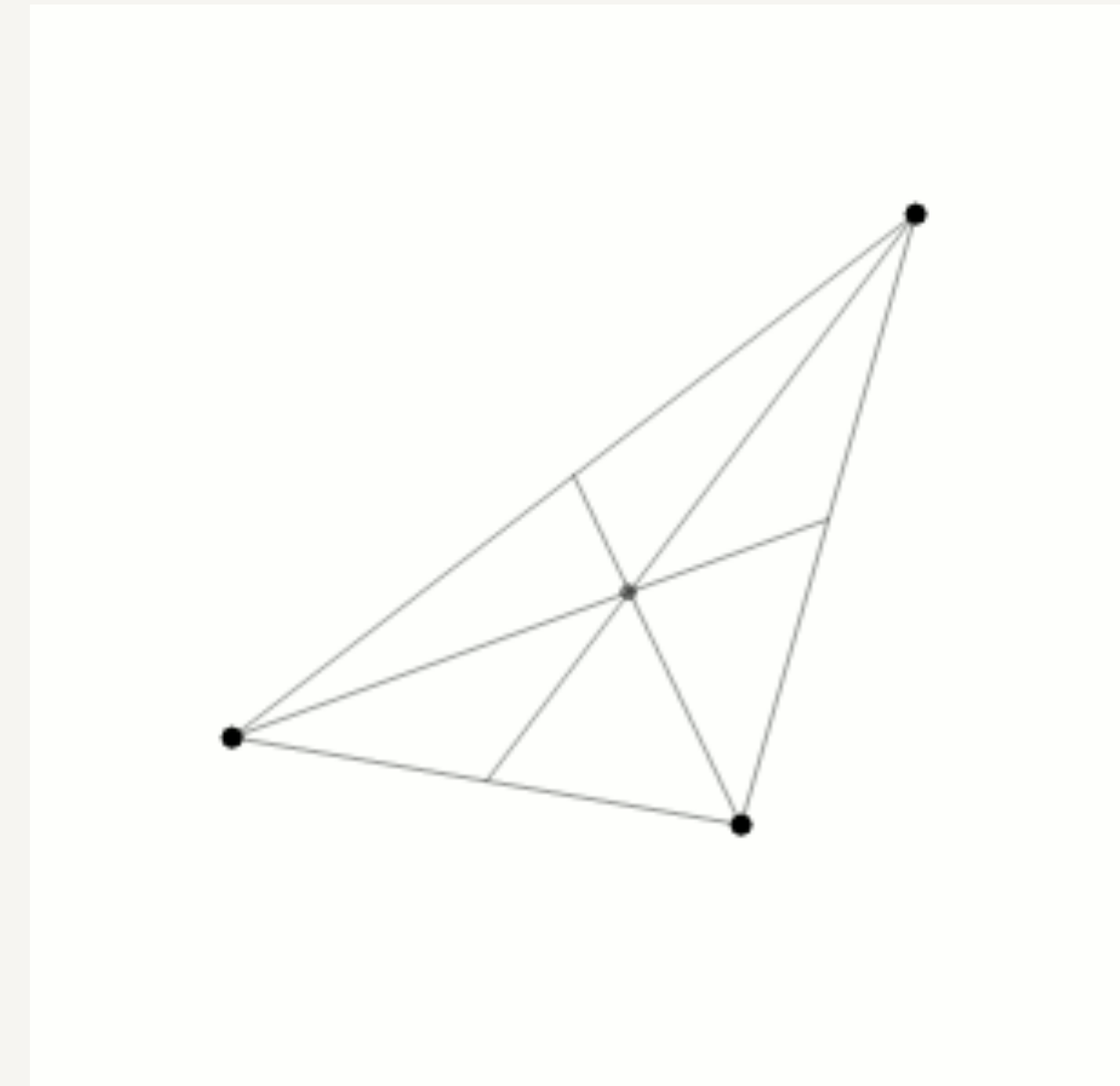
Physics of deep learning: Understanding grokking via the lens of physics



Ziming Liu, PhD student @ MIT, advised by Max Tegmark
April 27, 2023 @ Westlake University



ML Physics

A thick black double-headed arrow pointing left and right, positioned between the 'ML' and 'Physics' labels.

What is grokking in everyday life?

 grok (顿悟)

verb INFORMAL • US

gerund or present participle: **grokking**

understand (something) intuitively or by empathy.

"because of all the commercials, children grok things immediately"

- empathize or communicate sympathetically; establish a rapport.

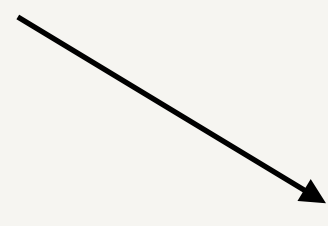
"nestling earth couple would like to find water brothers to **grok with** in peace"



What is grokking in science?

Apples fall to the ground.

Earth orbits around the Sun.



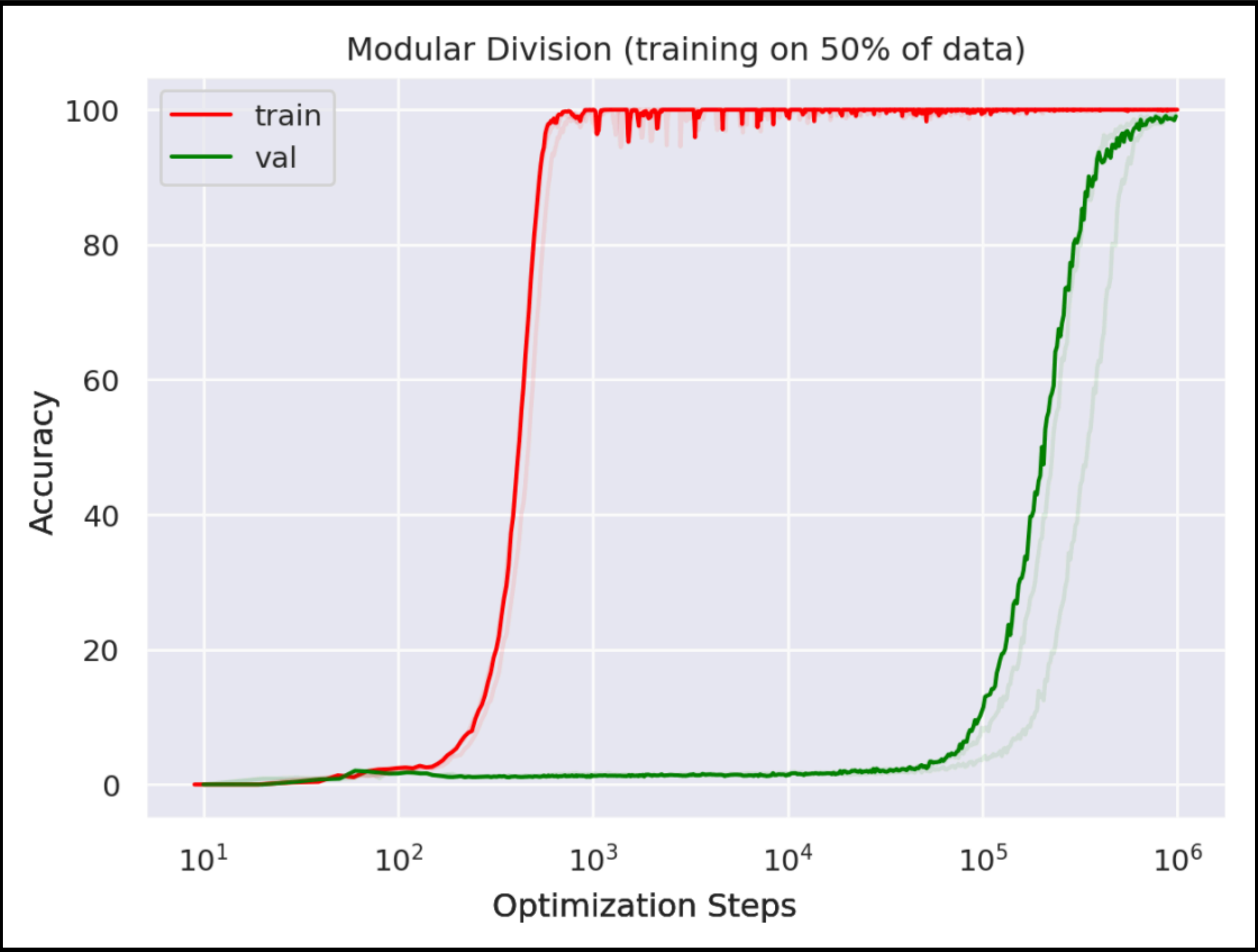
Universal gravitation

Generalisation !

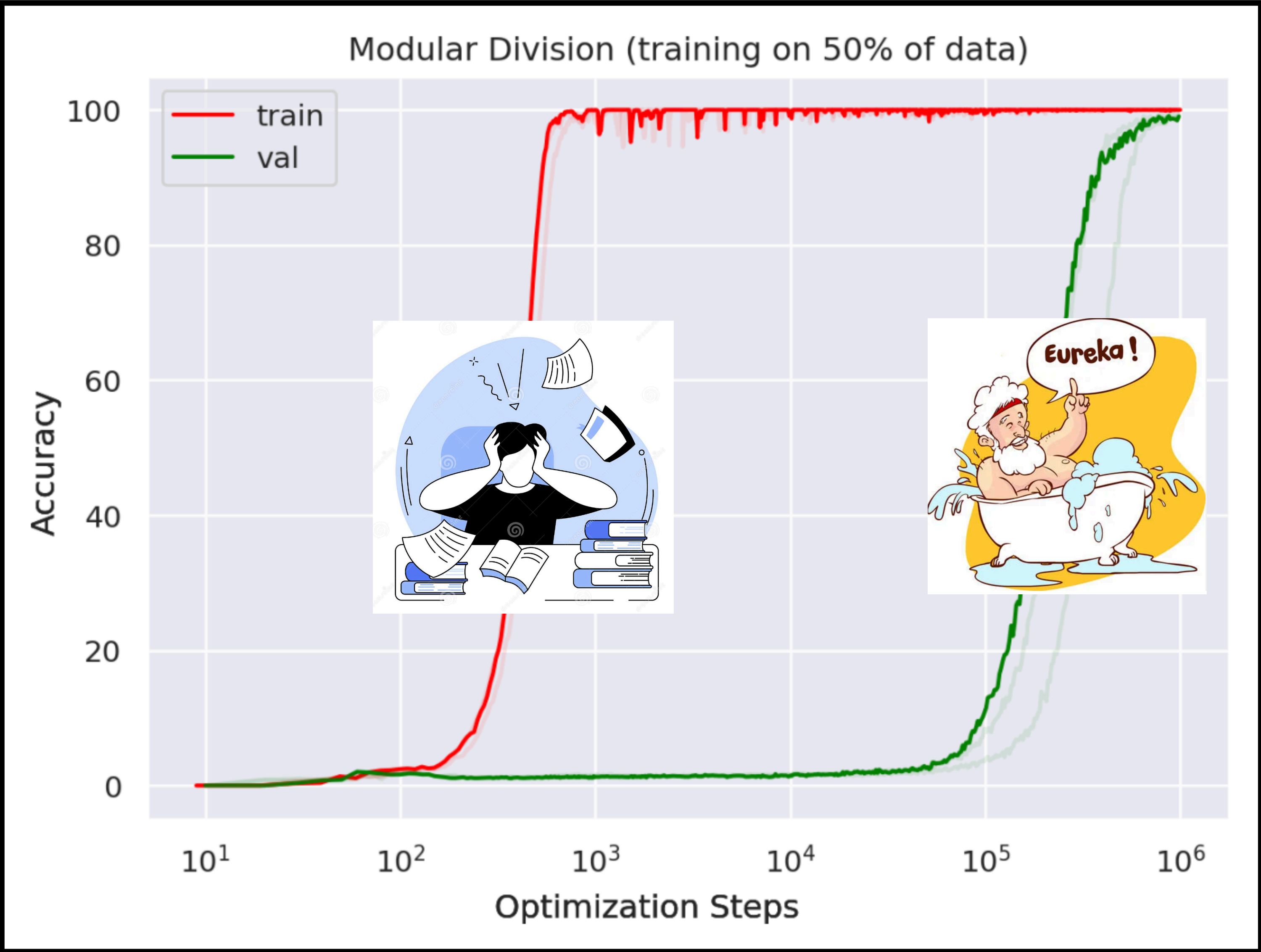
Sir Isaac Newton



What is grokking in ML?



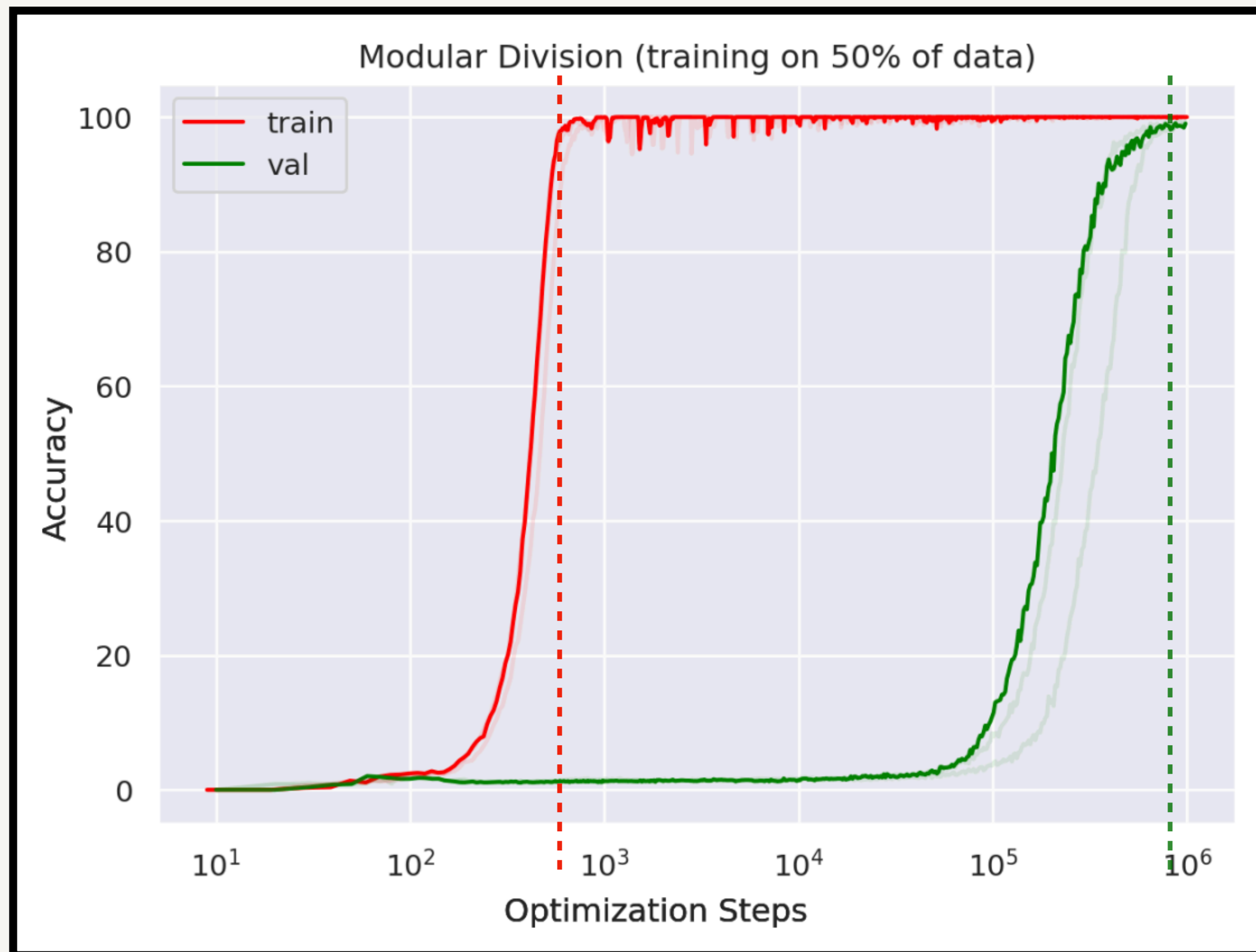
Grokking in everyday life and in ML



Grokking

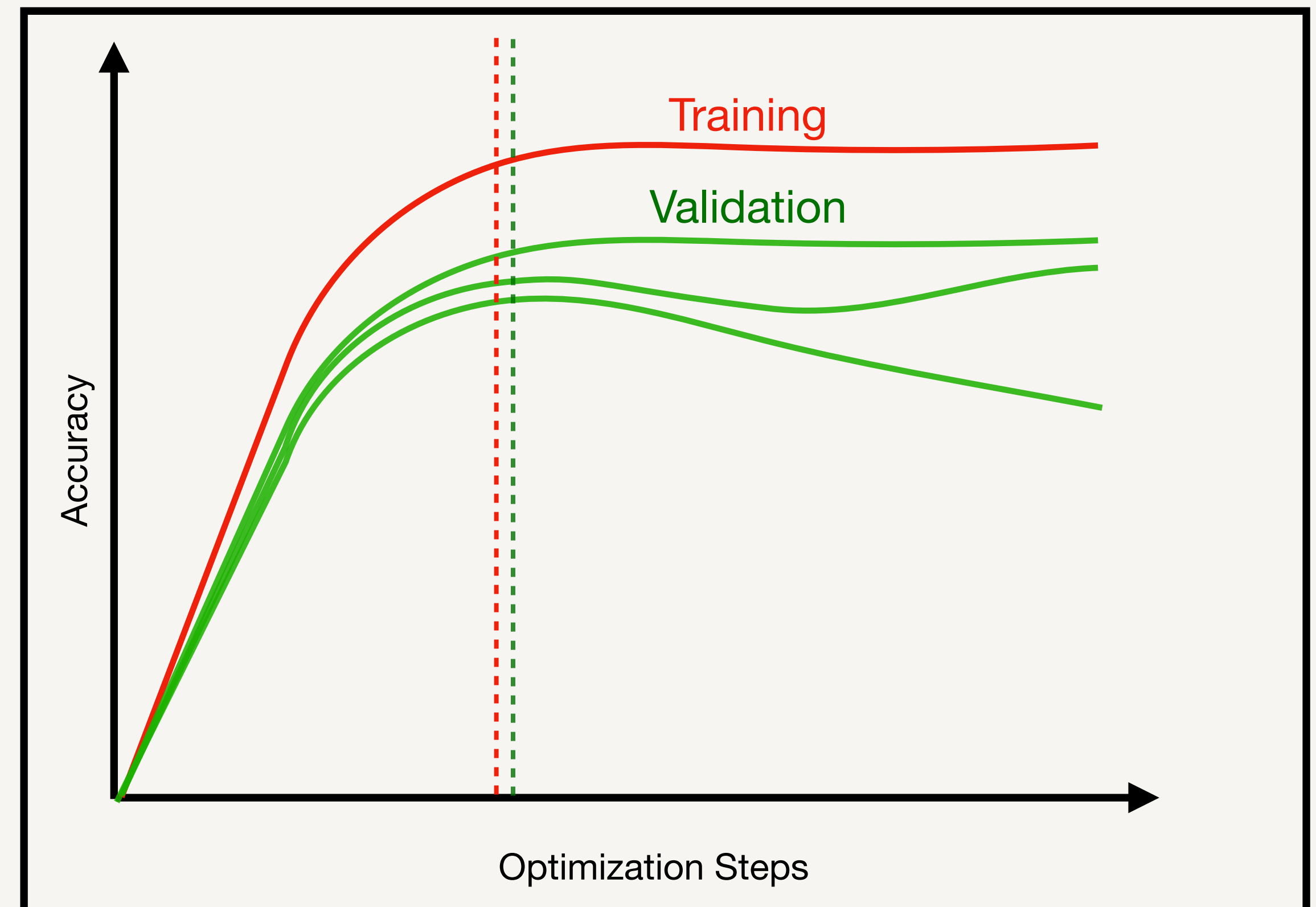
Puzzle 1: delayed generalization

Grokking



vs

Common training curves

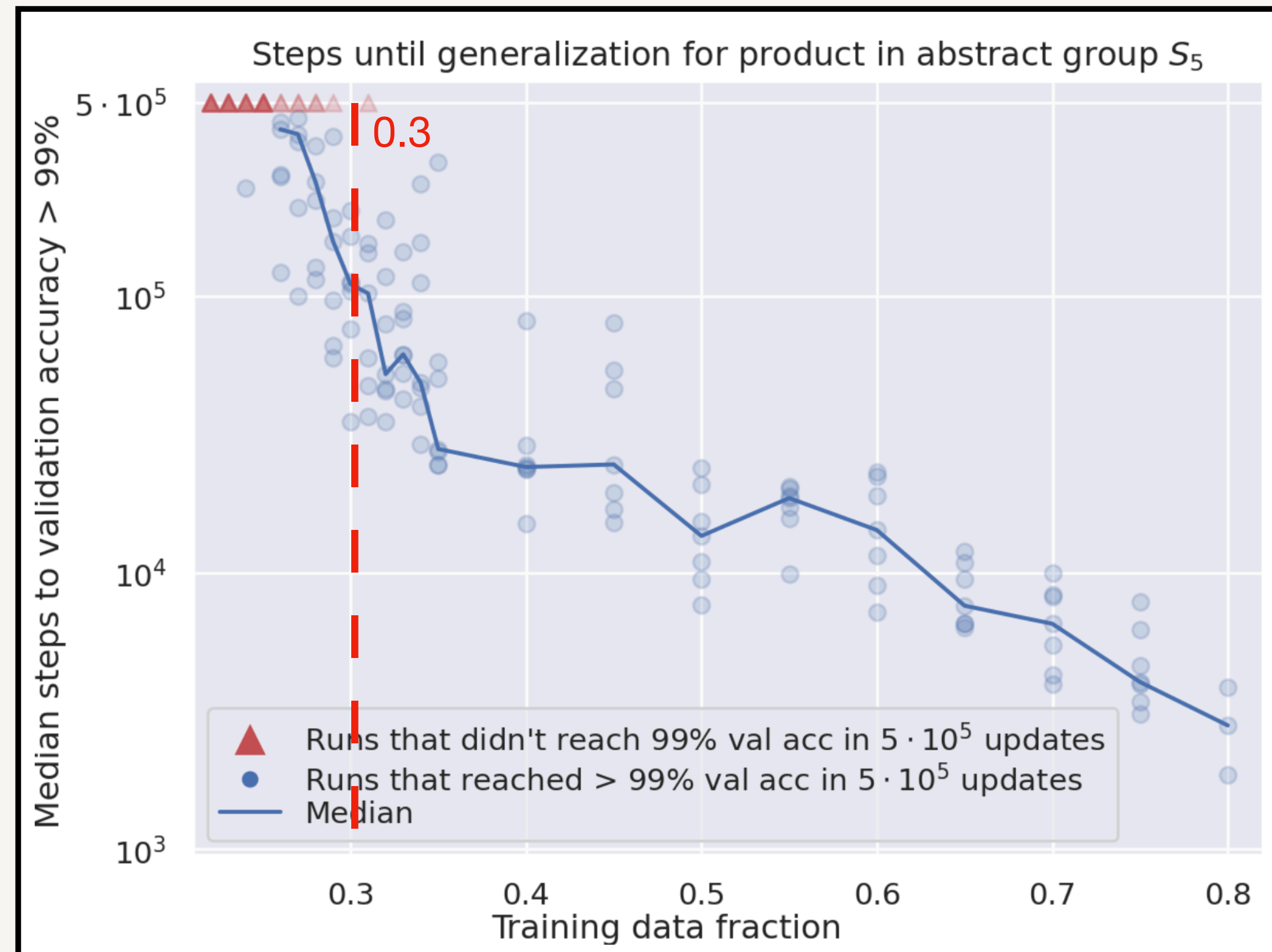


Validation accuracy is much **delayed** than training accuracy.

Training and validation accuracy go up **simultaneously**.

Grokking

Puzzle 2: dependence on training size



From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Grokking setup: Learning binary operation

$$\boxed{a} \circ \boxed{b} = \boxed{c}$$

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Grokking setup: Learning binary operation

Split the table into
train & **val** datasets

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

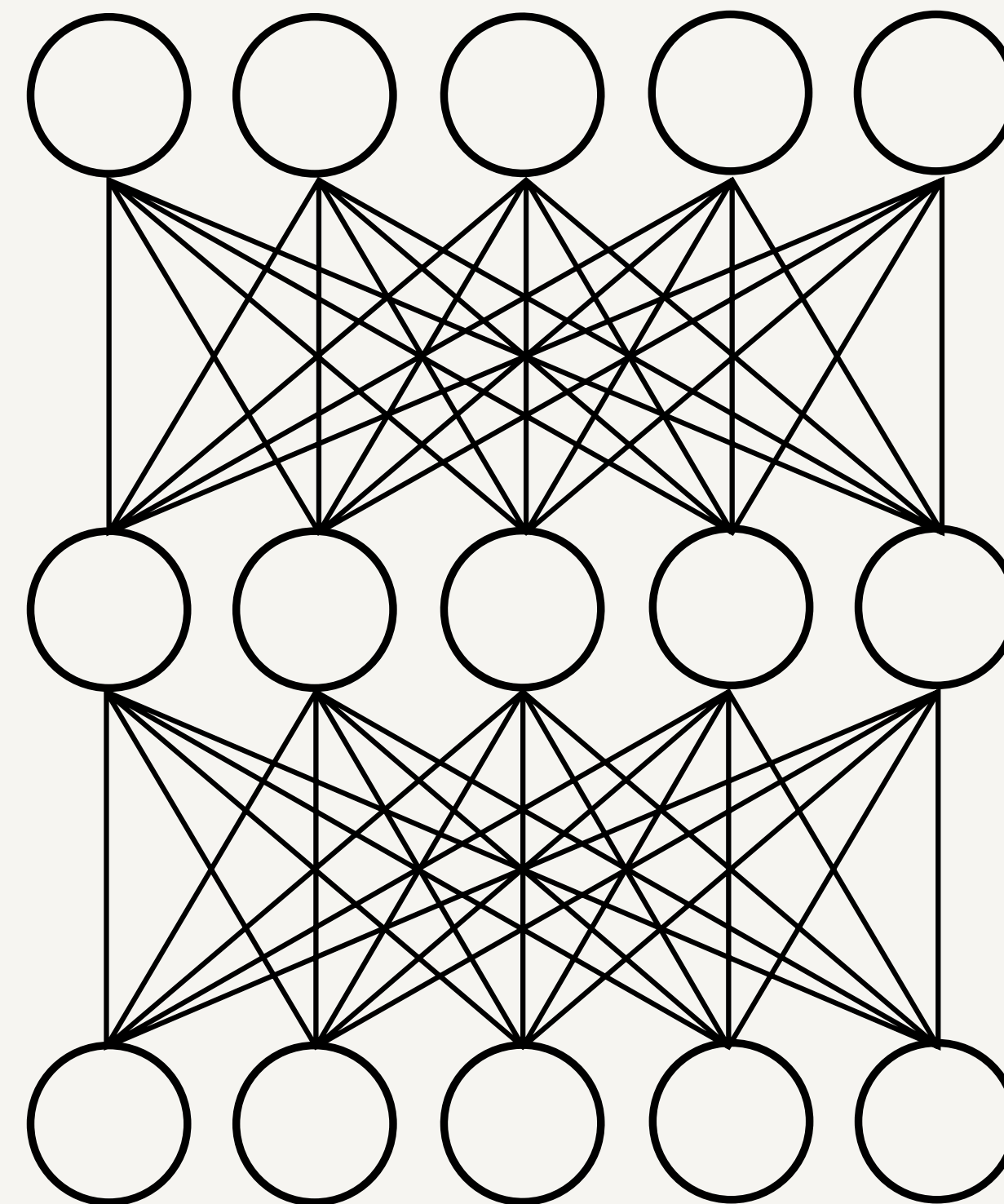
From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Grokking setup

Task: learn a binary operation

$$a + b \pmod{p} = c$$

Logits for a, b, c, ...

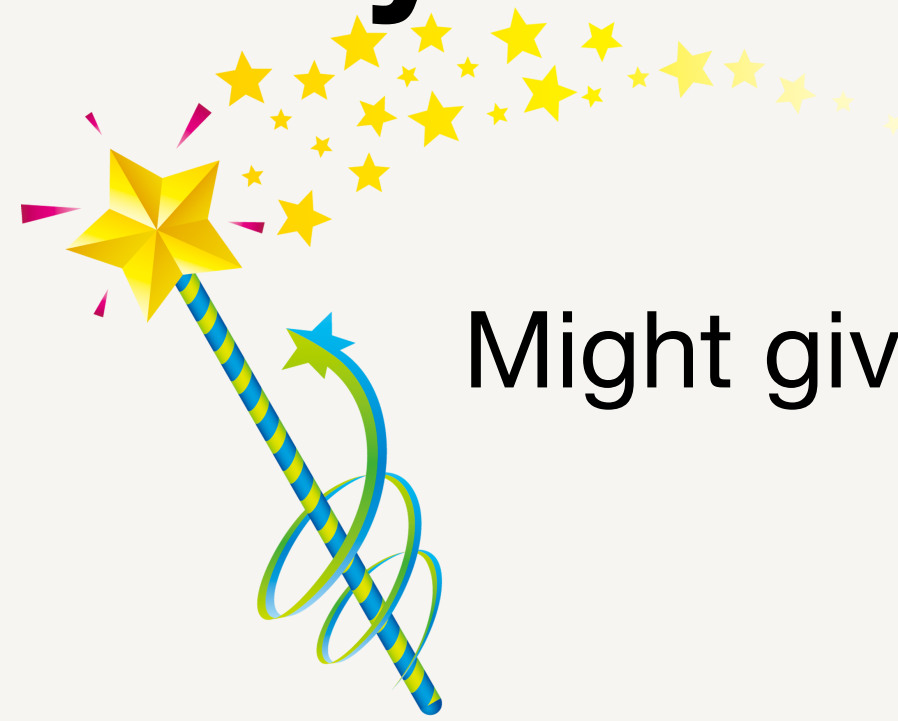


Decoder-only
Transformer
or MLP



← Trainable Embeddings

Why is Grokking interesting?



Might give practitioners hope that neural networks will eventually magically generalize



Alethea Power 1 month ago

"Did someone forget to turn off the computer?" 😅 That's exactly how it happened. One of my coworkers was training a network and he forgot to turn it off when he went on vacation. When he came back, it had learned. So we dug in and tried to figure out how and why it learned so long after we ...



305



REPLY

▼ [View 13 replies](#)

A comment from an author of the openai grokking paper, on YouTube.

Questions raised by *grokking*



No Magic!!!

1. How do networks generalize at all on algorithmic datasets?

- *Representation*

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a



2. Why does grokking (generalization) time depend strongly on the training set fraction?

- *Training size controls the speed of representation learning*

3. Under what conditions is generalization delayed?

- *Improper hyper-parameters that prohibit representation*



Towards Understanding Grokking: An Effective Theory of Representation Learning

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams
Department of Physics, Institute for AI and Fundamental Interactions, MIT
`{zmliu,kitouni,nnolte,ericjm,tegmark,mwill}@mit.edu`

Accepted by NeurIPS 2022 (Oral)



Ziming Liu



Ouail Kitouni



Niklas Nolte



Eric J. Michaud



Max Tegmark



Mike Williams

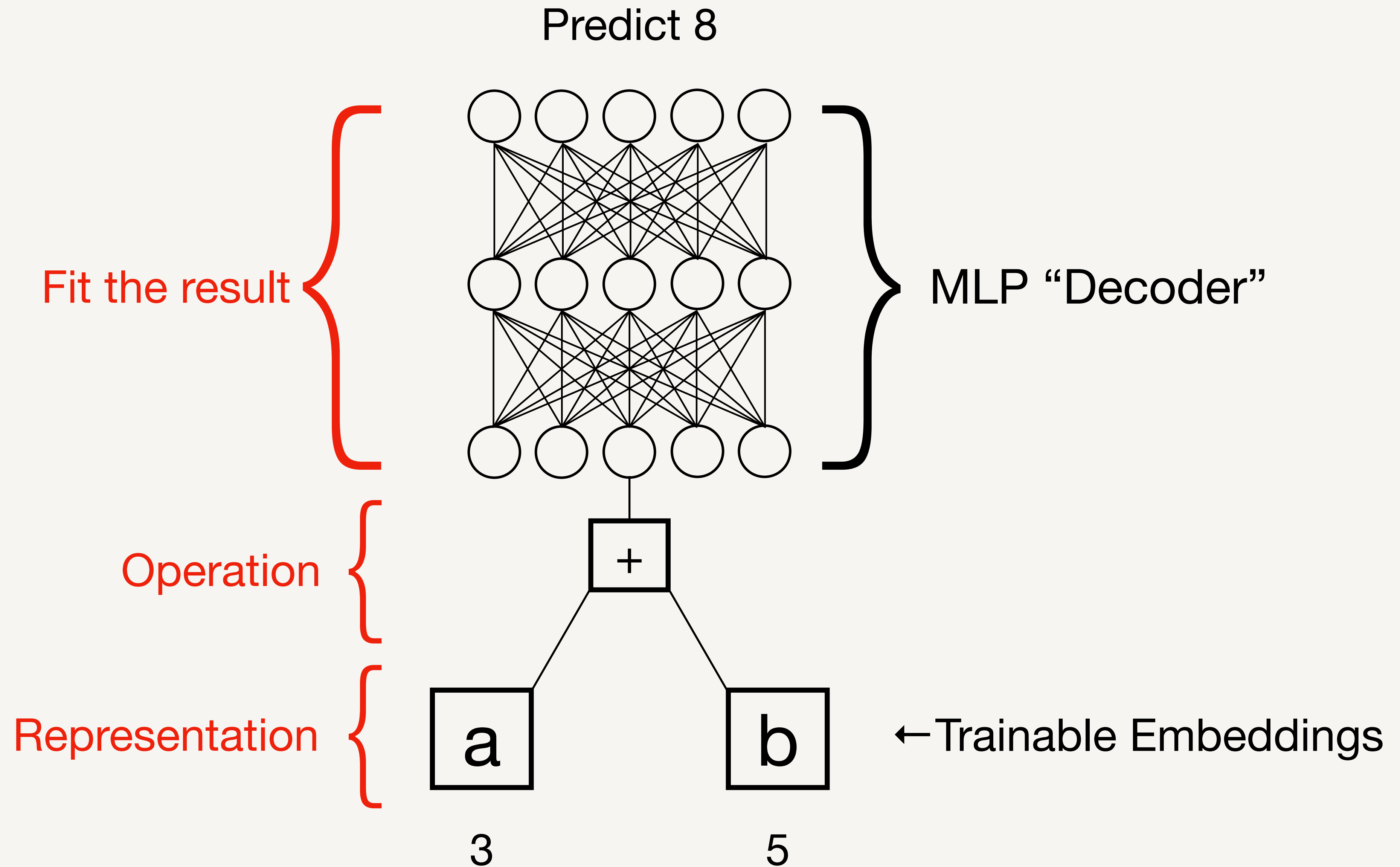


Q1: How do networks generalize at all on algorithmic datasets?

A1: *Representation.*

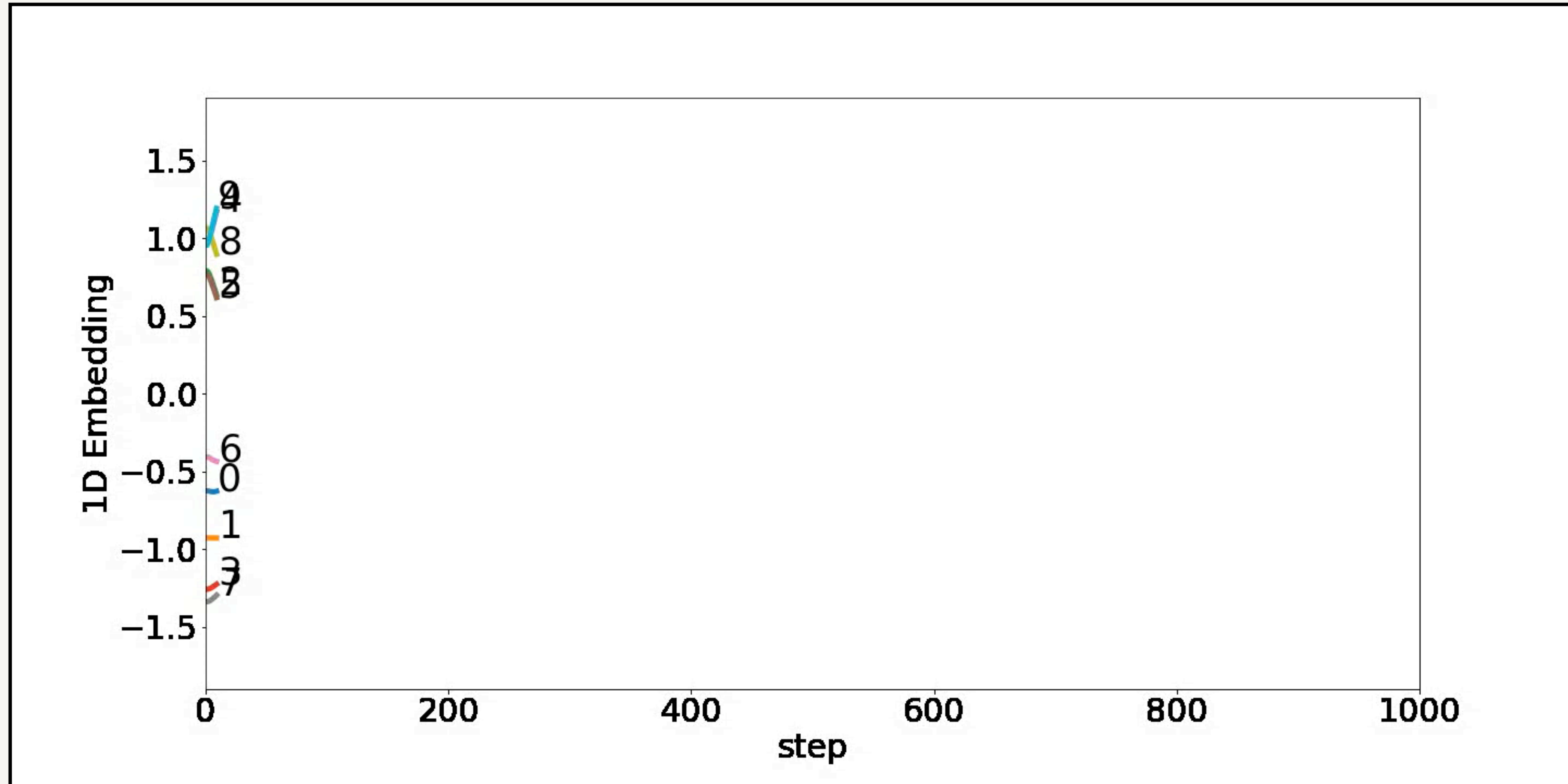
Toy Model

Addition dataset



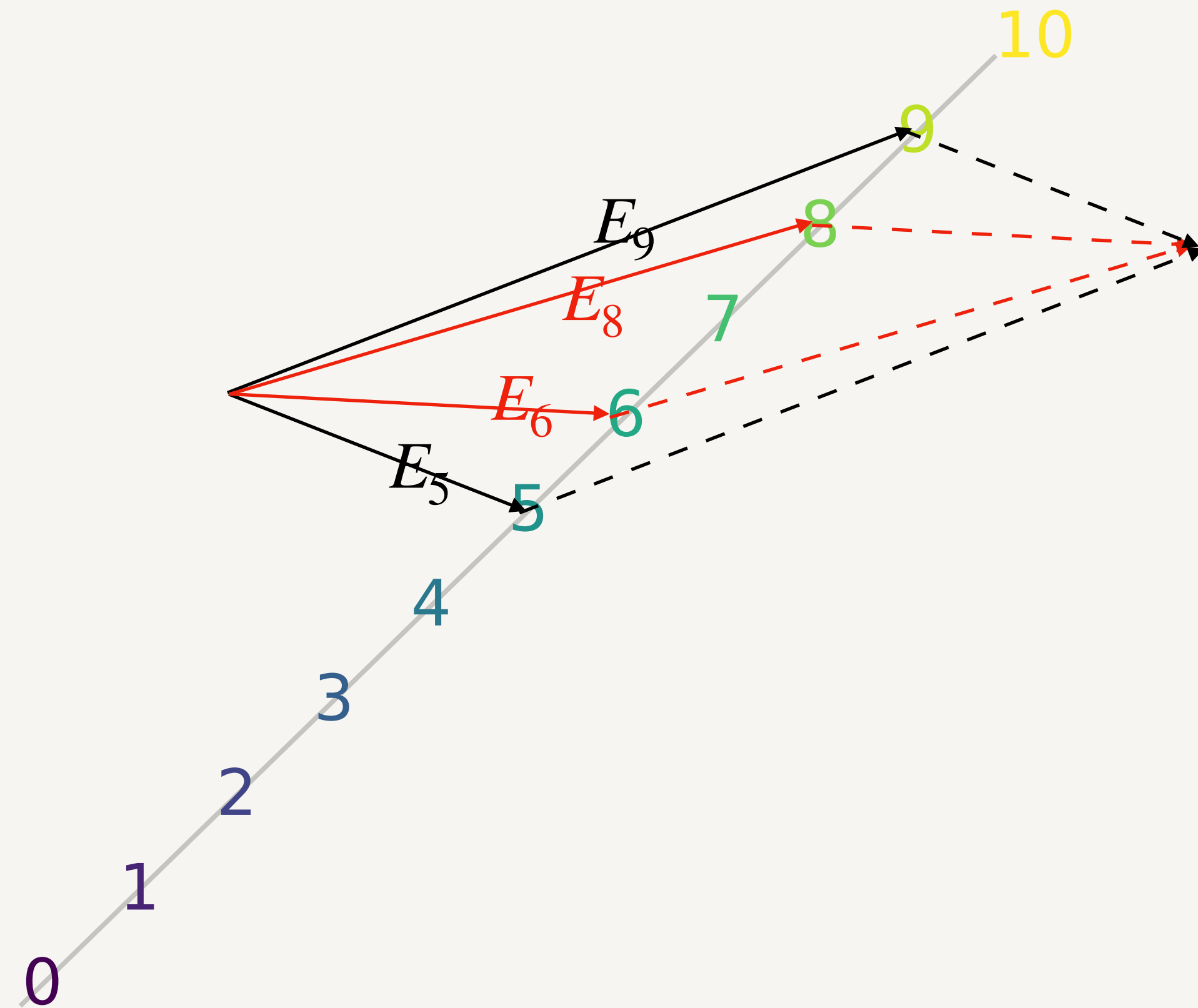
Peek in a generalisation case

Addition & toy model, 100% test accuracy



Representation is key to generalization!

Addition & toy model



If $5 + 9 = 14$

is in the train set

then the toy model will

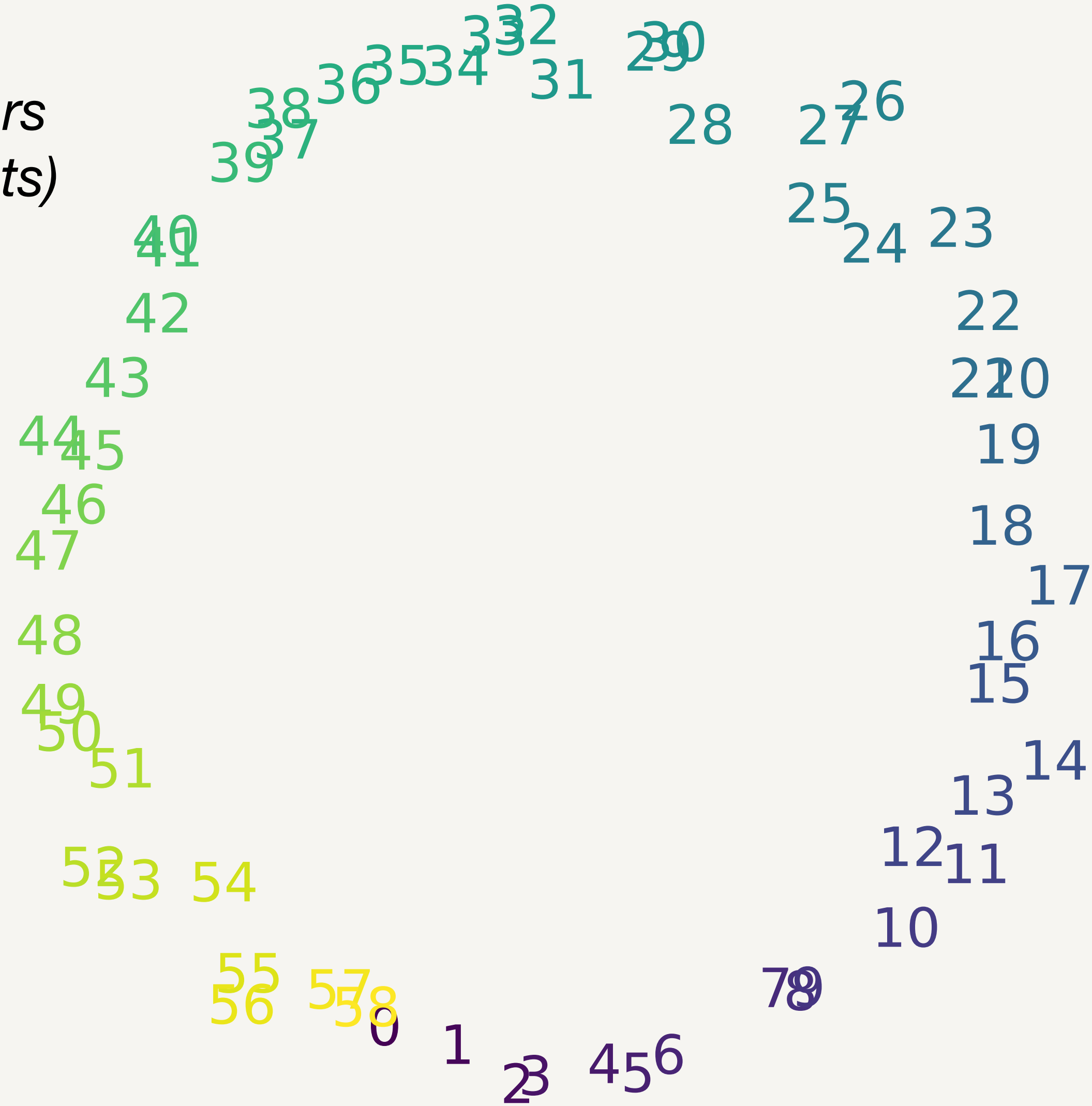
generalize to $6 + 8$

Because $E_5 + E_9 = E_6 + E_8$

Representation is key to generalization!

Modular addition & non-toy

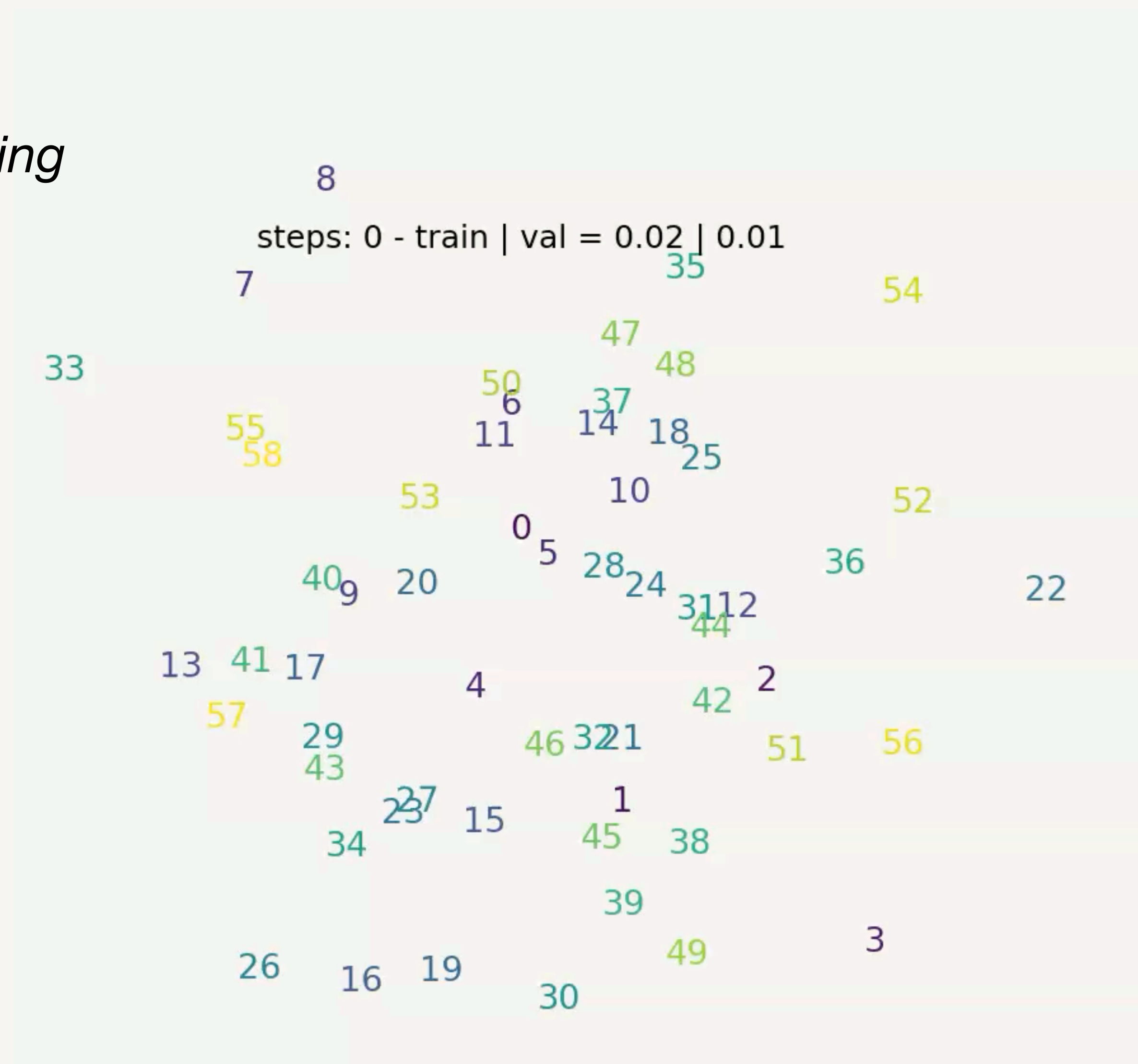
*Final embedding vectors
(first 2 PCA components)*



Representation is key to generalization!

Modular addition & non-toy

*embeddings over training
using PCA*

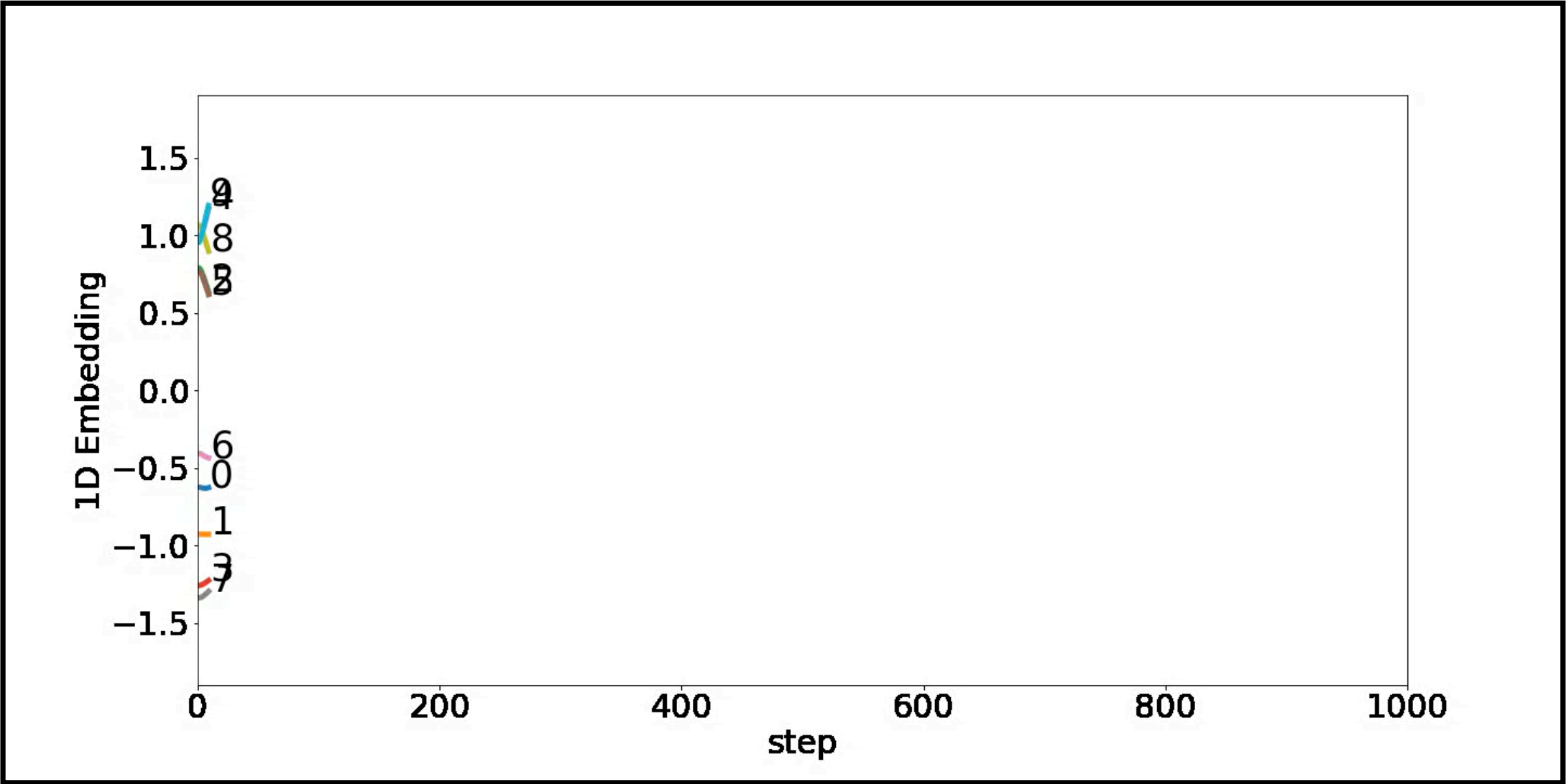


Q2: Why does grokking (generalization) time depend on training size?

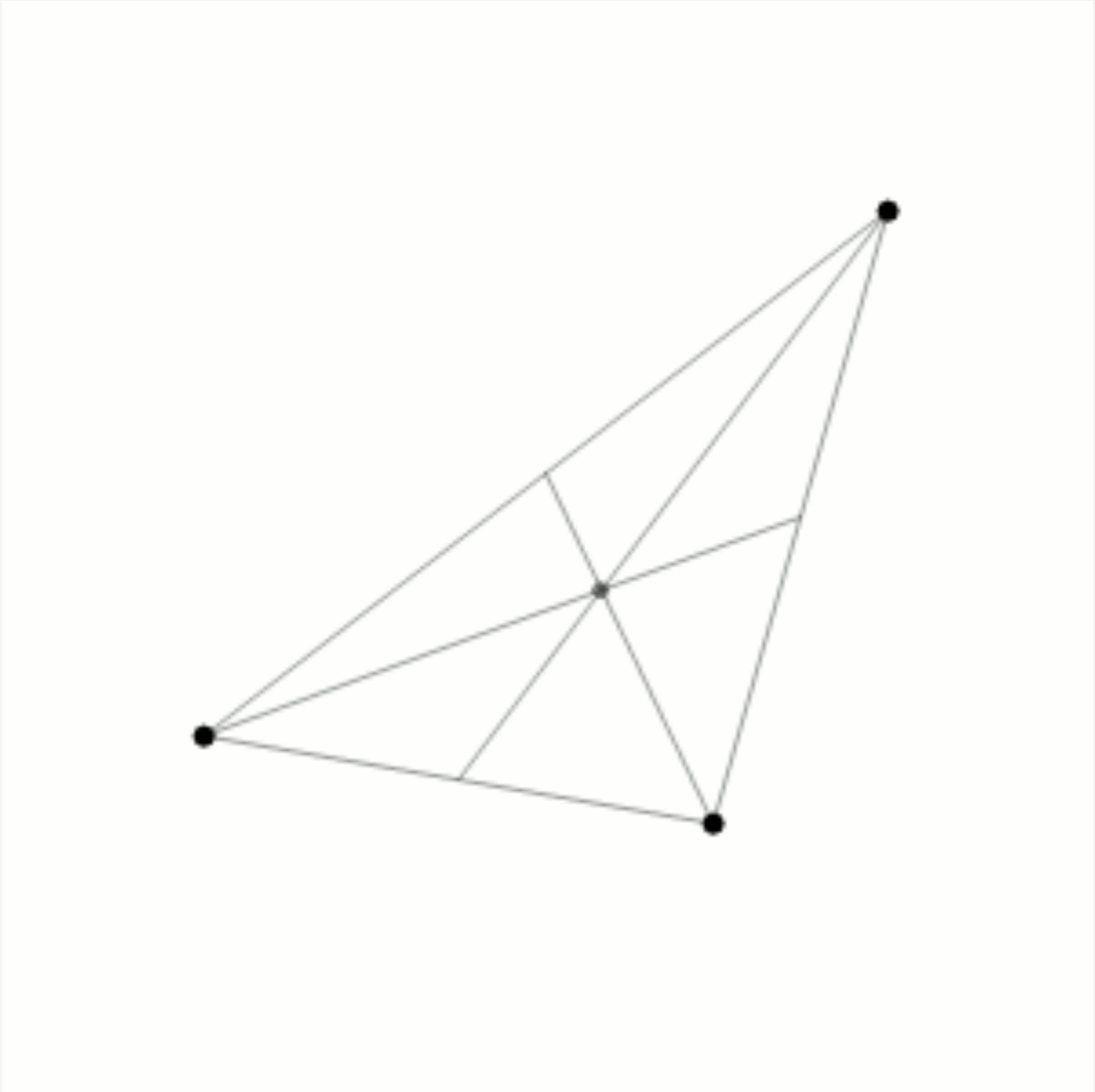
A2: Training size controls the speed of *representation* learning.

The dynamics of representation

Addition & toy model



ML Physics



Effective theory

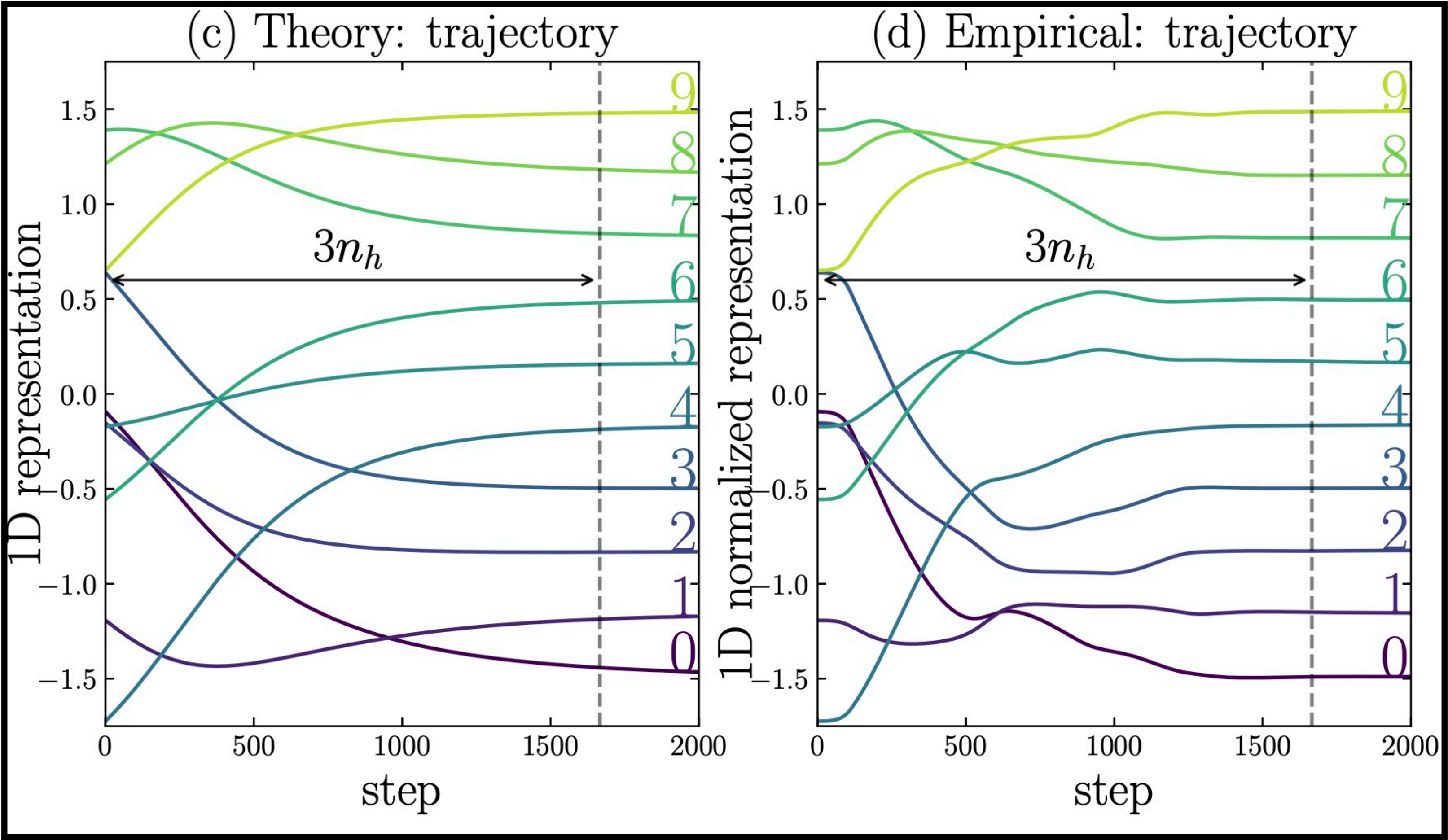
$$P_0(D) = \{(i, j, m, n) \mid (i, j) \in D, (m, n) \in D, i + j = m + n\}$$

$$\ell_{\text{eff}} = \frac{\ell_0}{Z_0}, \quad \ell_0 \equiv \sum_{(i,j,m,n) \in P_0(D)} |\mathbf{E}_i + \mathbf{E}_j - \mathbf{E}_m - \mathbf{E}_n|^2, \quad Z_0 \equiv \sum_k |\mathbf{E}_k|^2,$$

$$\frac{dE_i}{dt} = -\eta \frac{d\ell_{\text{eff}}}{dE_i}$$

Compare theory and experiment

Addition & toy model

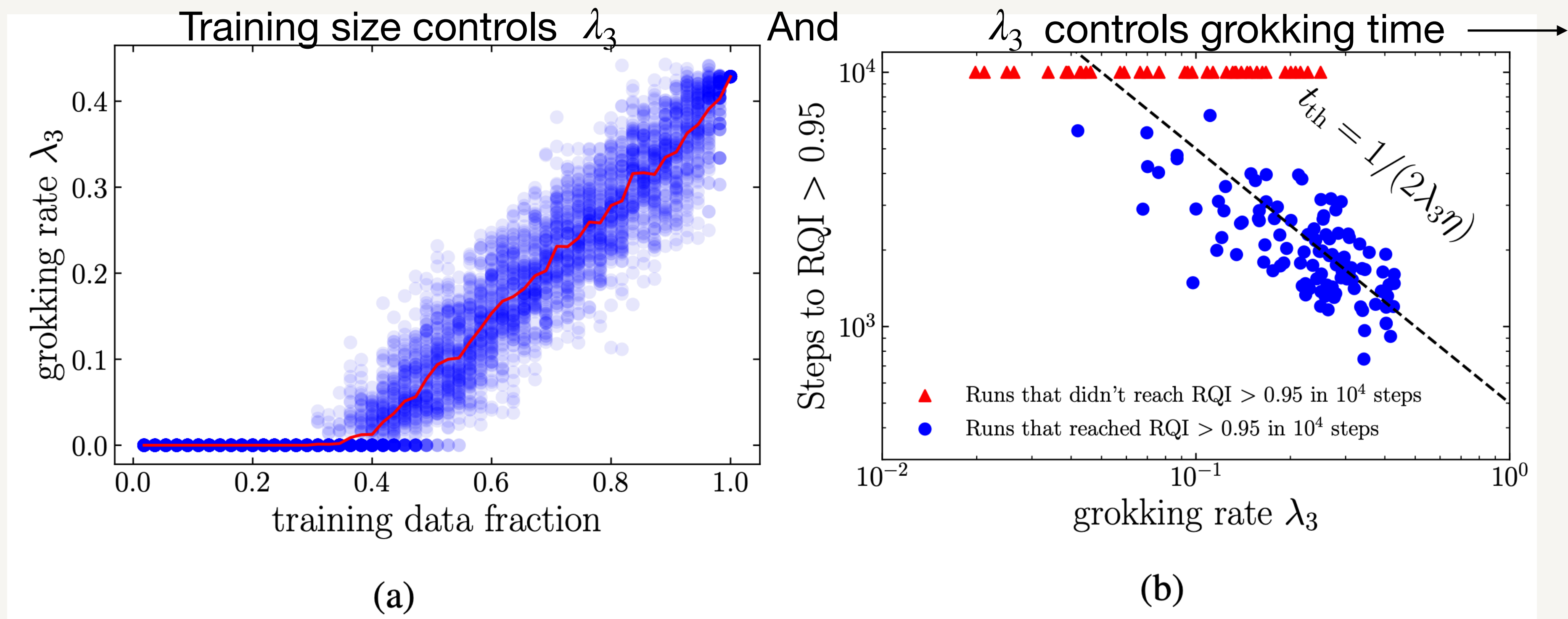


Grokking time dependence on train size

$$\ell_{\text{eff}} = \frac{\ell_0}{Z_0}, \quad \ell_0 \equiv \sum_{(i,j,m,n) \in P_0(D)} |\mathbf{E}_i + \mathbf{E}_j - \mathbf{E}_m - \mathbf{E}_n|^2, \quad Z_0 \equiv \sum_k |\mathbf{E}_k|^2,$$

Define Hessian $H_{ij} = \frac{\partial^2 \ell_0}{\partial E_i \partial E_j}$ with eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \dots$ where $\lambda_1 = \lambda_2 = 0$

Proposition: Grokking time is proportional to λ_3^{-1}

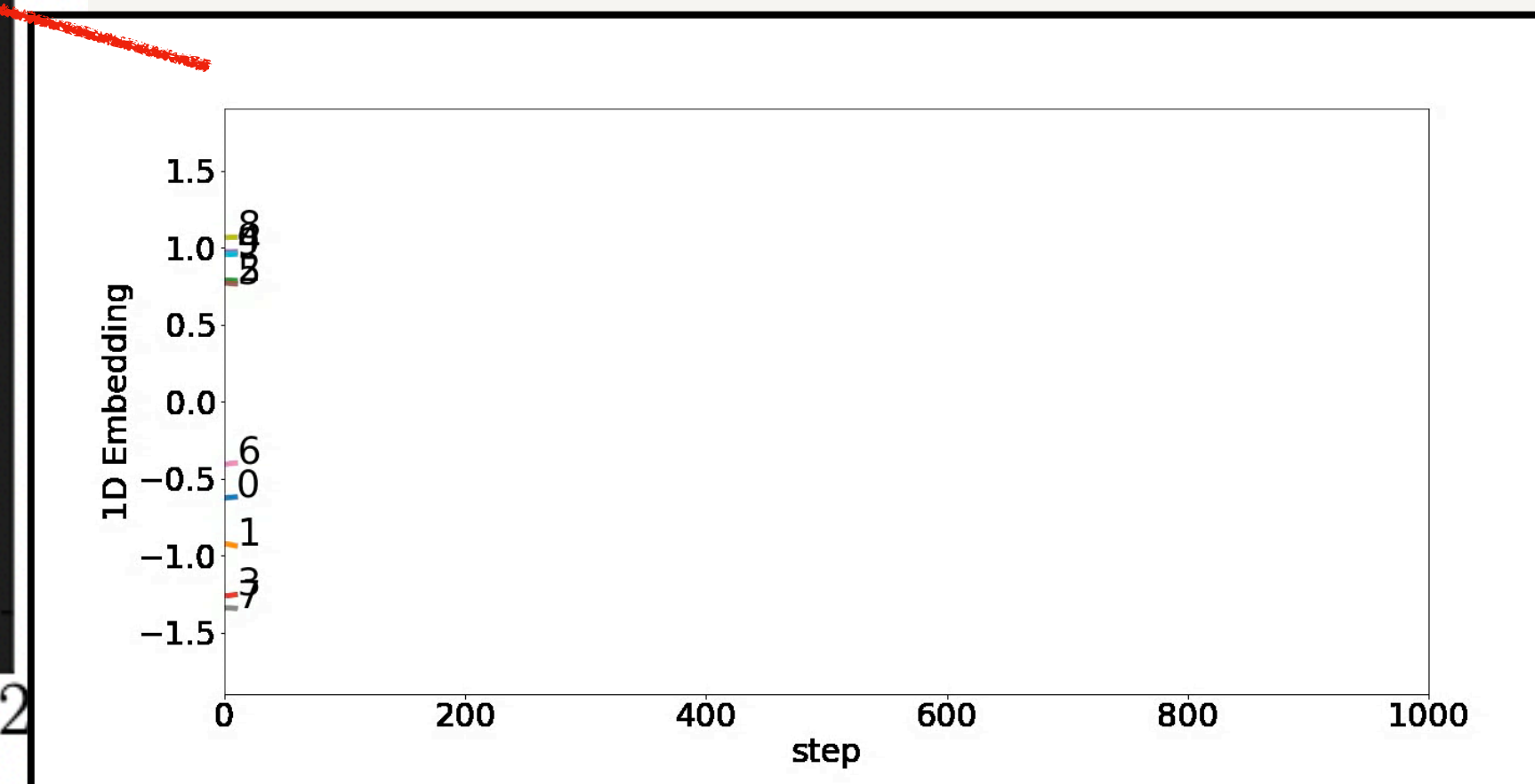
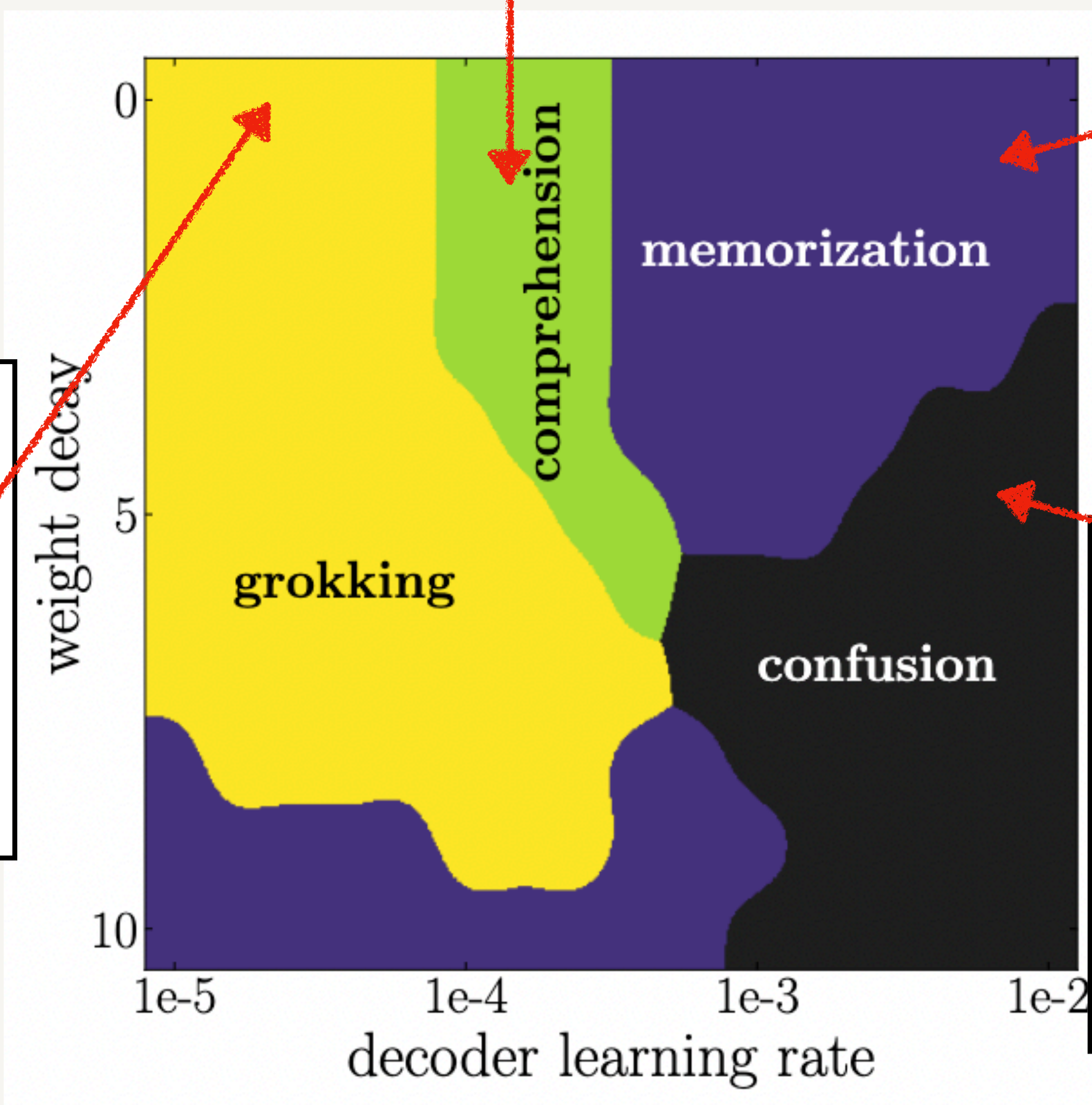
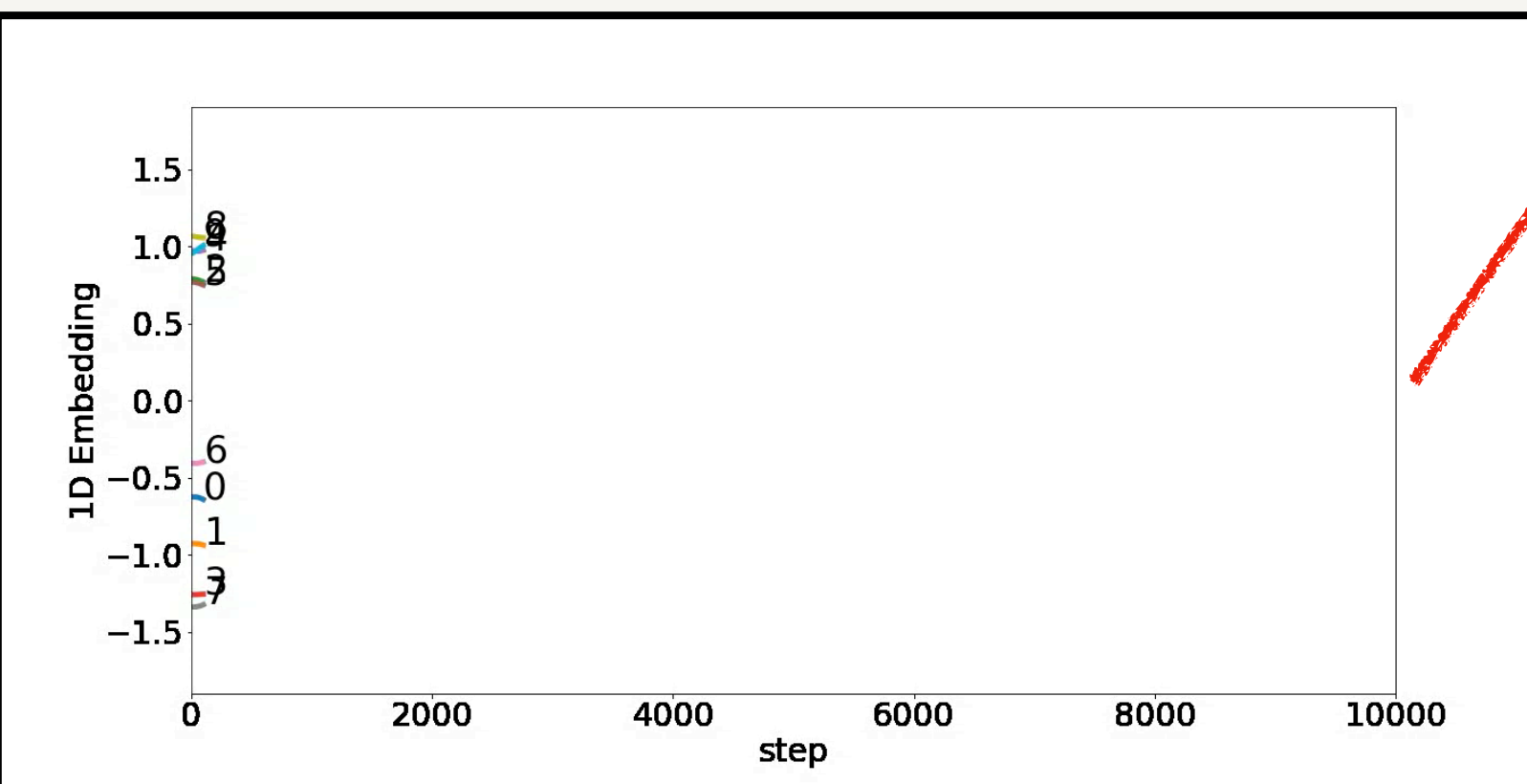
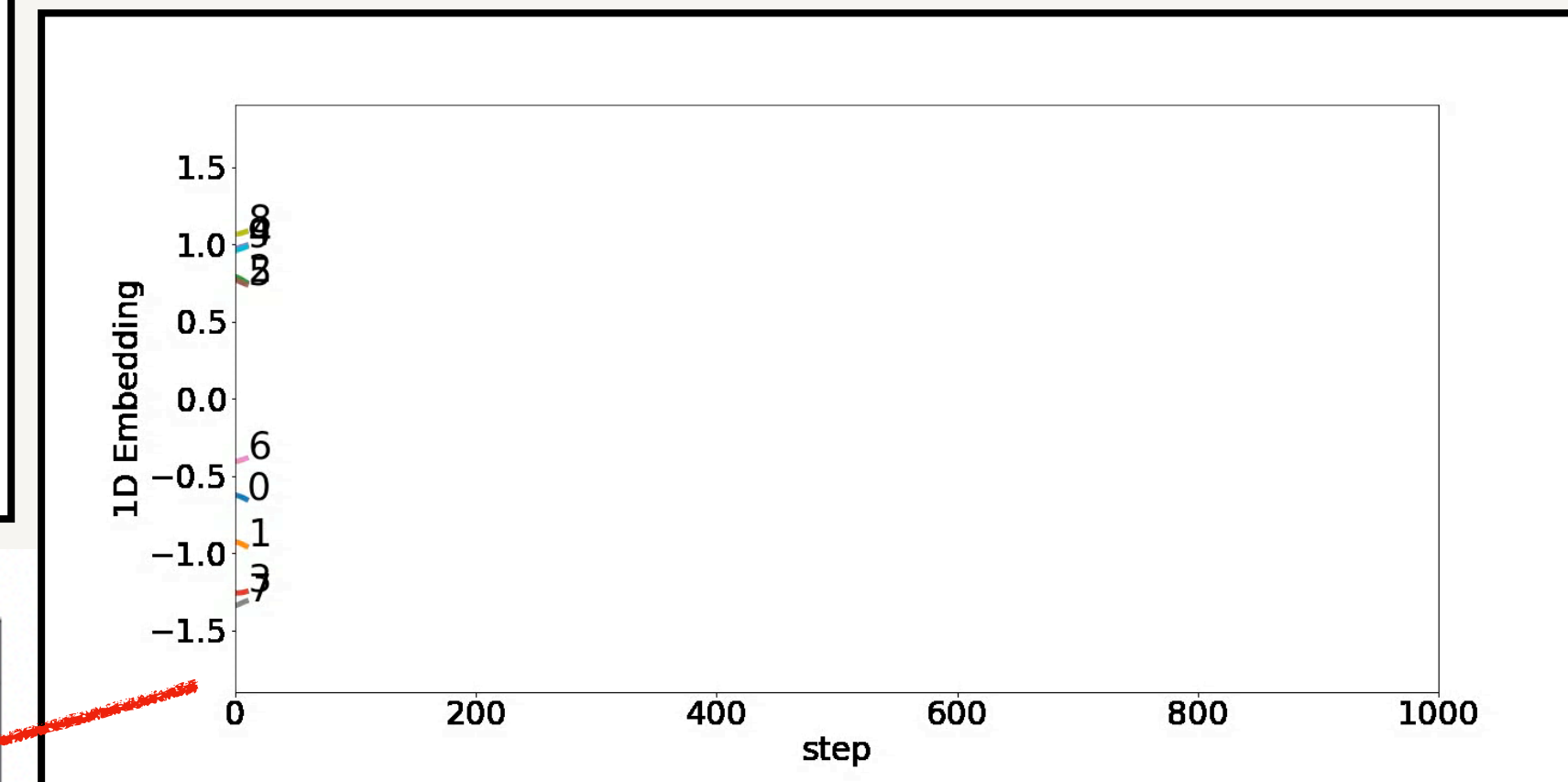
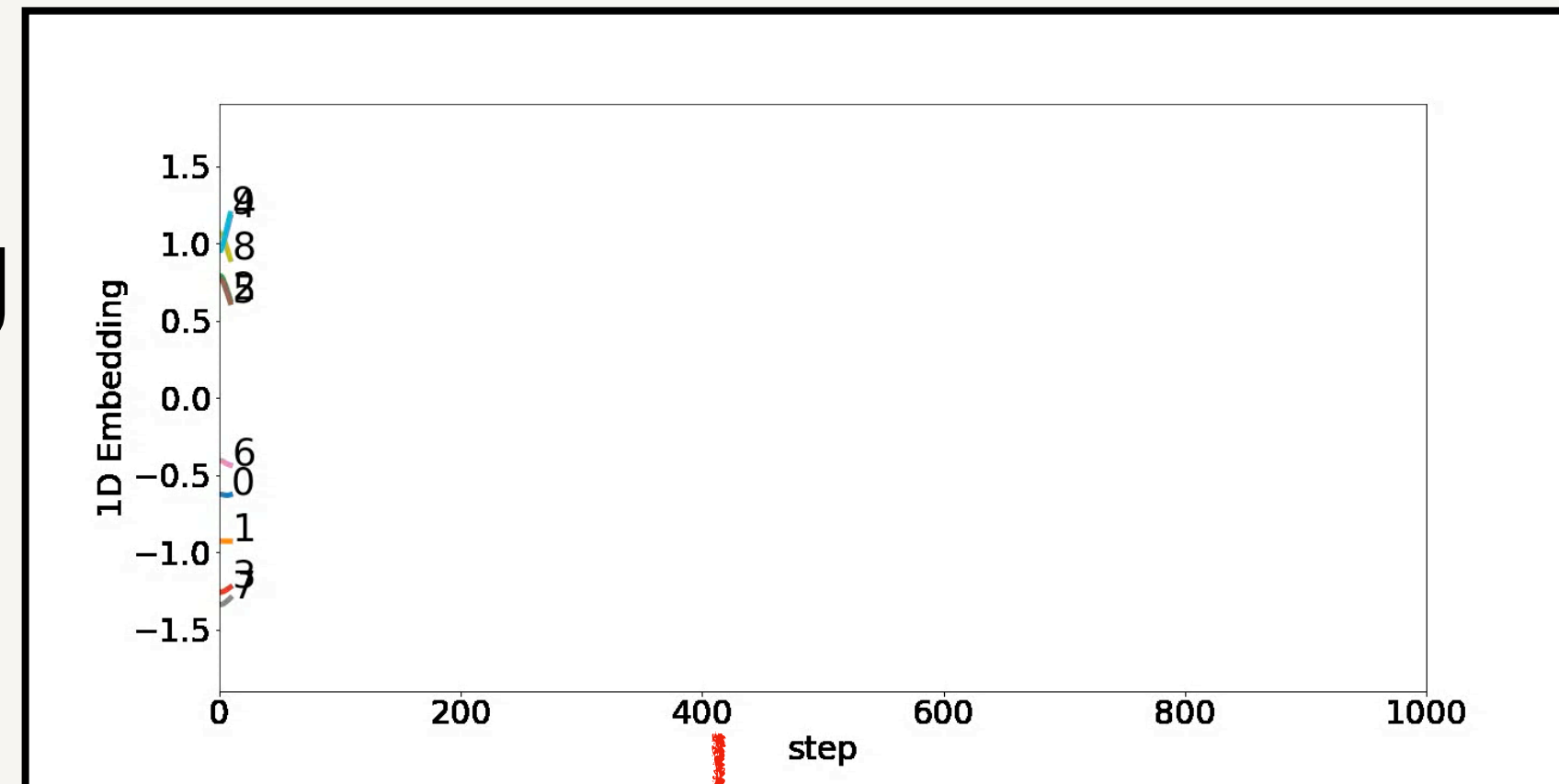


Q3: Under which conditions is generalization delayed?

A3: Improper hyper-parameters that prohibit *representation*.

Grokking from slow representation learning

Addition & toy model



Summary

1. Observed that generalization is associated with the model learning *structured representations*.
2. Developed an *effective theory* for learning dynamics of representations (embeddings) in a toy setting. Our theory exhibits a phase transition in train data fraction.
3. Made *phase diagrams* describing how learning dynamics depend on hyperparameters, allowing us to control grokking.

Still, we want to understand:

Q1: The origin of grokking from dynamics on loss landscape: Why is generalization much delayed after overfitting?

LU mechanism

Q2: The prevalence of grokking: Can grokking occur on datasets other than algorithmic datasets?

Yes



Omnigrok: Grokking Beyond Algorithmic Data

Accepted to ICLR 2023 (Spotlight)



Ziming Liu

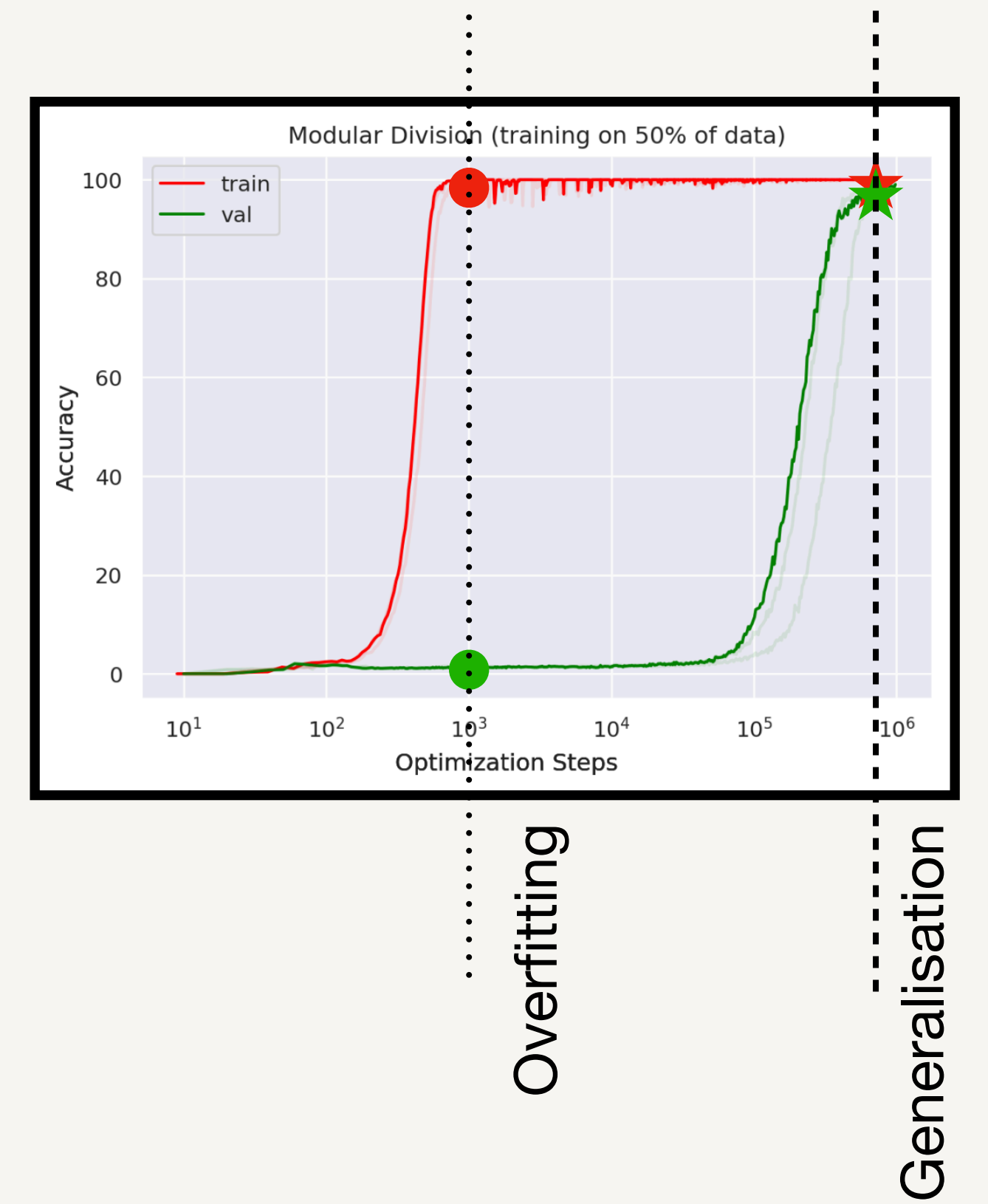
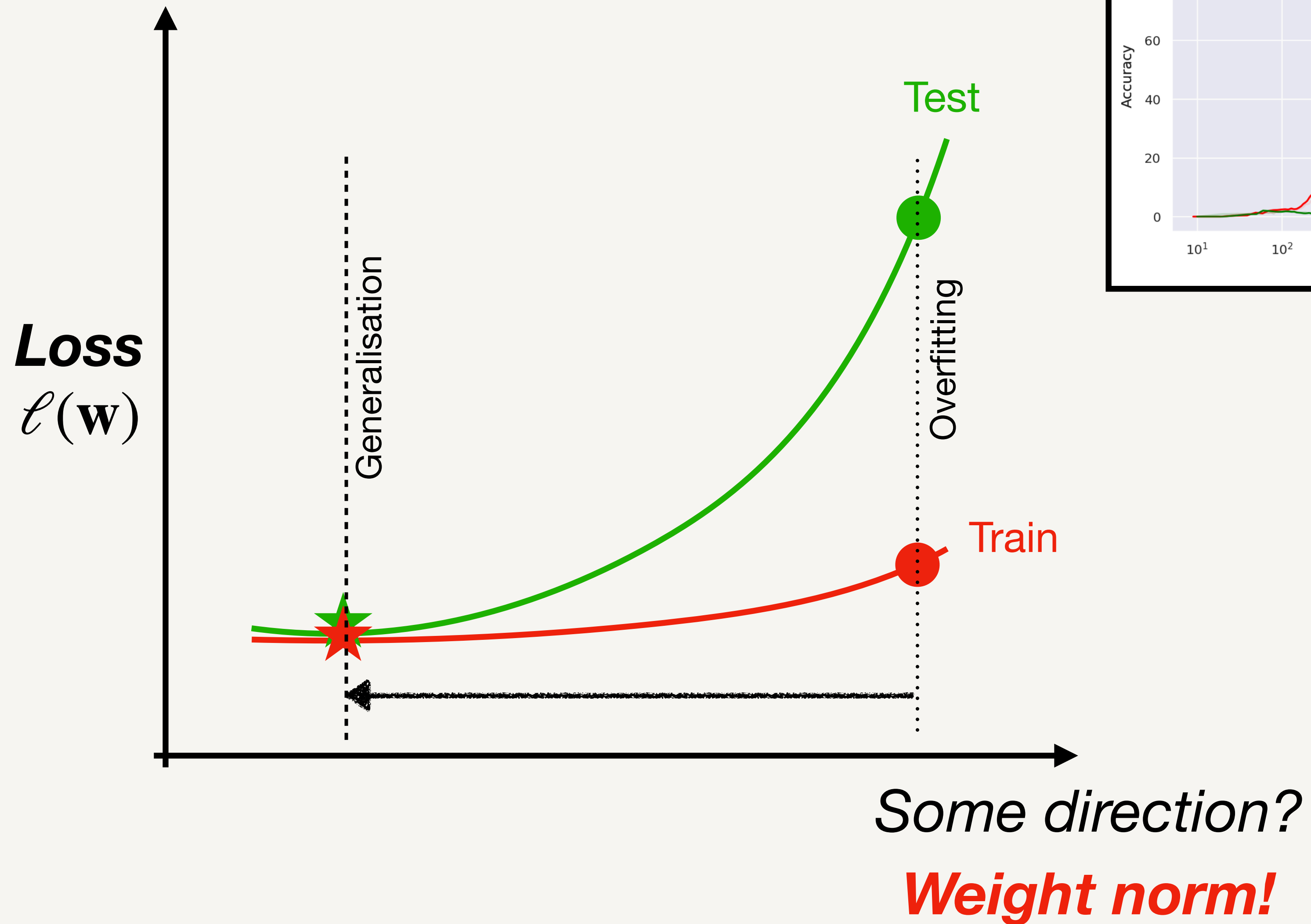


Eric J. Michaud

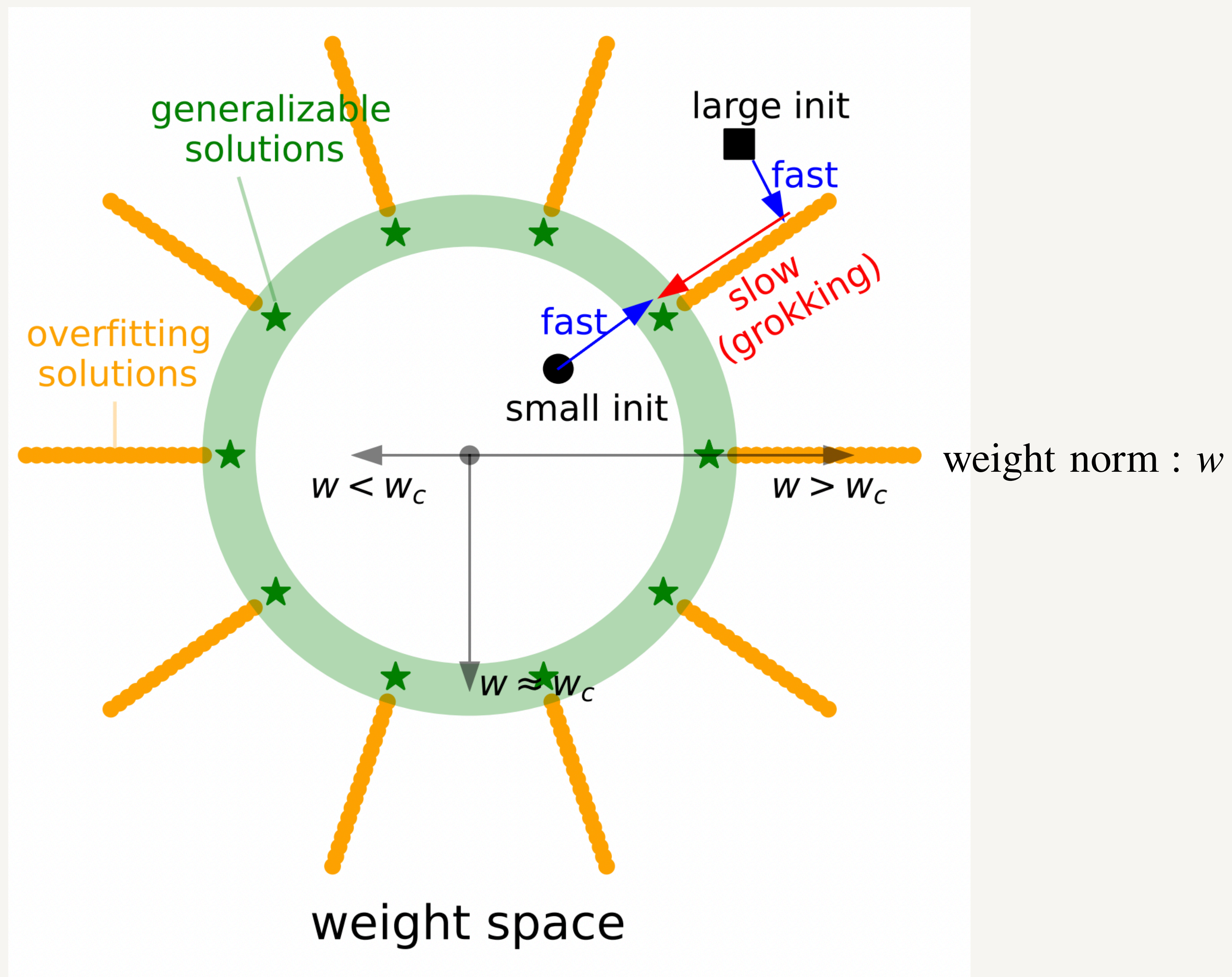


Max Tegmark

Grokking due to train/test mismatch



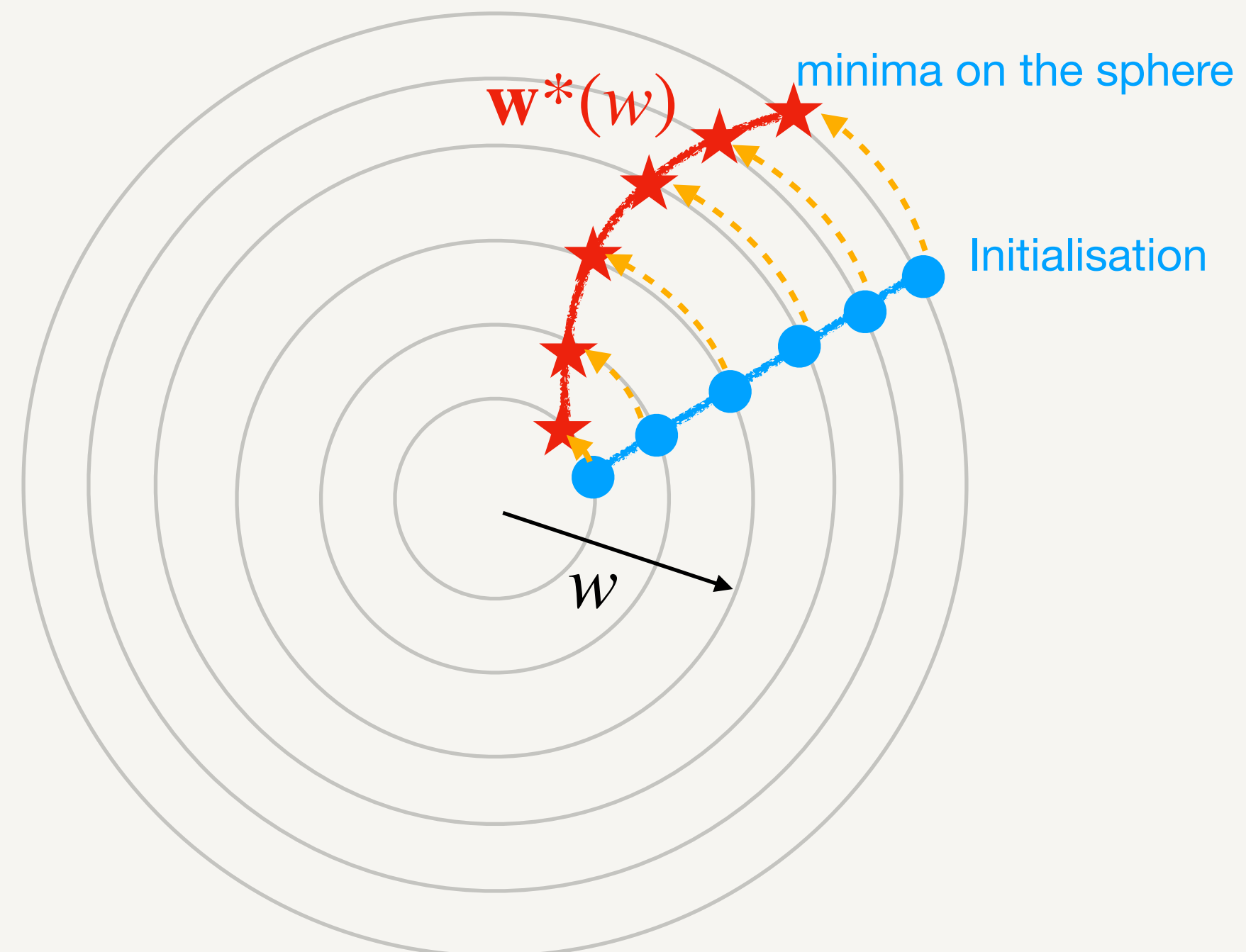
Loss Landscape



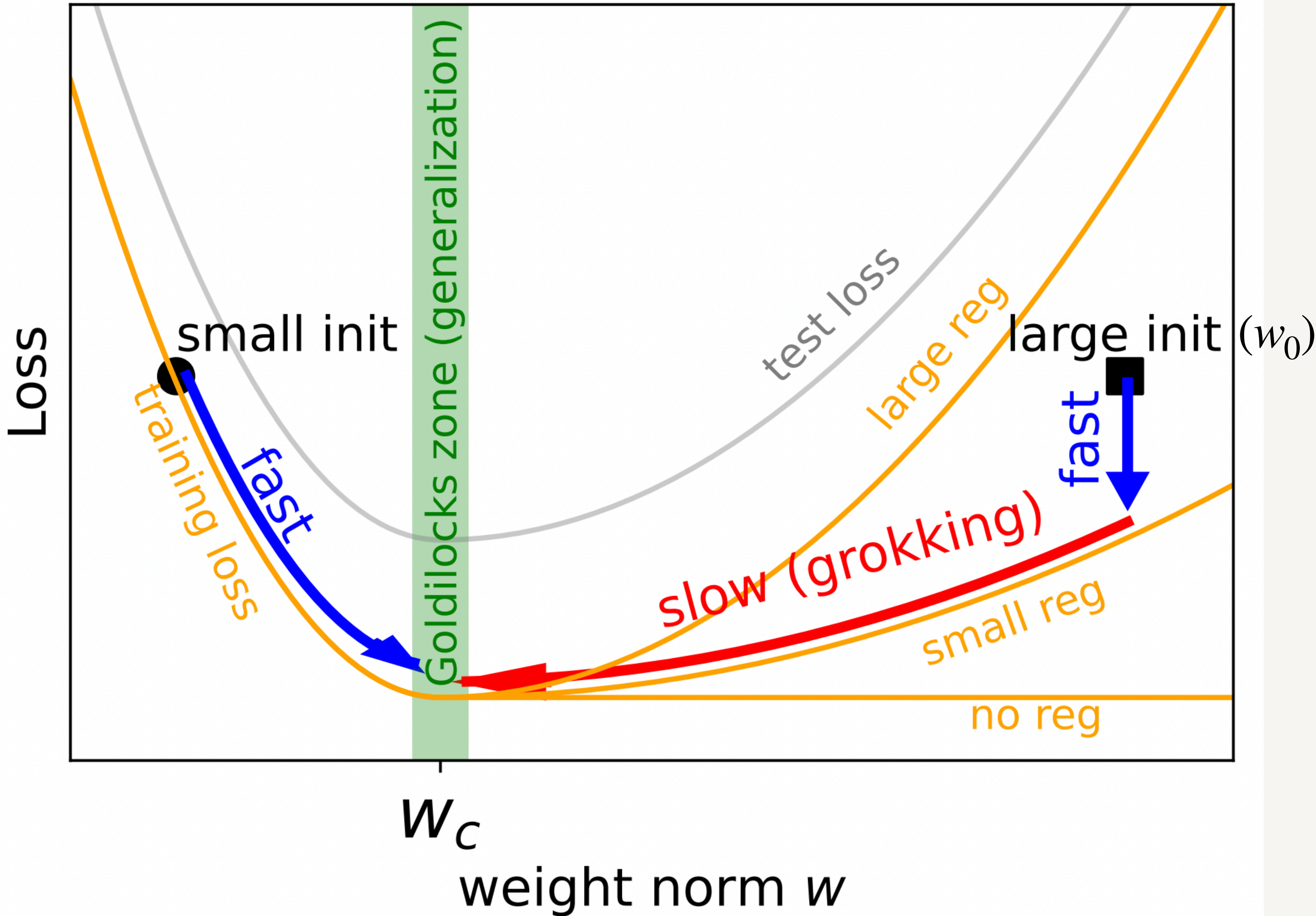
Reduced 1D landscape

$$\tilde{f}(w) \equiv f(\mathbf{w}^*(w)), \quad \text{where } \mathbf{w}^*(w) \equiv \underset{\|\mathbf{w}\|_2=w}{\operatorname{argmin}} l_{\text{train}}(\mathbf{w})$$

↓
Any quantity of interest, e.g., train/test loss/error.



LU mechanism



γ : weight decay

$$\frac{dw}{dt} = -\gamma w - \frac{\partial \tilde{\ell}_{\text{train}}(w)}{\partial w}$$

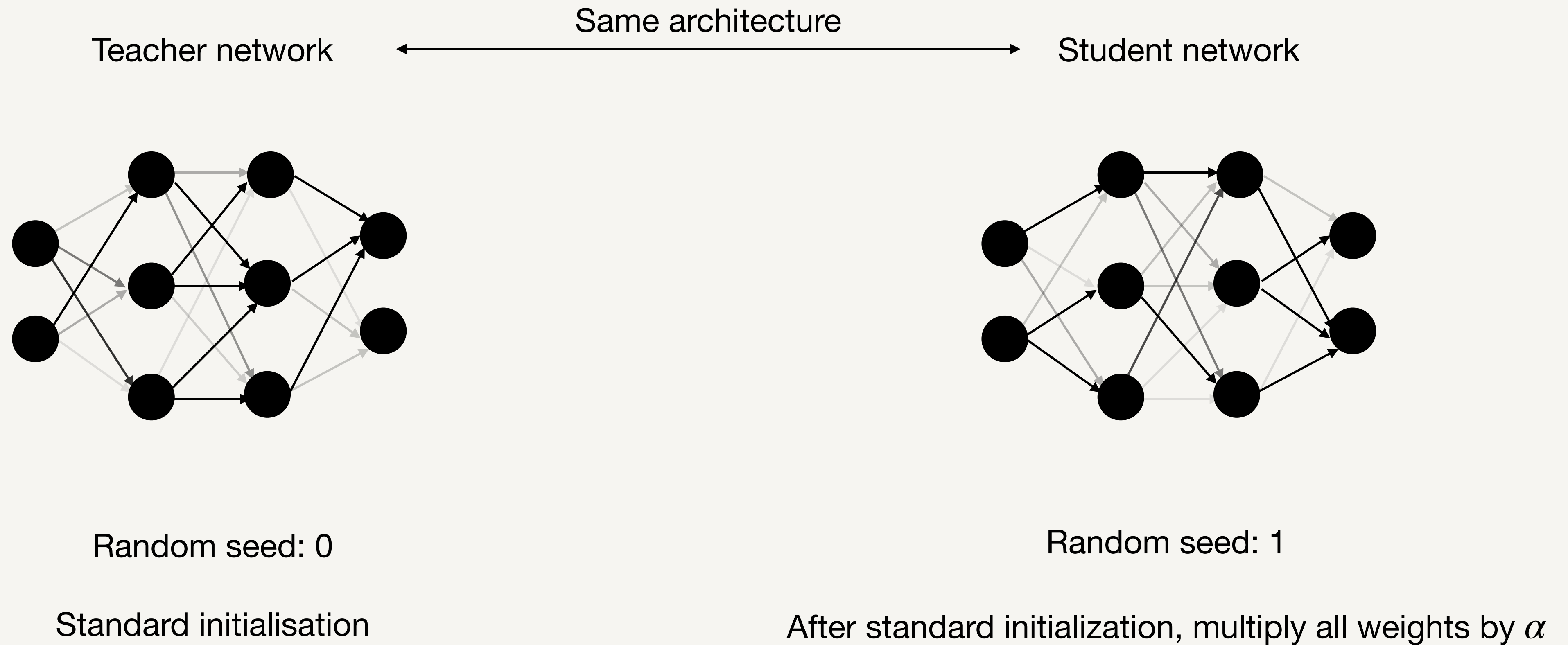
$$\downarrow w > w_c, \frac{\partial \tilde{\ell}_{\text{train}}(w)}{\partial w} = 0$$

$$\frac{dw}{dt} = -\gamma w$$

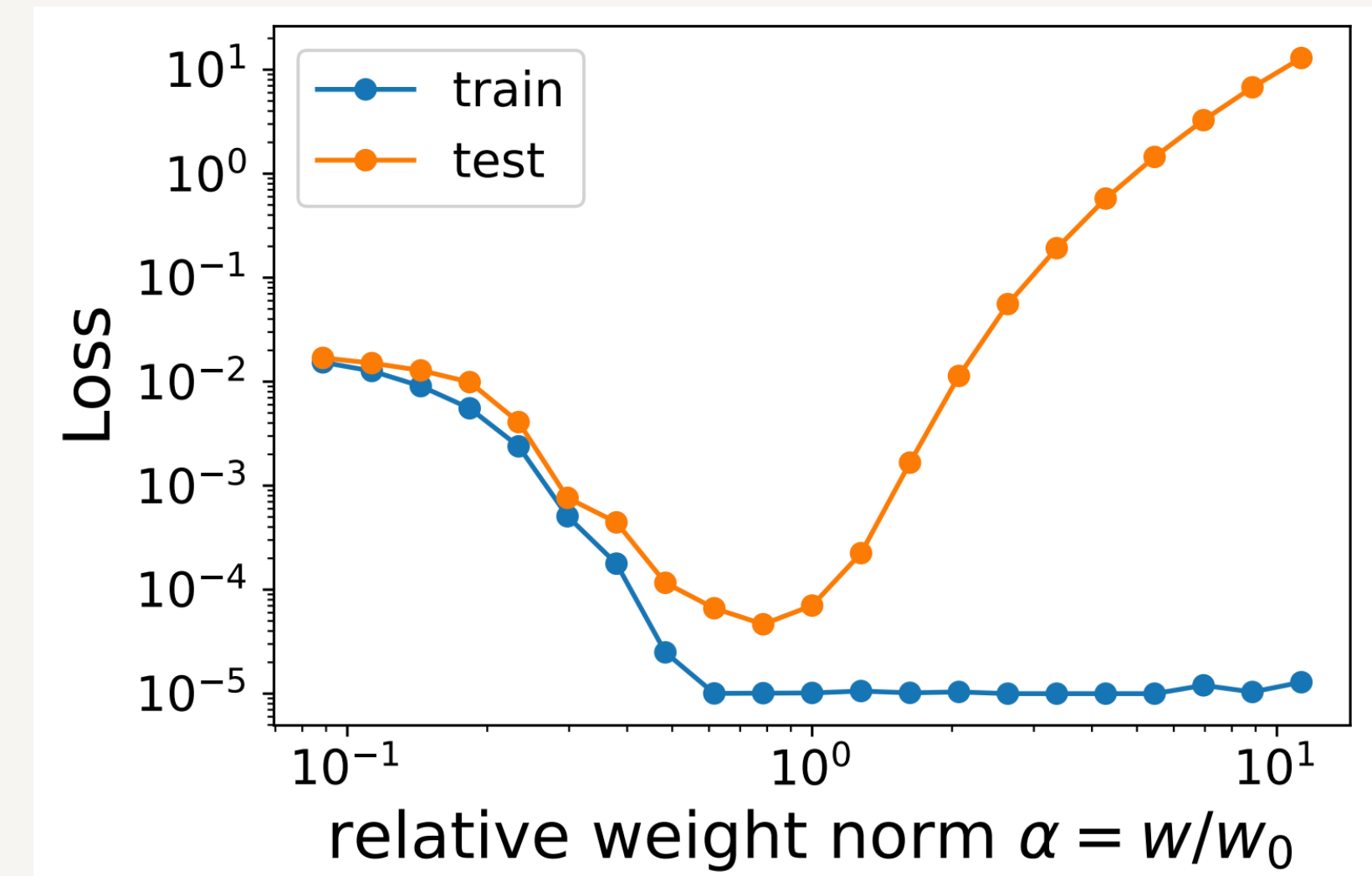
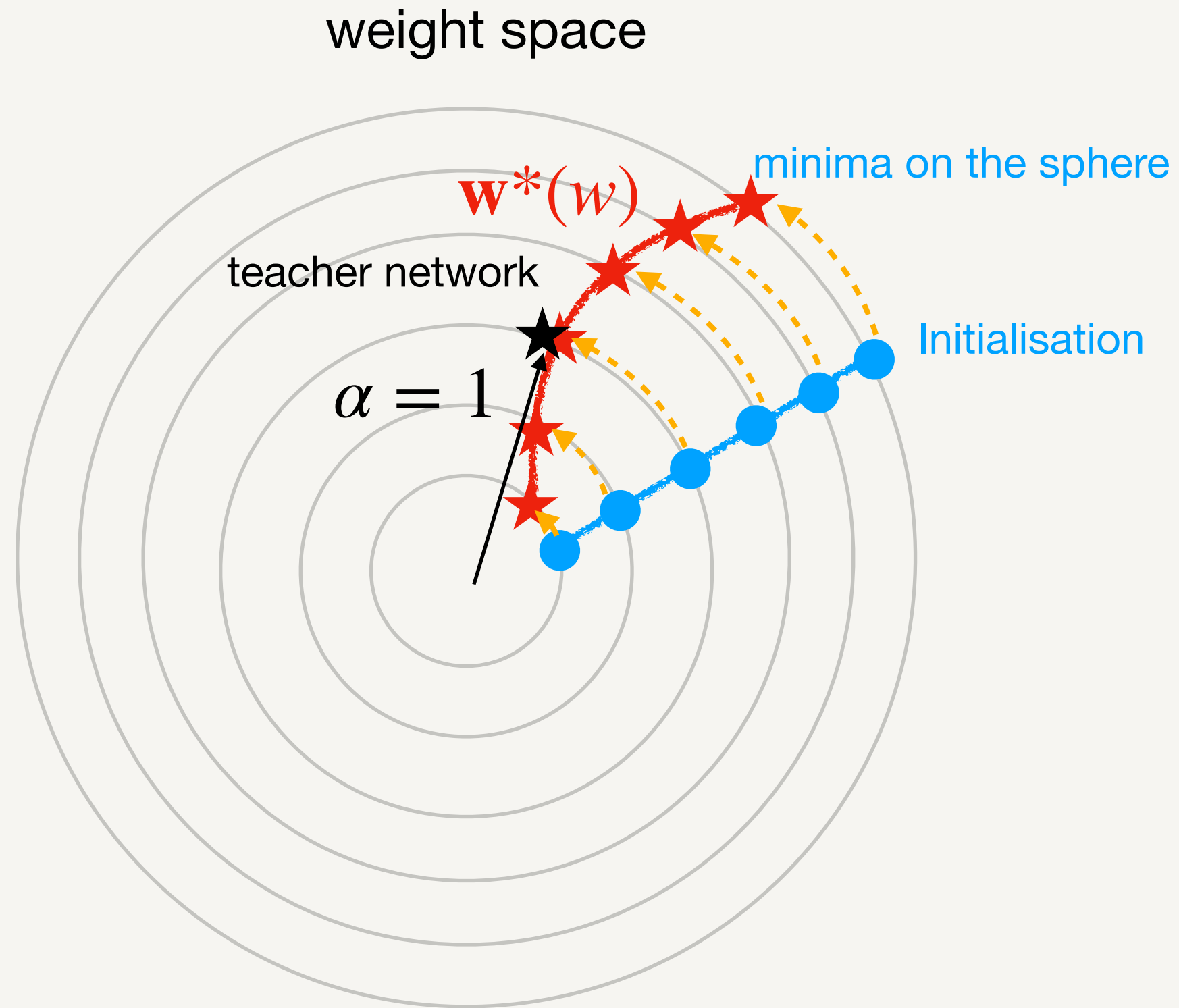
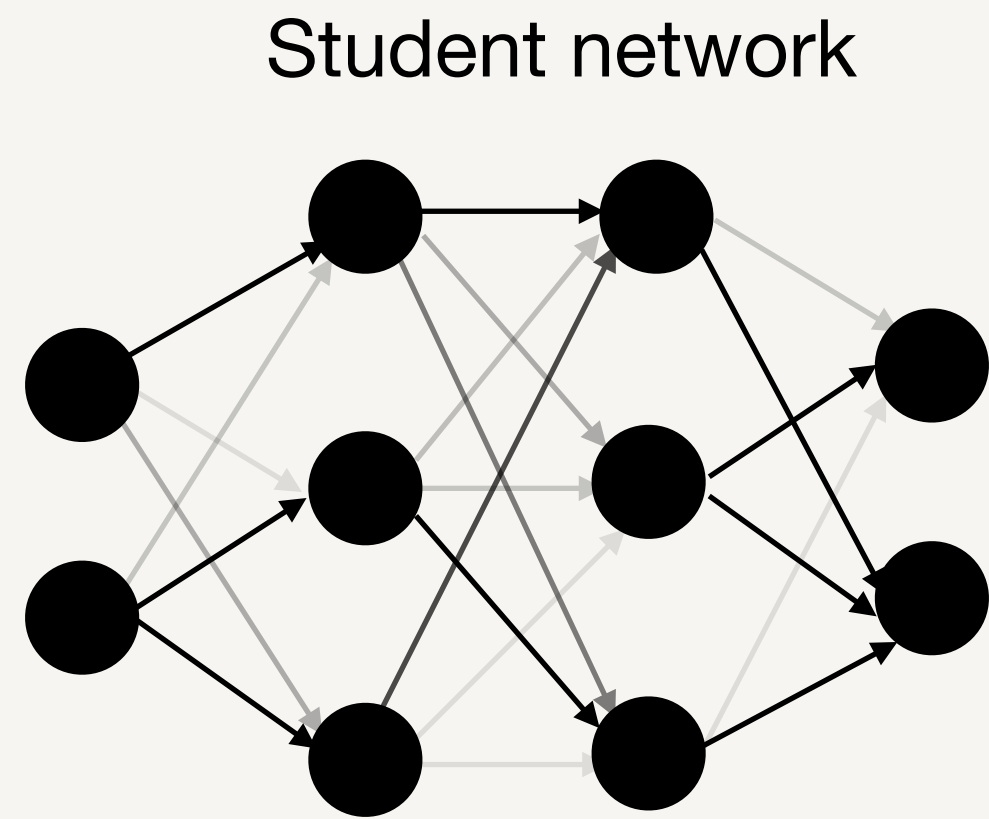
$$w(t) = \exp(-\gamma t)$$

$$t(w_0 \rightarrow w_c) = \log\left(\frac{w_0}{w_c}\right) / \gamma \propto \gamma^{-1}$$

Toy: Teacher-student



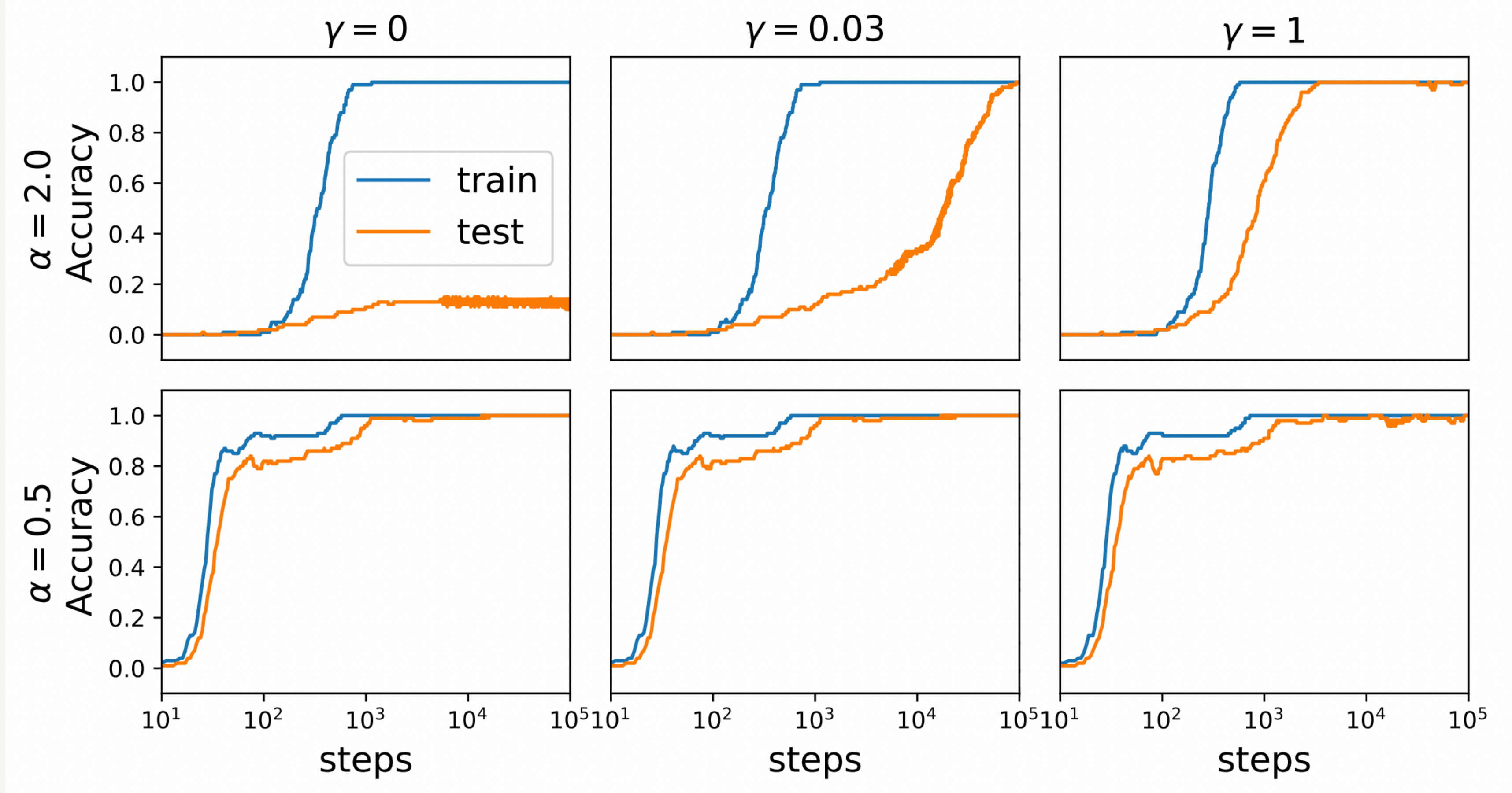
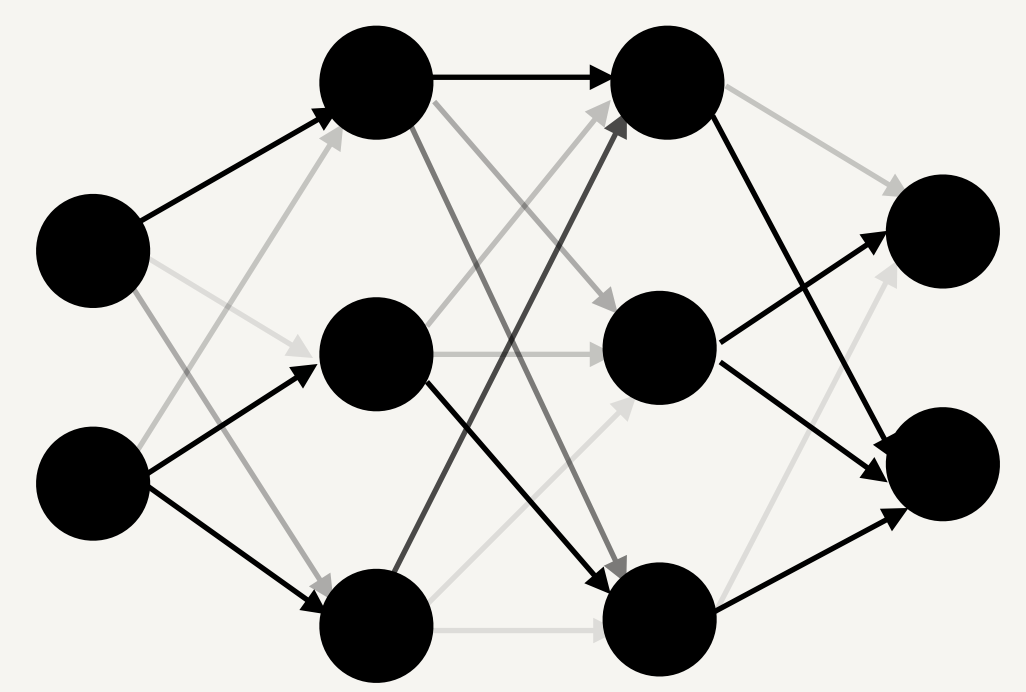
Teacher-student: Landscape



Teacher-student: Grokking

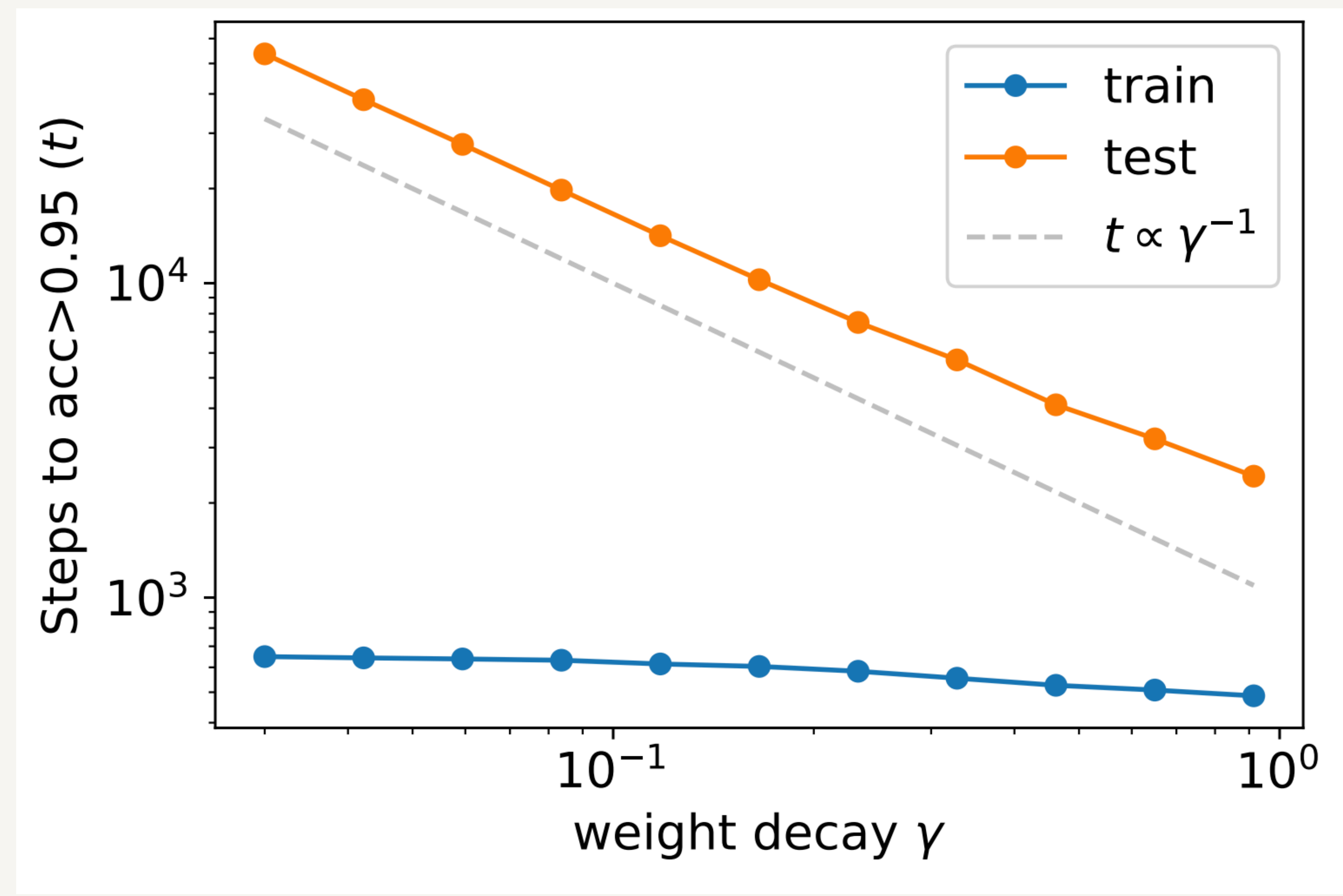
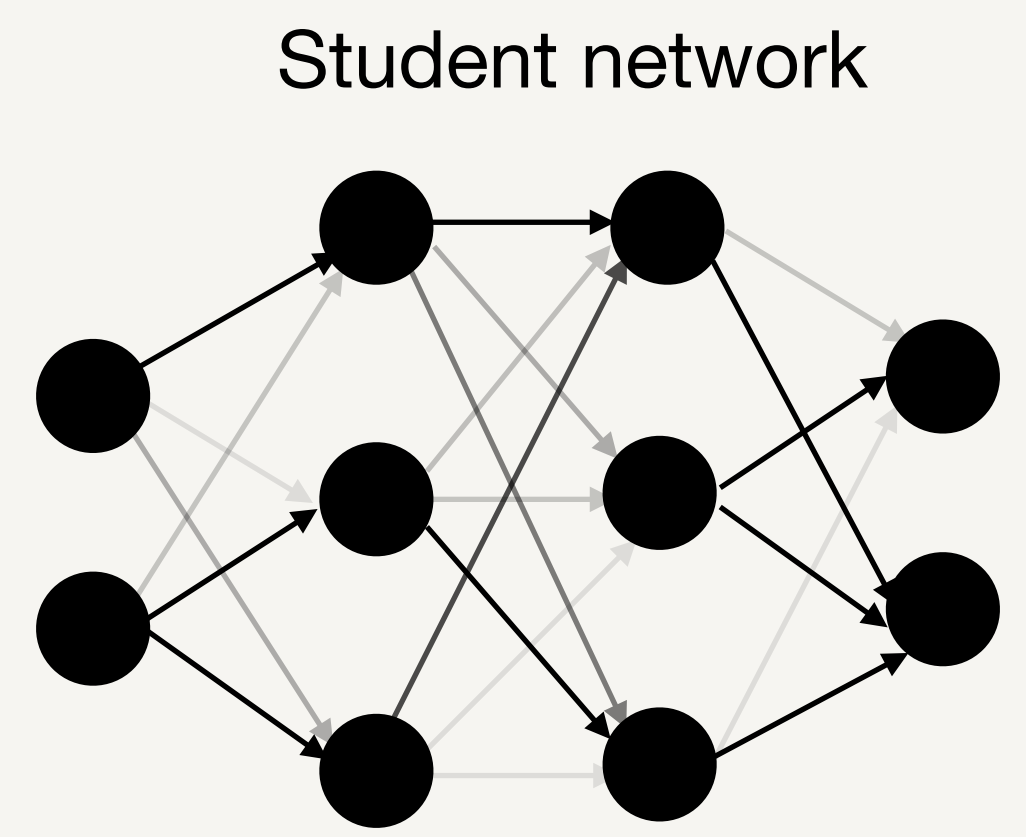
Note: weight norm is not constrained here.

Student network



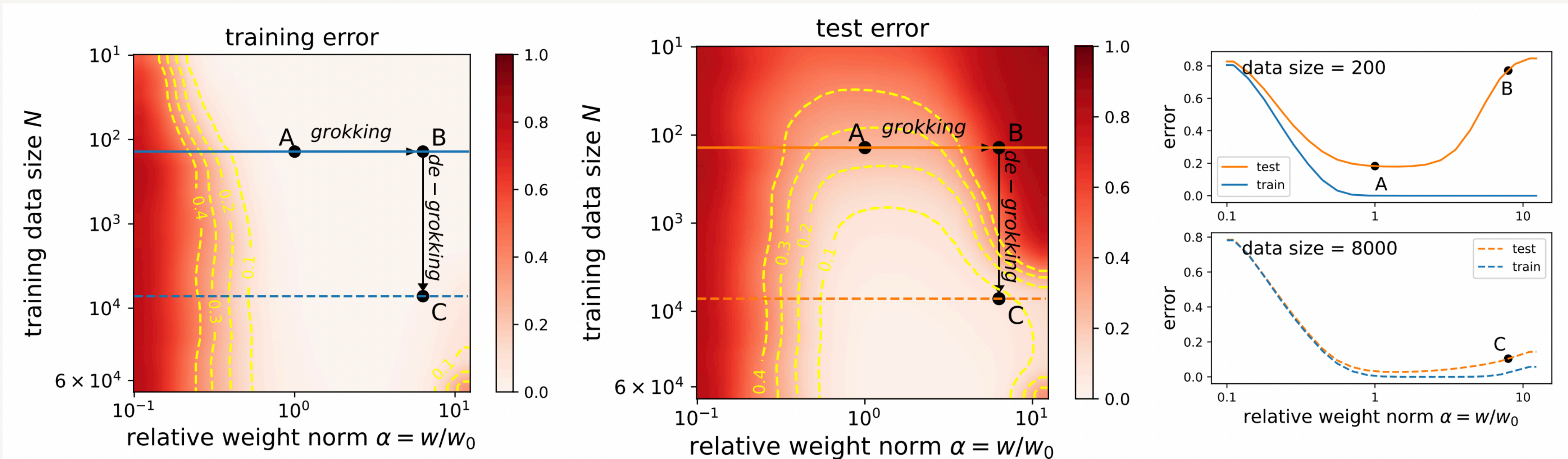
Teacher-student: Grokking

Note: weight norm is not constrained here.

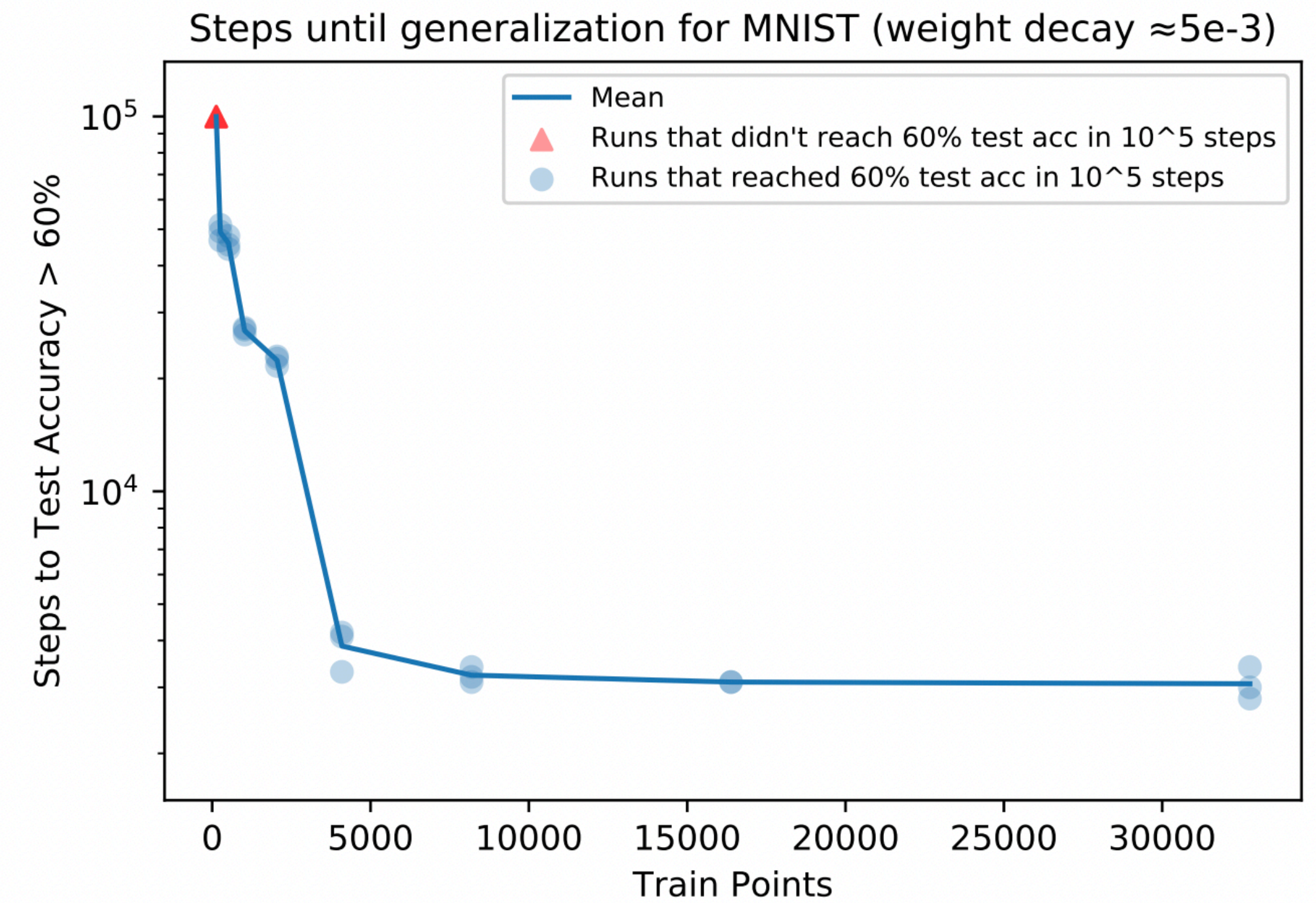
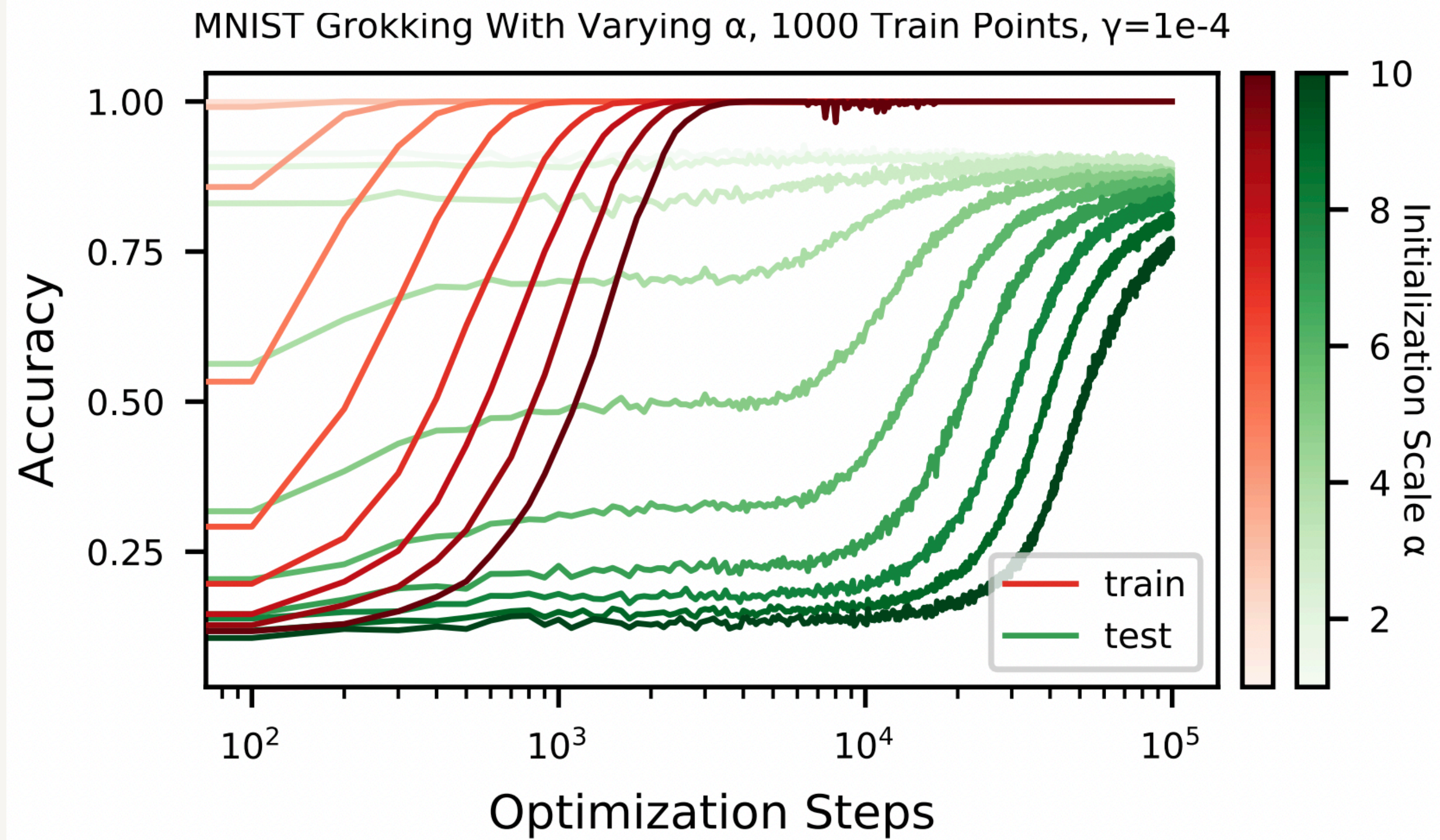


MNIST: landscape analysis

Model: MLP

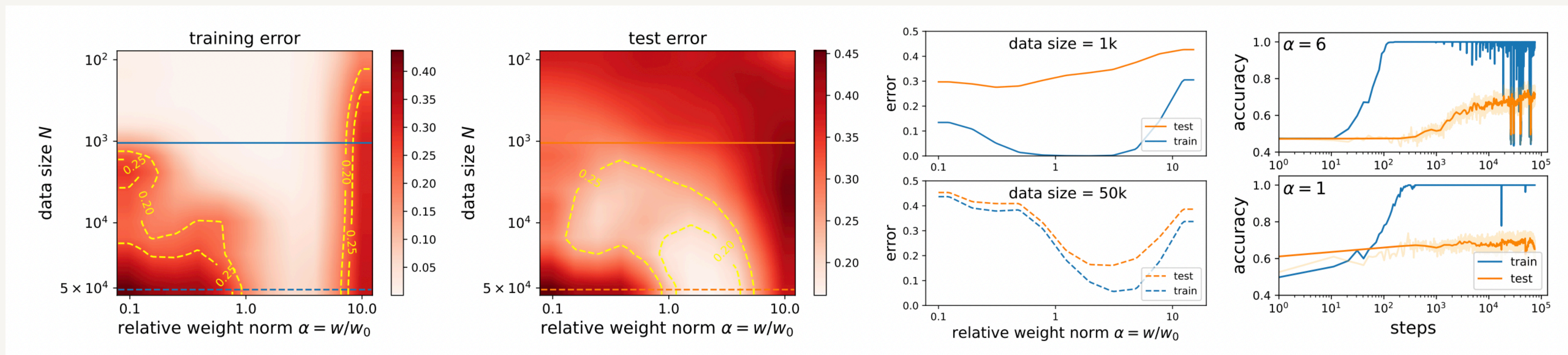


MNIST: Grokking

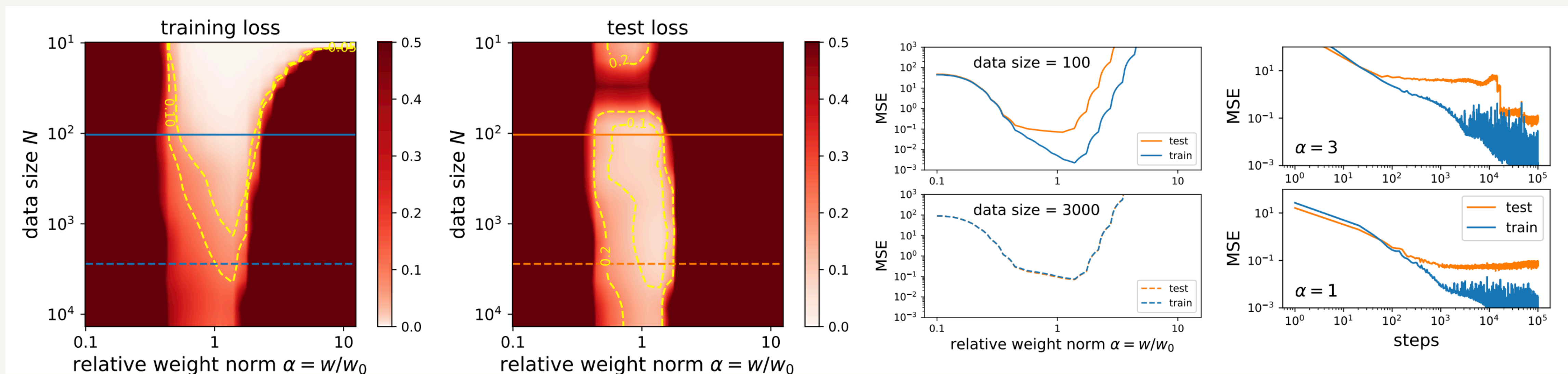


More datasets

IMDb (Sentiment Analysis) + LSTM



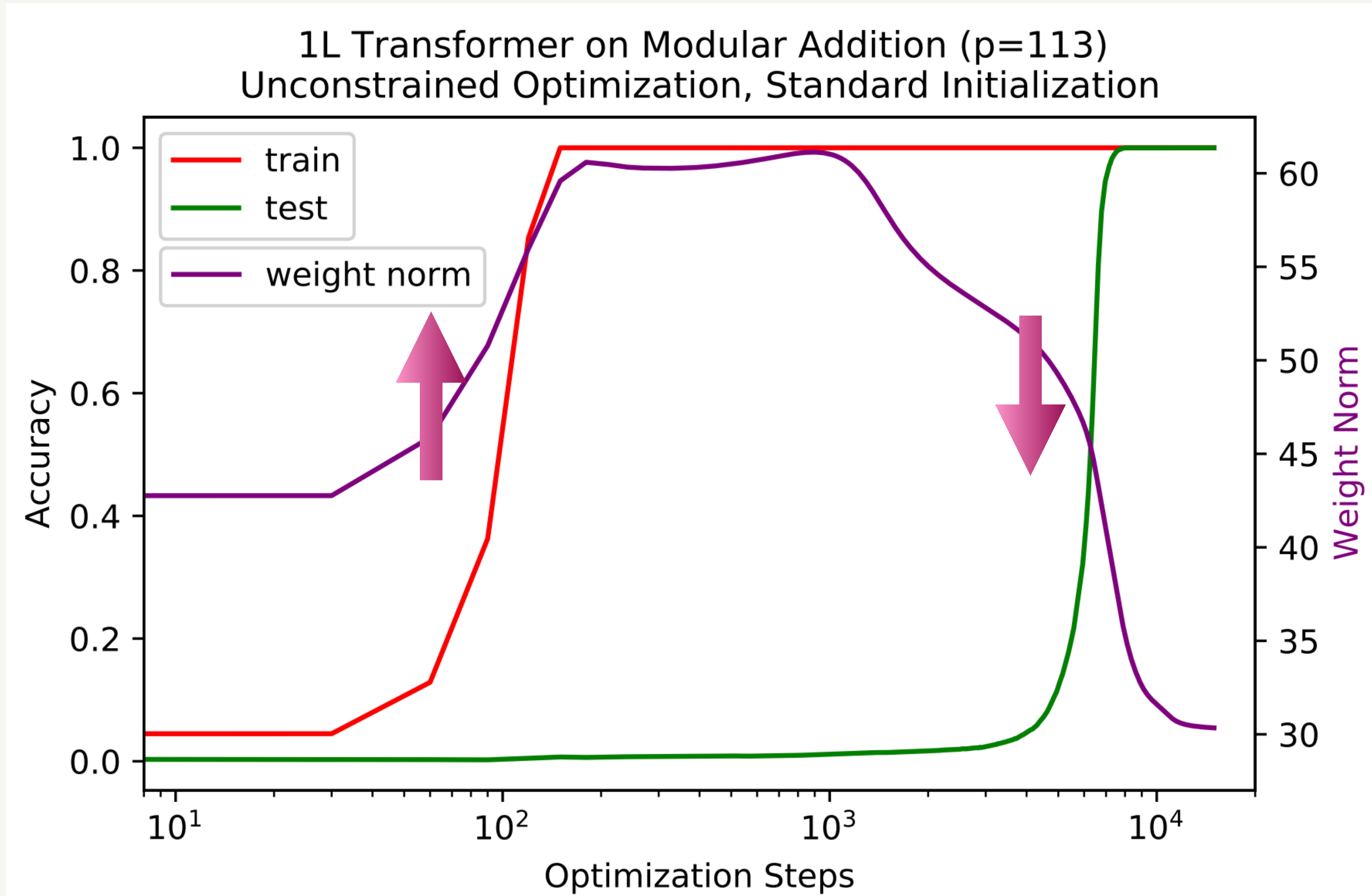
QM9 (Molecule) + Graph Convolutional Neural Network



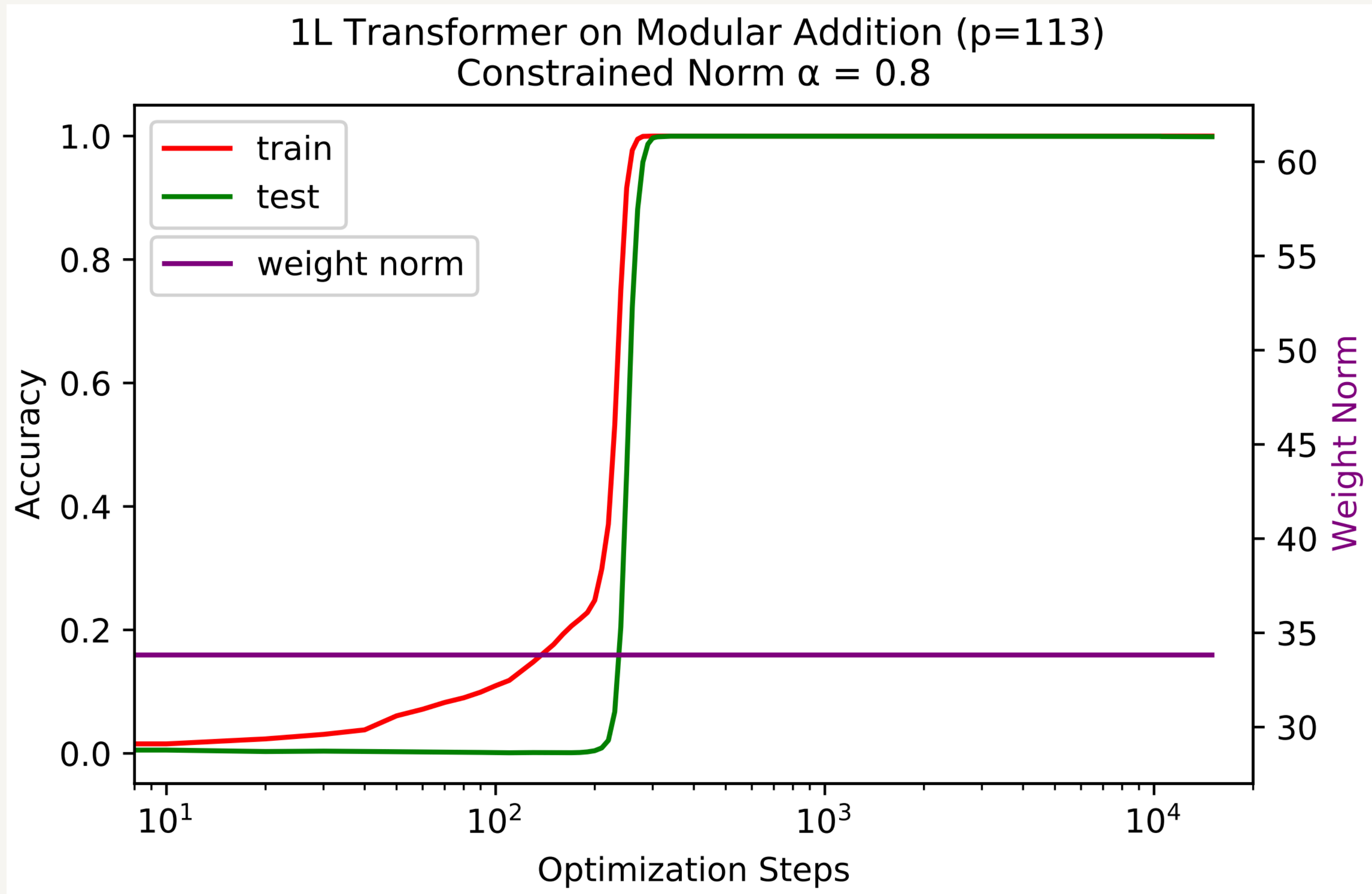
Wait a second ...

1. For algorithmic datasets, standard initialisation is sufficient to produce grokking. But on standard datasets we induce grokking by manually increasing initialisation scale.
2. Since we can induce grokking on standard datasets, we you remove grokking from algorithmic datasets?

Modular addition: Weight norm evolution



Remove grokking by small & constrained scale



Outlook

1. Grokking in large language models?
2. More applications of reduced loss landscape?
3. Theory of LU mechanism?
4. Task-dependent Initialisation?

Physics & Deep Learning

Grokking

- Knowledge**
1. Thermodynamics (phase diagrams)
 2. Classical mechanics (particle interaction)
- Approach**
3. Identifying useful variables (weight norm)
 4. Toy examples & controlled experiments

Backup:

Representation learning vs grokking

Algorithmic: Representation learning

representation messiness m

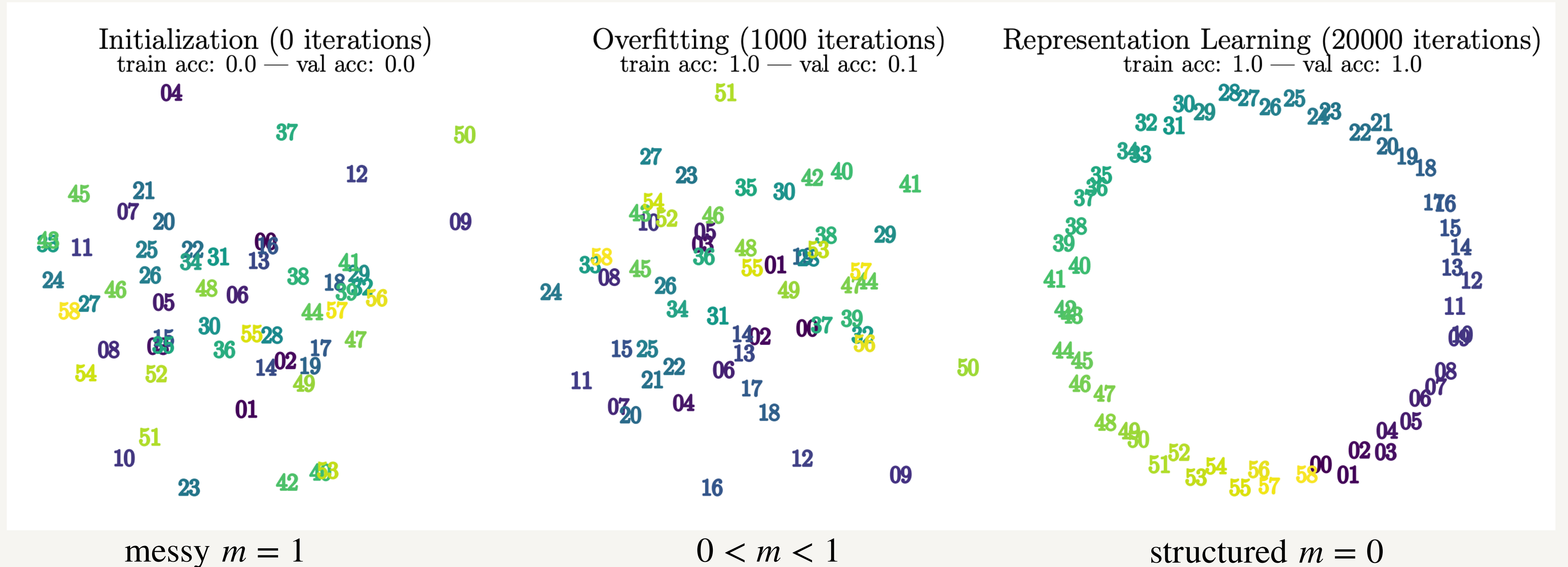
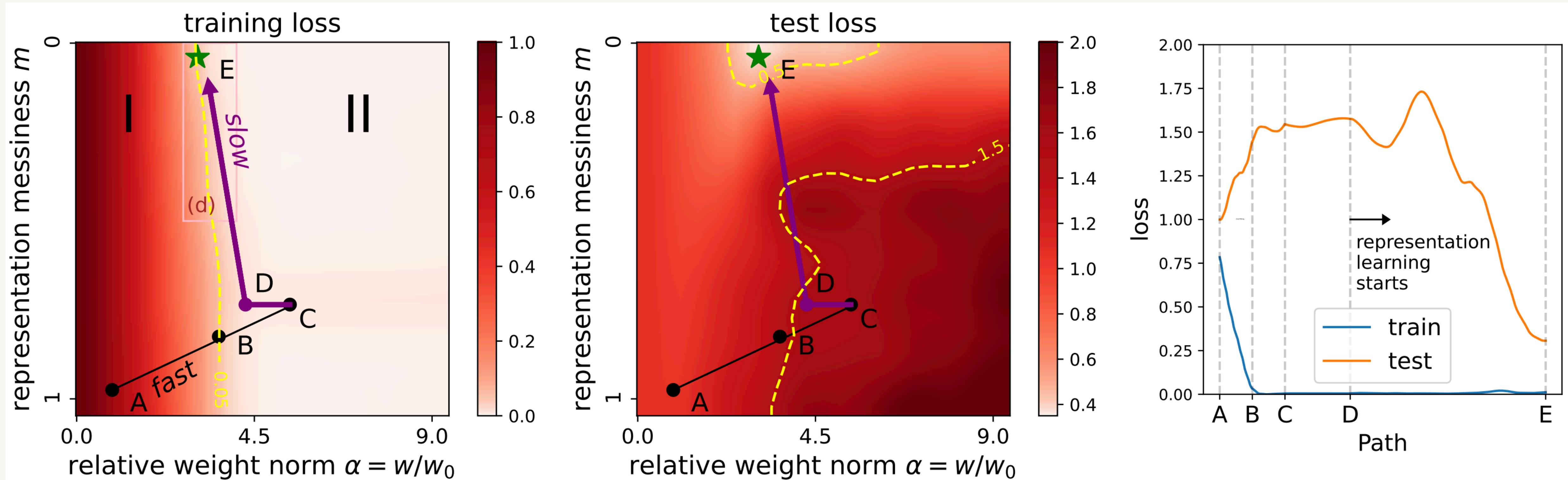


Figure 1 of “Towards Understanding Grokking: An Effective Theory of Representation Learning”, NeurIPS 2022. Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams

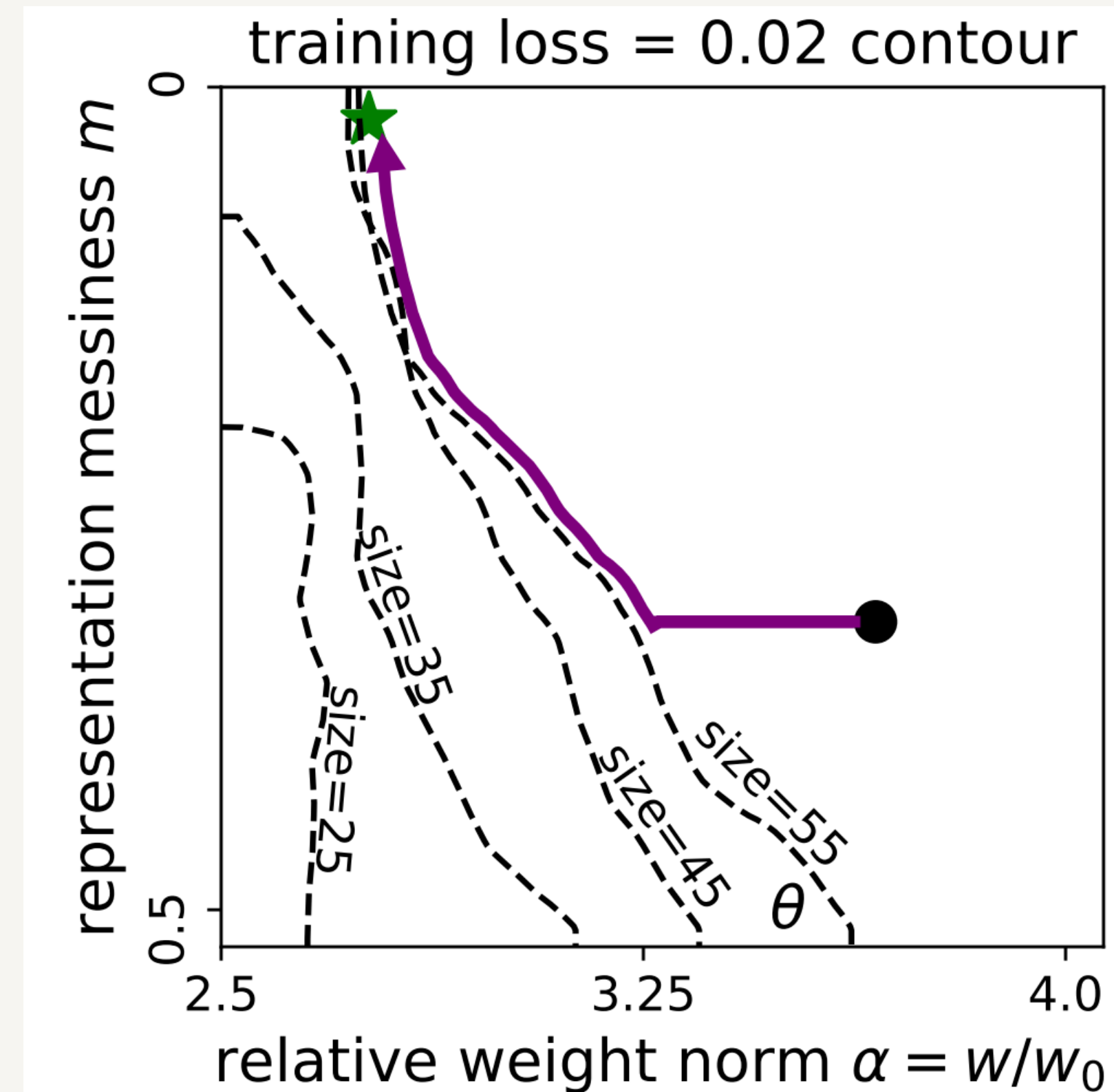
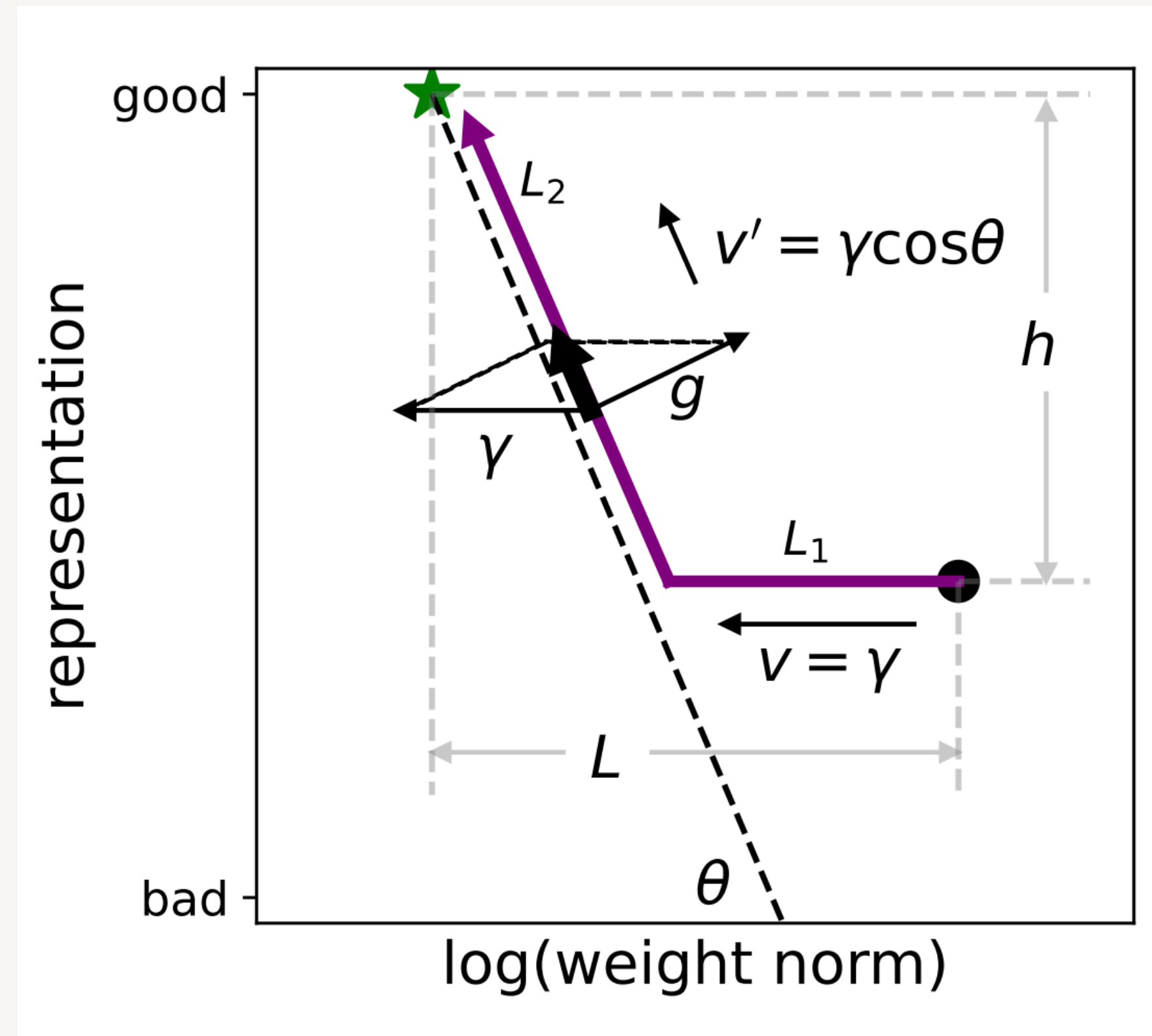
Algorithmic: Landscape analysis

$$\mathbf{w}^*(w, m) \equiv \operatorname{argmin}_{\|\mathbf{w}\|_2=w} l_{\text{train}}(\mathbf{w}, m), \quad \tilde{l}_{\text{train}}(w, m) \equiv l_{\text{train}}(\mathbf{w}^*, m), \quad \tilde{l}_{\text{test}}(w, m) \equiv l_{\text{test}}(\mathbf{w}^*, m)$$

$$\frac{dw}{dt} = -\eta_D \left(\frac{\partial \tilde{l}_{\text{train}}}{\partial w} + \gamma w \right), \quad \frac{dm}{dt} = -\eta_R \frac{\partial \tilde{l}_{\text{train}}}{\partial m},$$



Algorithmic: data size/weight decay dependence



$$t = \frac{L + h \tan \theta}{\gamma}$$

\nearrow data size $\uparrow \rightarrow \theta \downarrow \rightarrow t \downarrow$
 \searrow $\gamma \uparrow \rightarrow t \downarrow$

MNIST: landscape analysis

$$\mathbf{R} = m\mathbf{R}_{\text{raw}} + (1 - m)\mathbf{R}_{\text{linear}},$$

