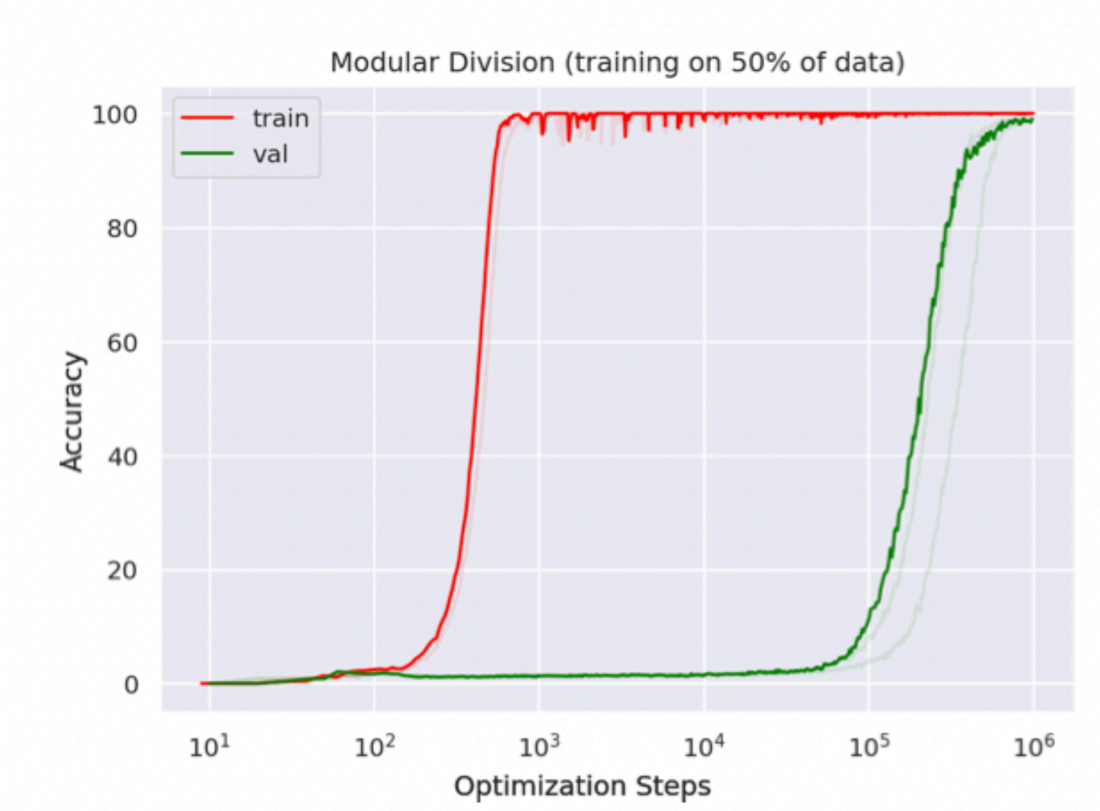
Grokking

Ibragimov Aidar 2023

What is Grokking?



Long after severely overfitting, validation accuracy sometimes suddenly begins to increase from chance level toward perfect generalization. We call this phenomenon ‘**grokking**’

Algorithmically generated datasets

These datasets are typically not collected from the real world but are instead generated through simulations, mathematical models, or other computational methods. The purpose of creating such datasets is to have control over various parameters, conditions, and data characteristics to test, train, or evaluate machine learning models, algorithms, or systems.

Algorithmically generated datasets

Examples

- binary operation tables of the form a ○ b = c where a, b, c are discrete symbols with no internal structure, and ○ is a binary operation. Examples of binary operations include addition, composition of permutations, and bivariate polynomials:

  x+y (mod p), xy (mod p), permutations

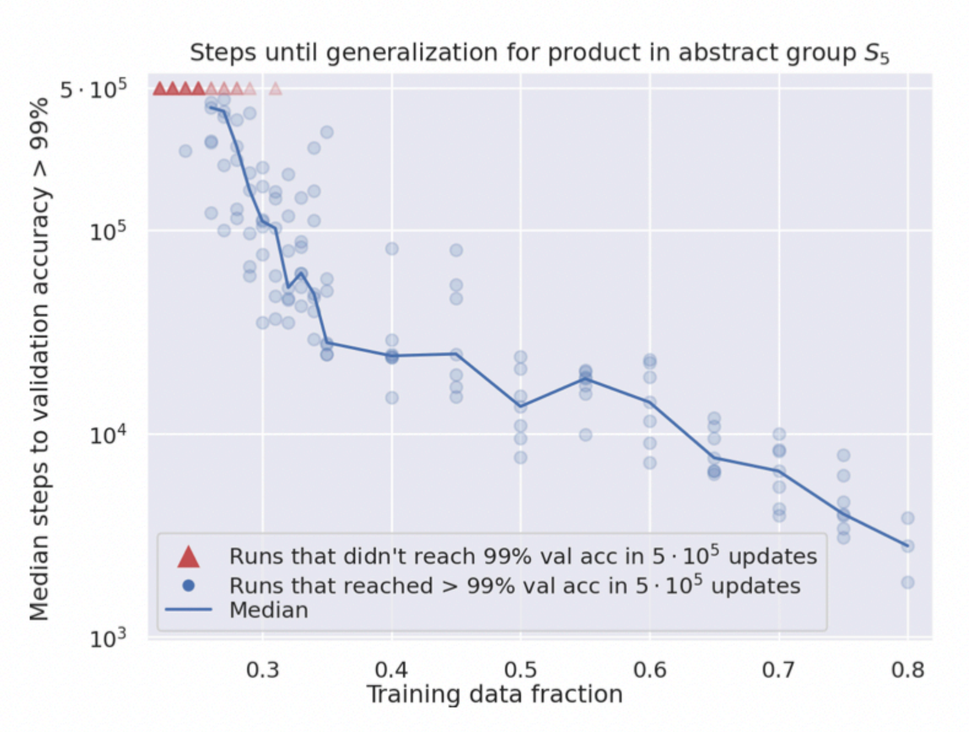
★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

Algorithmically generated datasets

Examples

- 5 digit addition: 1|3|4|5|2|+|5|8|3|2|1|=|0|7|1|7|7|3
- Predicting Repeated Subsequences: 7 2 8 3 1 9 3 **8 3 1 9** 9 2 5 END
- Finding the max element in a sequence: START 0 4 7 2 4 **14** 9 7 2 5 3 END

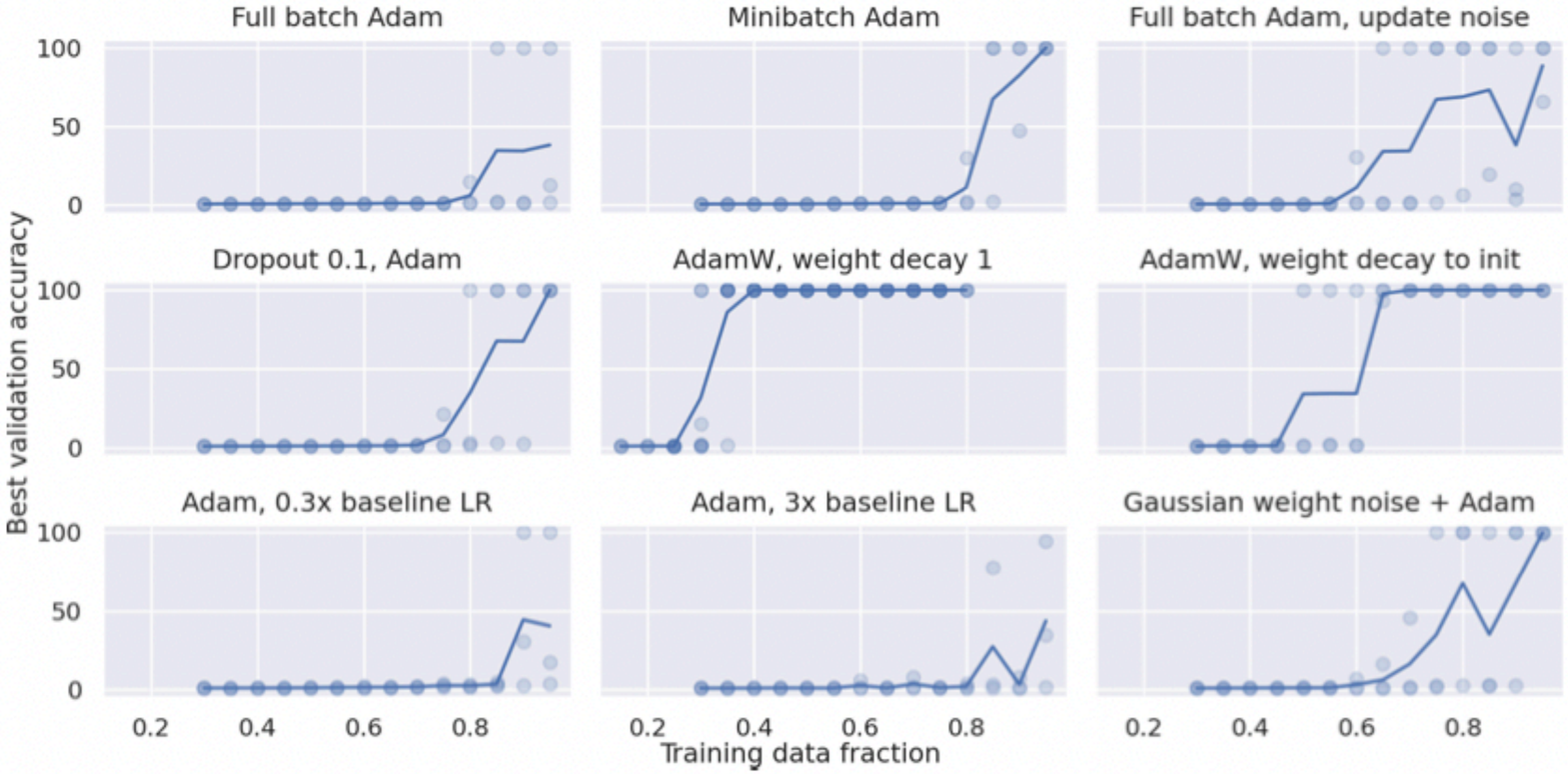
Grokking experiments



In the vicinity of 25-30% of data, a decrease of 1% of training data leads to an increase of 40-50% in median time to generalization. While the number of steps until validation accuracy > 99% grows quickly as dataset size decreases, the number of steps until the train accuracy first reaches 99% generally trends down as dataset size decreases and stays in the range of $10^3 - 10^4$ optimization steps.

Exponential increase in optimization time until reaching generalization as dataset size decreases on all the algorithmic tasks

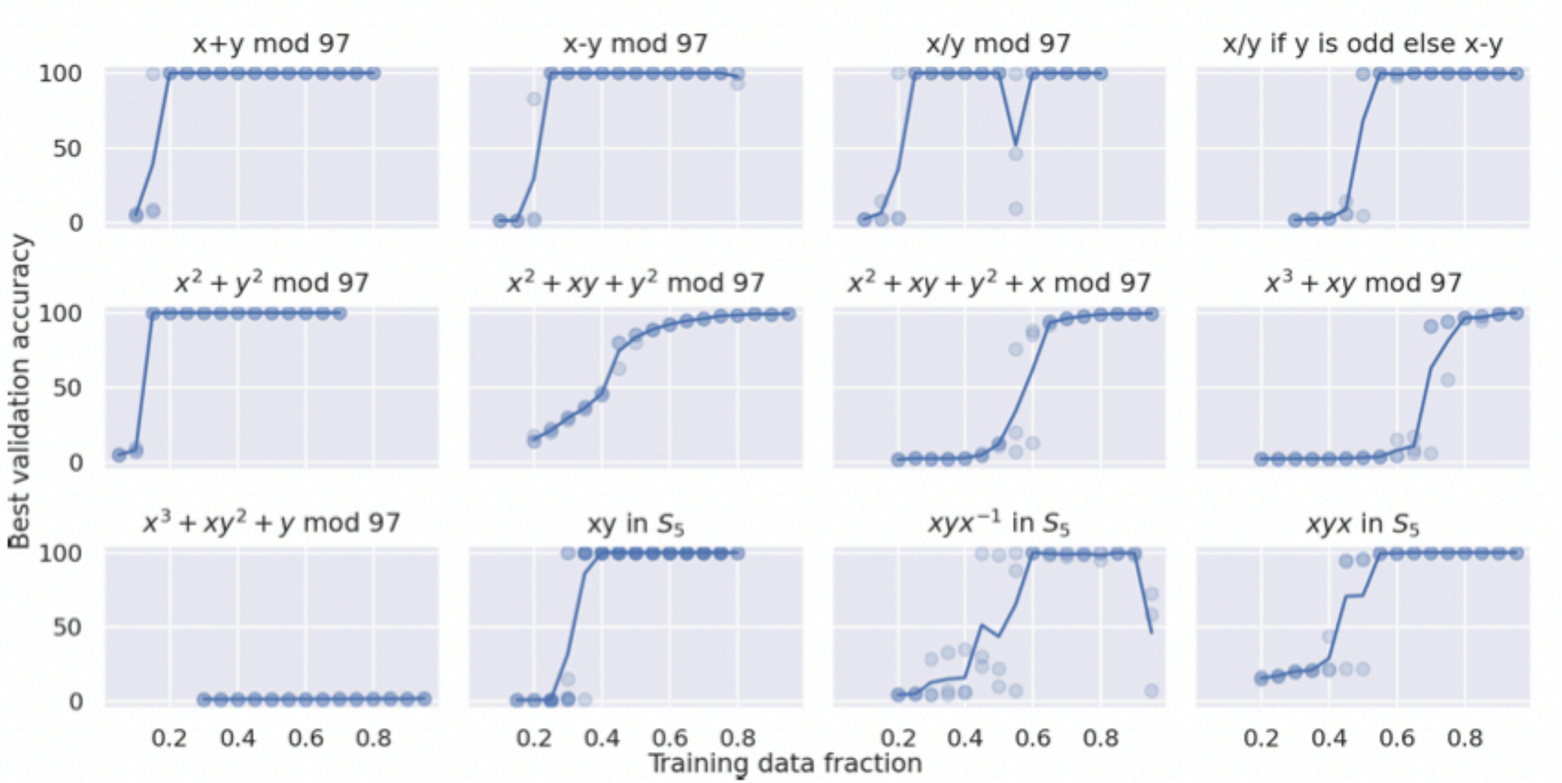
Grokking experiments



Optimization budget of 10^5steps for the problem of learning the product in the abstract group S_5

Weight decay improves generalization the most, but some generalization happens even with full batch optimizers and models without weight or activation noise at high percentages of training data.

Grokking experiments



Generalization happens at higher percentages of data for intuitively more complicated and less symmetrical operations.

Grokking experiments

x - y (mod 97), x / y (mod 97) - is very similar.

97 is prime, so every nonzero residue modulo a prime can be represented as a power of a primitive root.

x / y (mod 97) $\sim$ x - y (mod 96) $\sim$ x - y (mod 97)

Grokking experiments

The matrix of the output layer for the case of S_5

clusters of permutations, and each cluster is a coset of the subgroup $\langle (0, 3)(1, 4), (1, 2)(3, 4) \rangle$ or one of its conjugates.

Grokking explanations

Possible explanations:

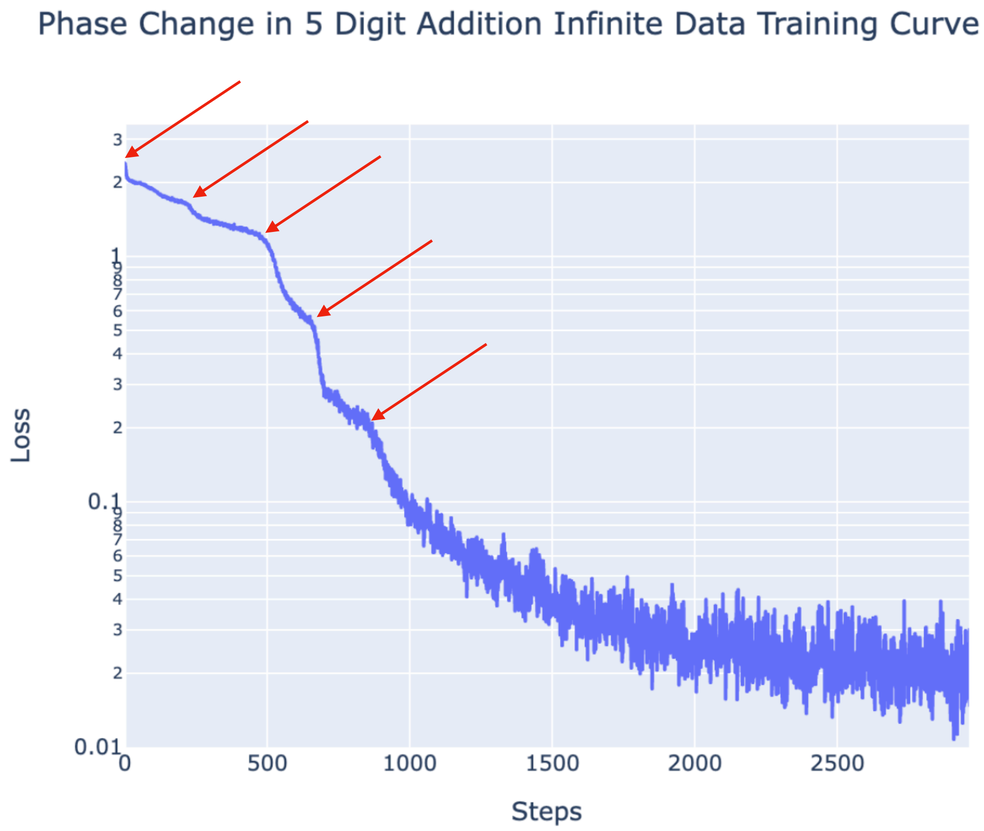
- Grokking has a deep relationship to **phase changes.**
- Grokking can result from a mismatch between training and test loss against model **weight norm**.

Grokking and phase changes

Phase changes, ie a sudden change in the model's performance for some capability during training, are a general phenomena that occur when training models, that have also been observed in large models trained on non-toy tasks. For example, the sudden change in a transformer's capacity to do in-context learning when it forms induction heads.

One finding in DeepMind's AlphaZero Interpretability paper was that there is a phase change in the model's capabilities, where it learns to represent a lot of chess concepts around step 32,000.

Grokking and phase changes



Grokking and phase changes

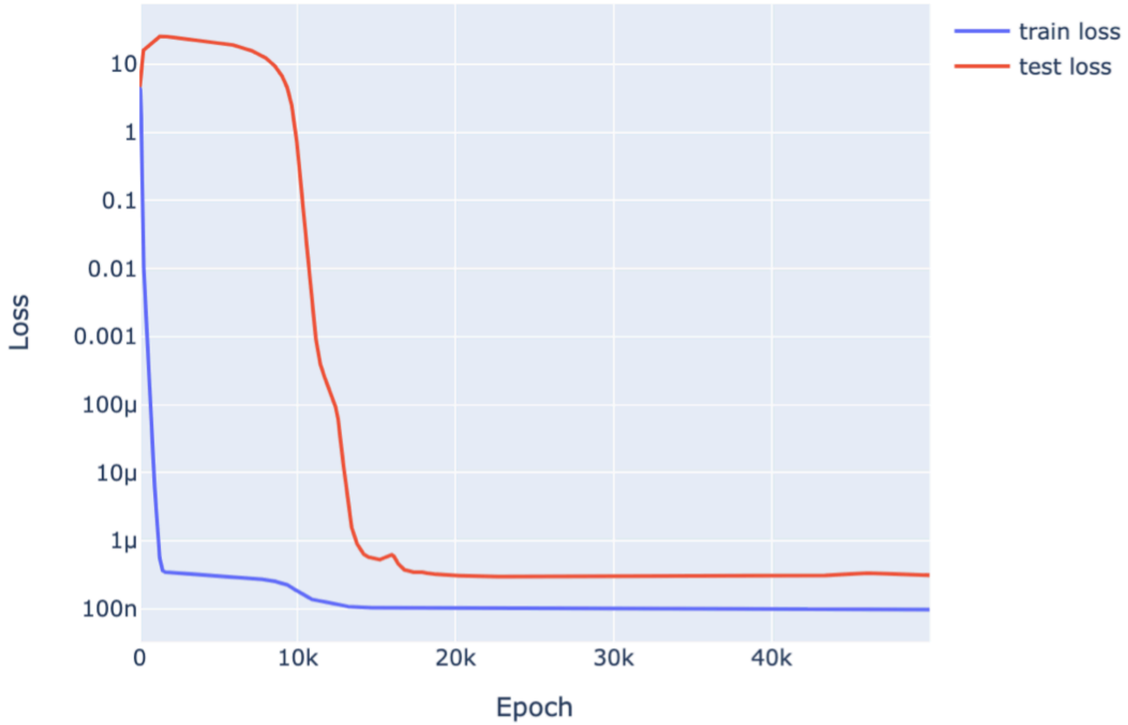
Intuition

- The memorised solution is «easier to reach» on small data, and so is reached first
- Due to the regularisation, the model still prefers the generalising solution to the memorised solution
- Phase change indicates that the generalising solution is «hard to reach» in some sense

} **Grokking**

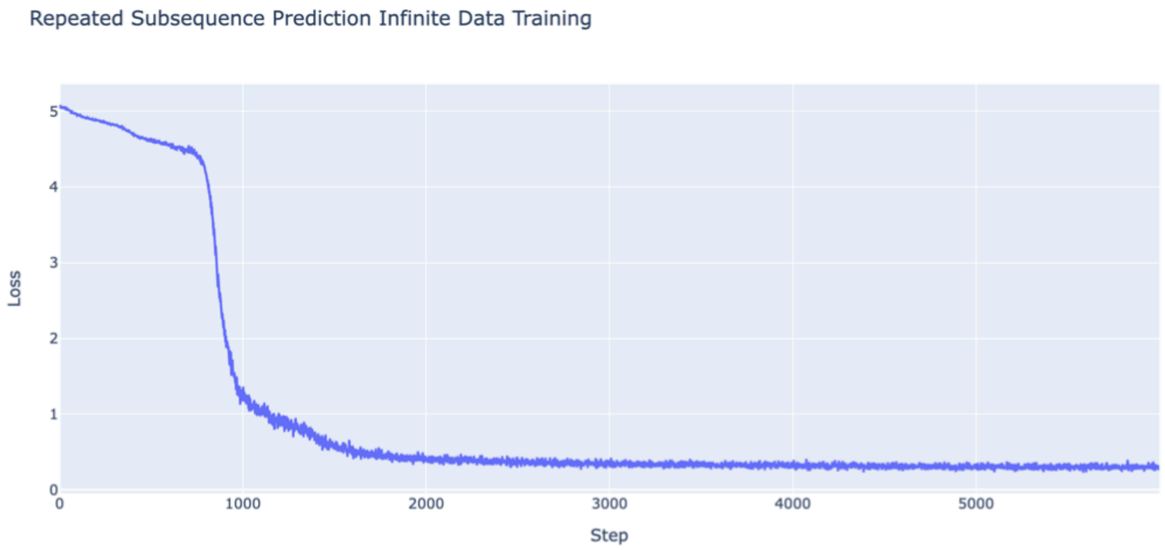
Train a model on a problem that exhibits phase changes even when given infinite training data, and train it with regularisation and limited data.

Grokking and phase changes

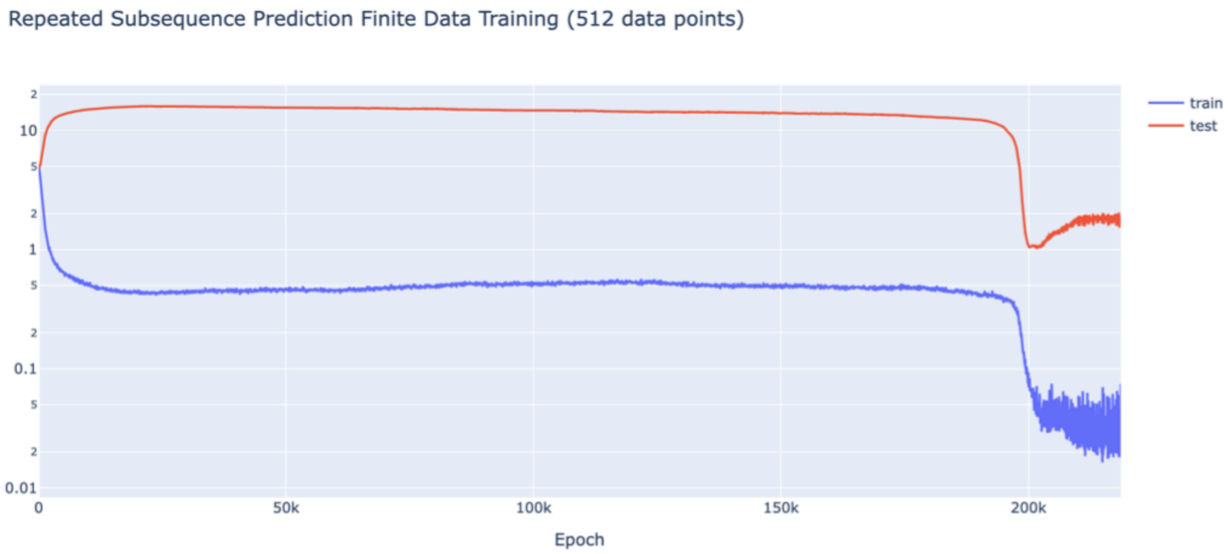


A training curve for a 1L Transformer trained to do addition mod 113, trained on 30% of the 113^2 pairs - it shows clear grokking

Grokking and phase changes

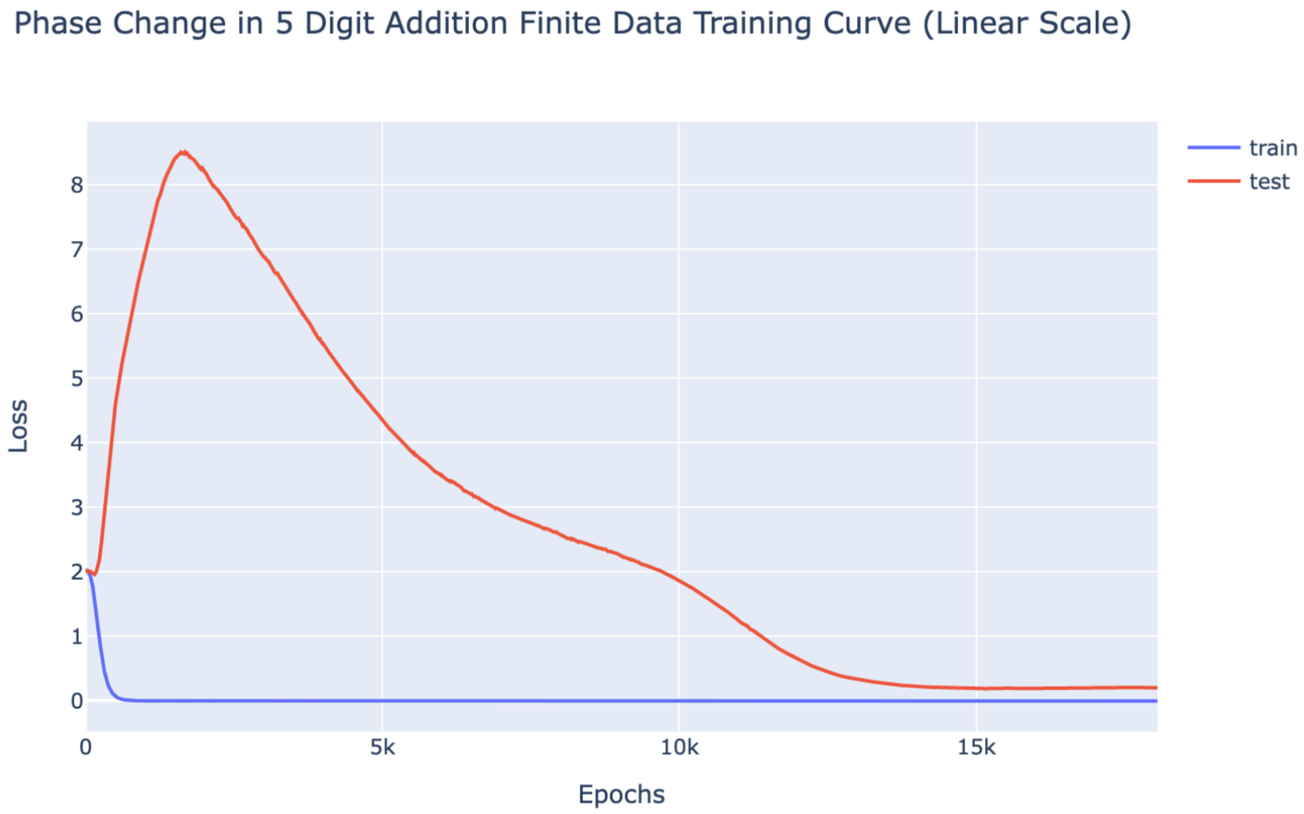


Loss curve for predicting repeated subsequences in a sequence of random tokens in a 2L attention only transformer on infinite data - shows a phase change



Loss curve for predicting repeated subsequences in a sequence of random tokens in a 2L attention-only transformer given 512 training data points - shows clear grokking.

Grokking and phase changes



Grokking shown when training 1L Transformer to do 5 digit addition on 700 training examples. Shown with a linear y axis to make grokking visually clear, train loss plateaus at 3×10^{-8}, test loss plateaus at 0.2

Grokking and weight norm

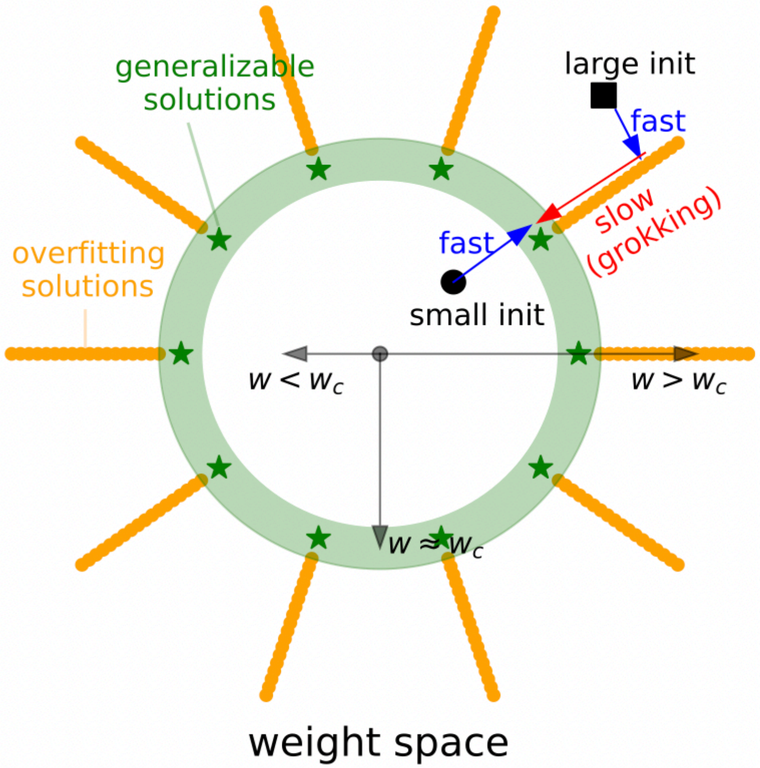
model weights - w

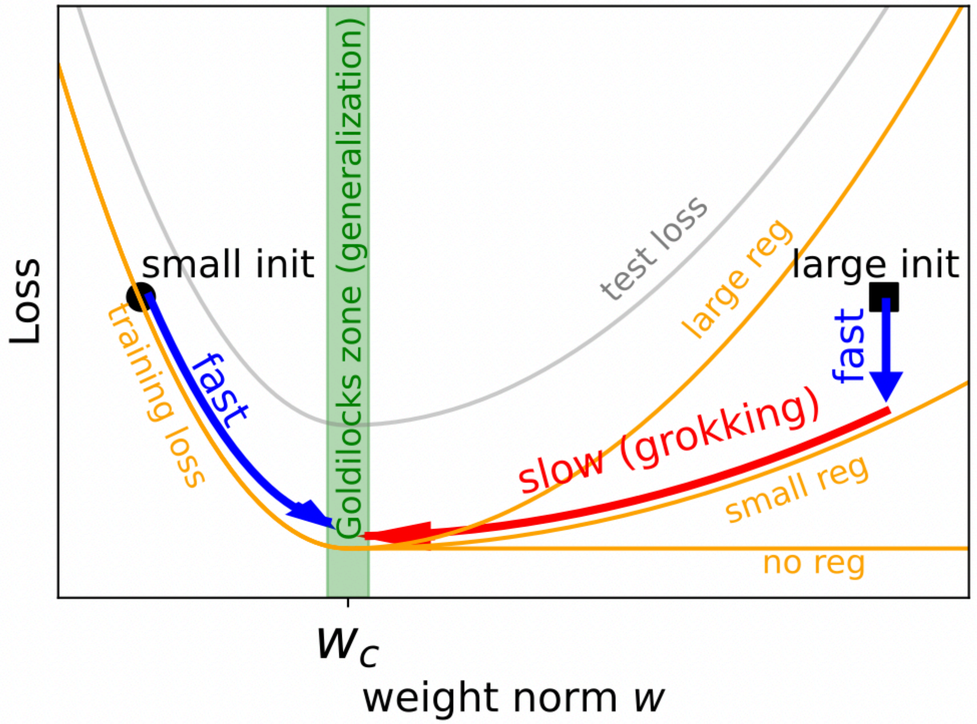
weight norm $-\ \omega = ||w||_2$

weight decay $-\ \gamma$

initial weight multiplier - α (multiply initial weight by alpha)

Grokking and weight norm





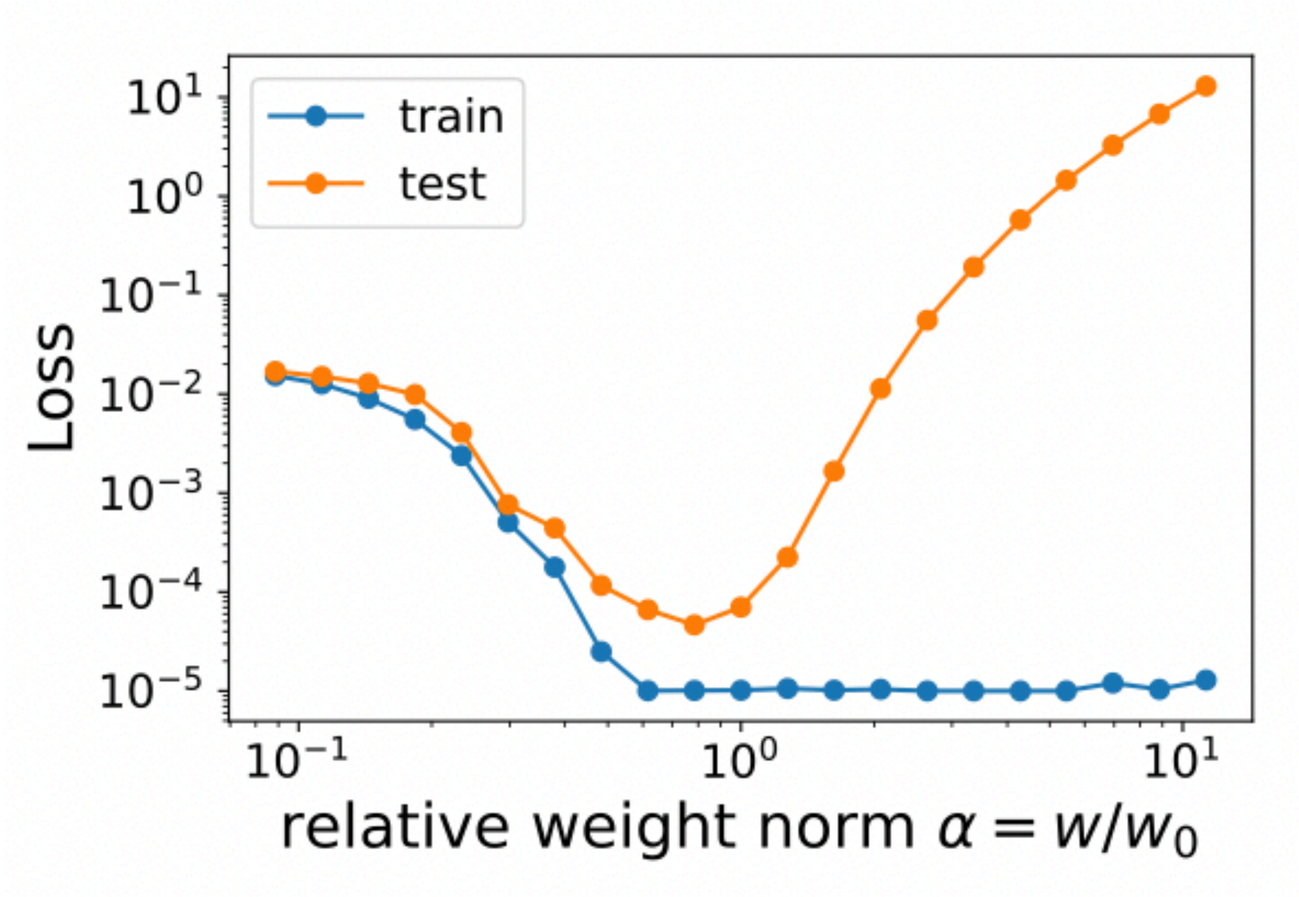
Goldilocks zone - generalization zone

$$\omega(t) \approx exp(-\gamma t)\omega_0$$

The standard initialization schemes typically initialize ω no larger than ω_c

Grokking and weight norm

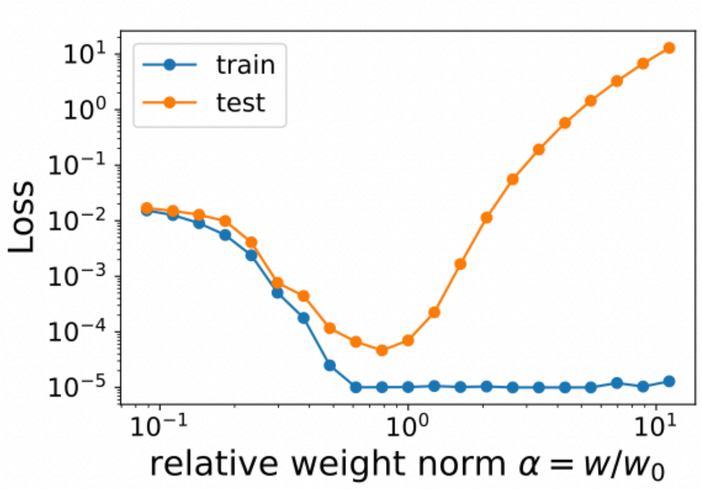
LU-mechanism

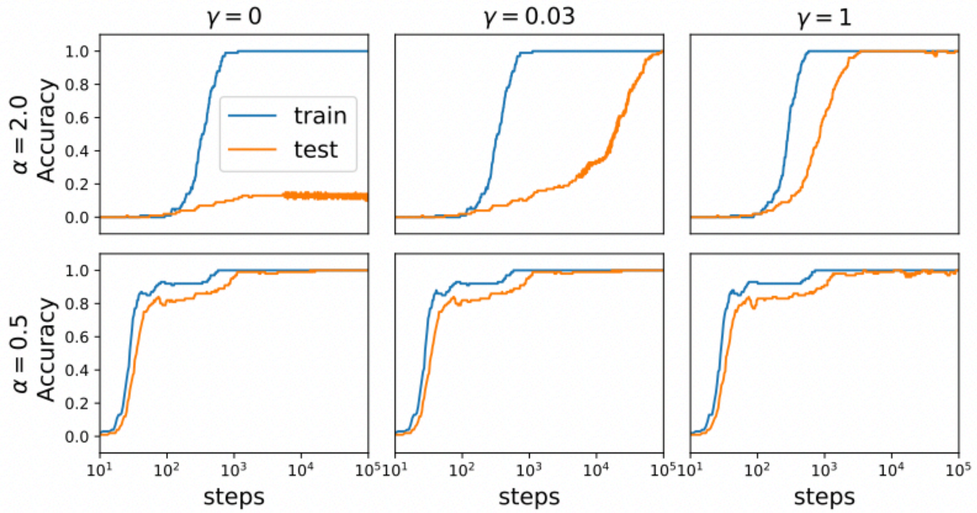


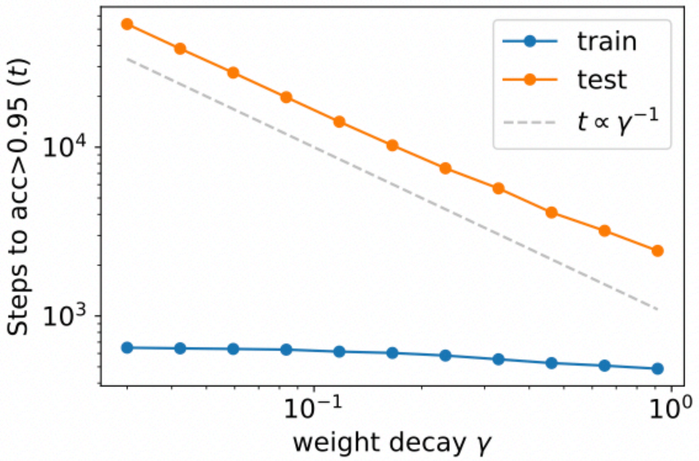
Usual situation, without regularization.

Loss on train look like «L», loss on test look like «U»

Grokking and weight norm



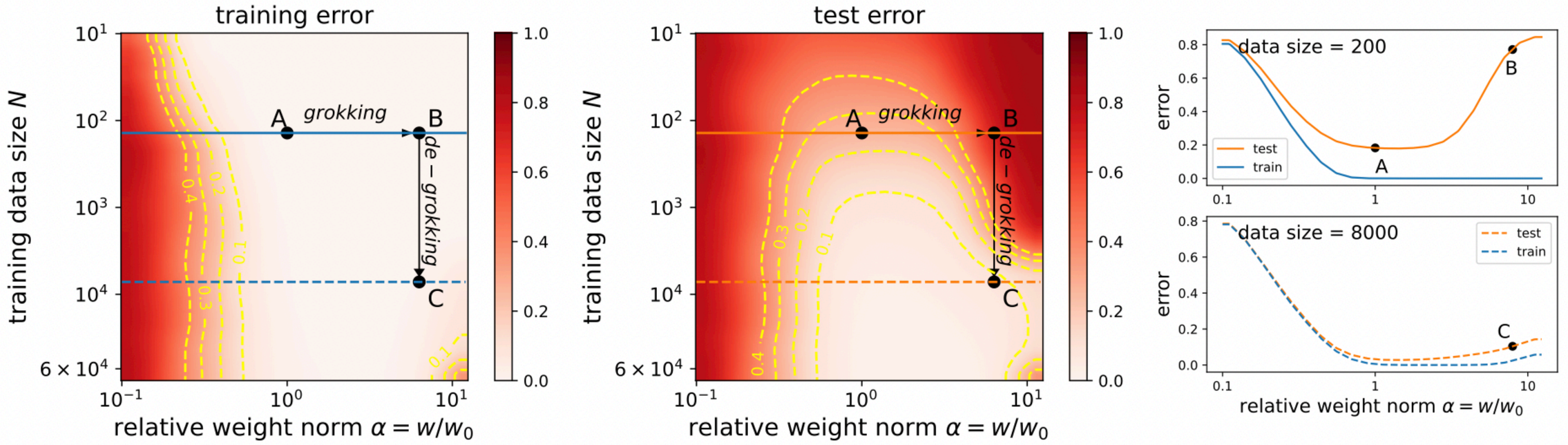




Teacher-student setup. The student network is trained with the Adam optimizer (learning rate 3×10^{-4}) for 10^5 steps.

Grokking and weight norm

induce grokking on MNIST

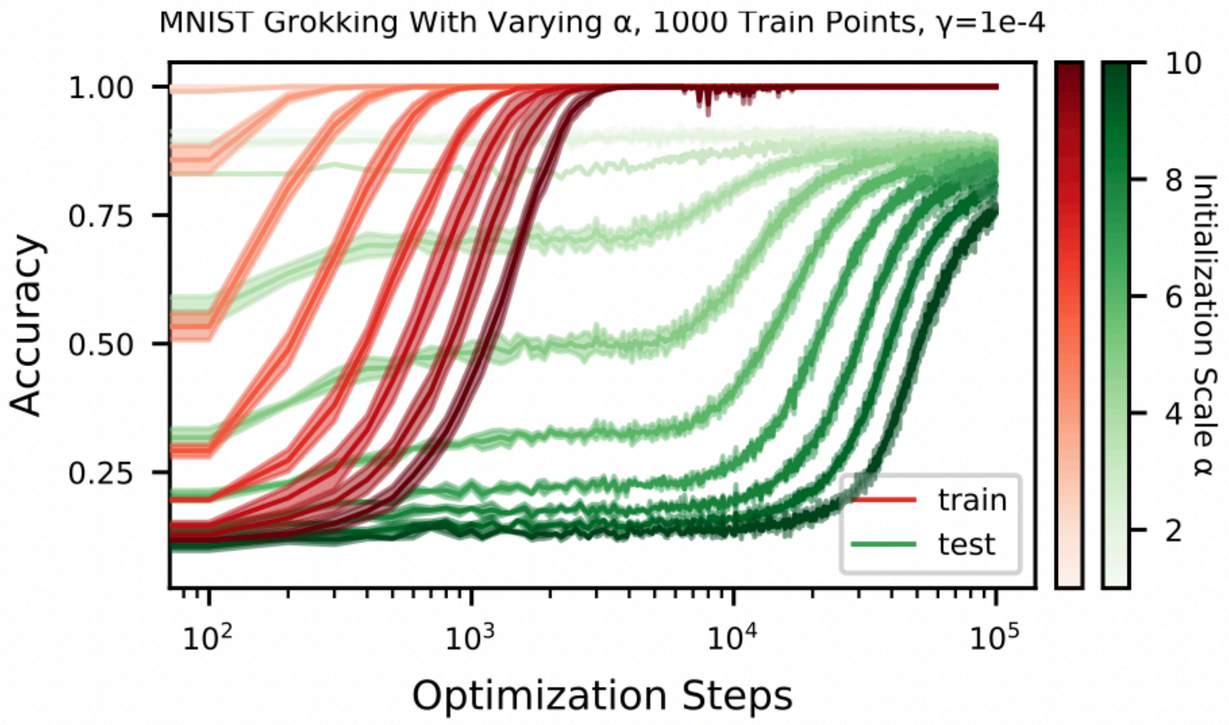


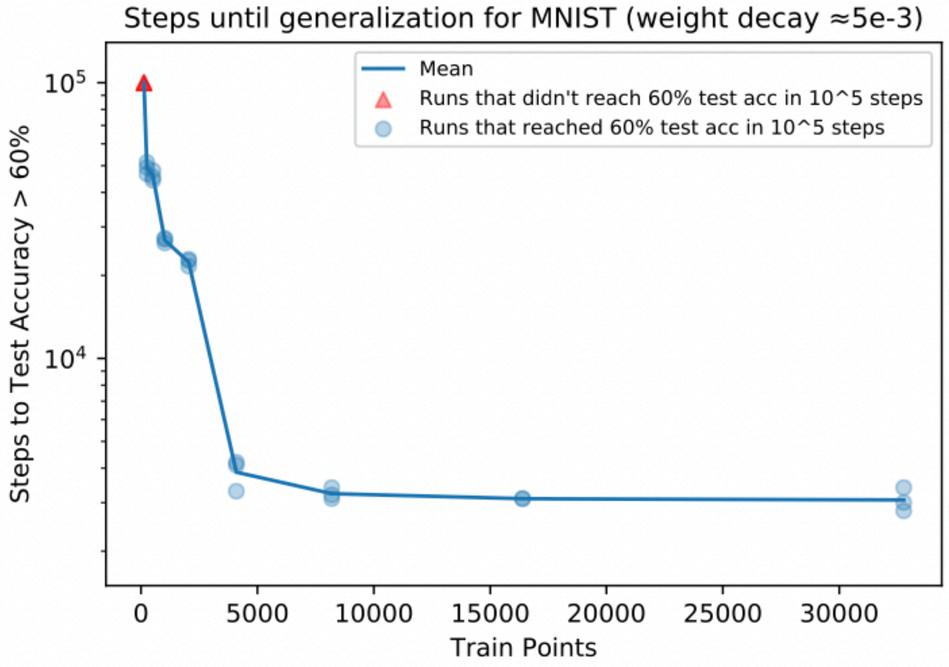
$\gamma = 0$

critical training set size below which generalization is impossible ≈ 200

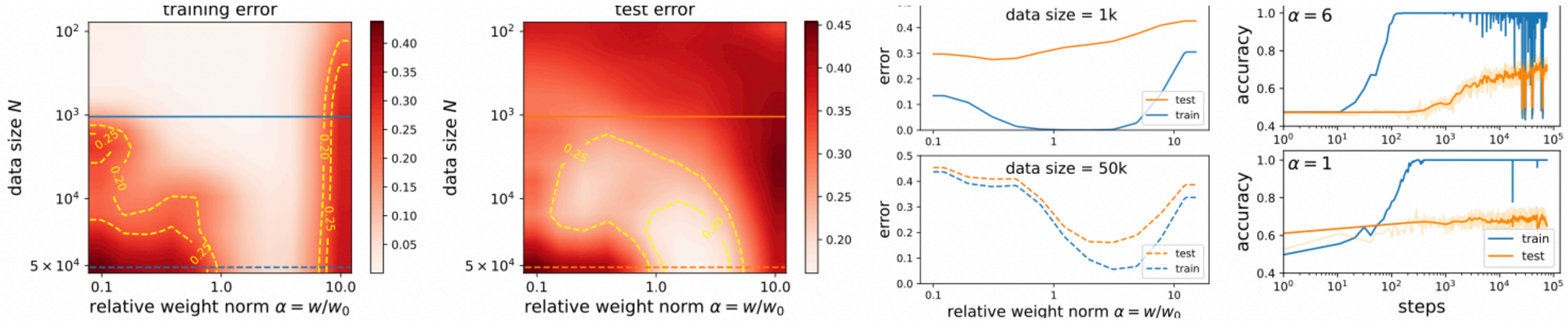
Grokking and weight norm

induce grokking on MNIST



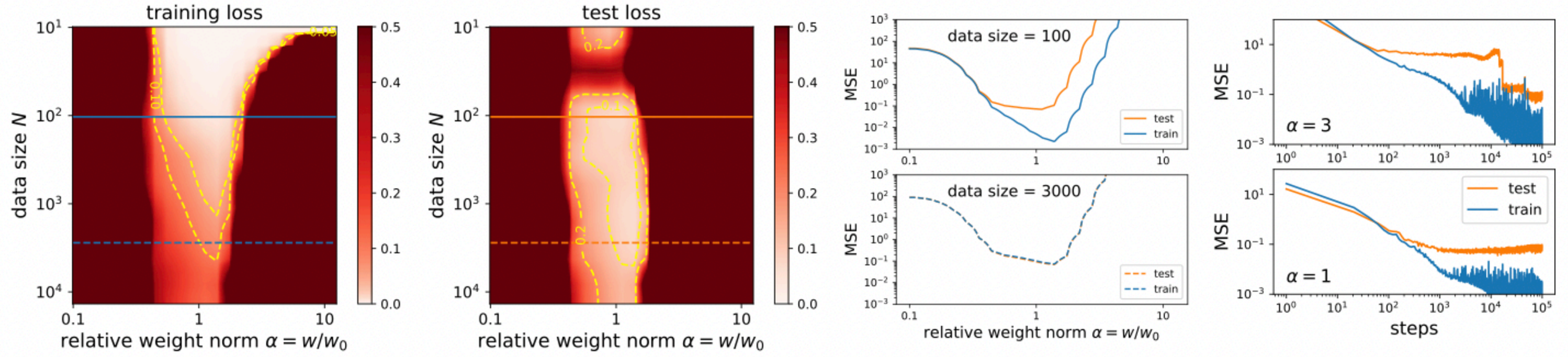


Grokking and weight norm



Sentiment analysis of text.

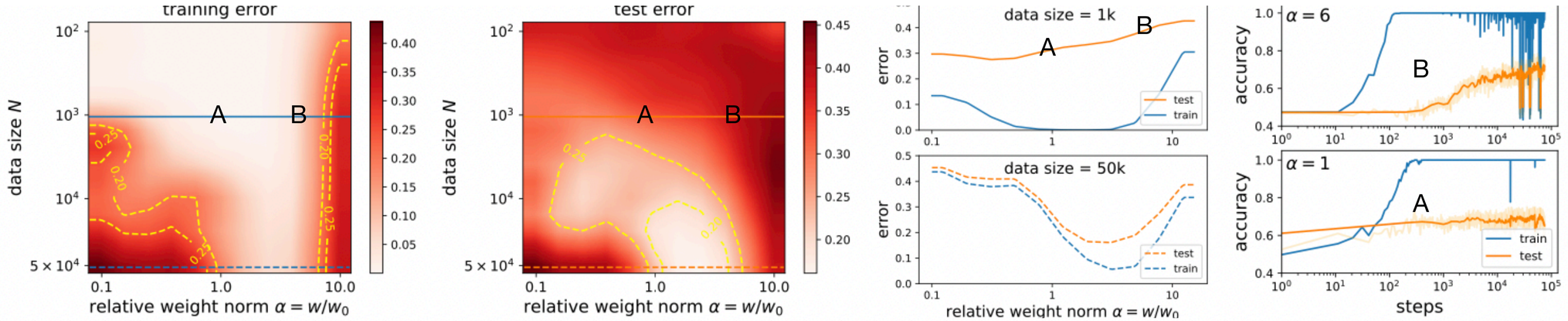
Grokking using LSTMs on IMDb dataset



Molecules.

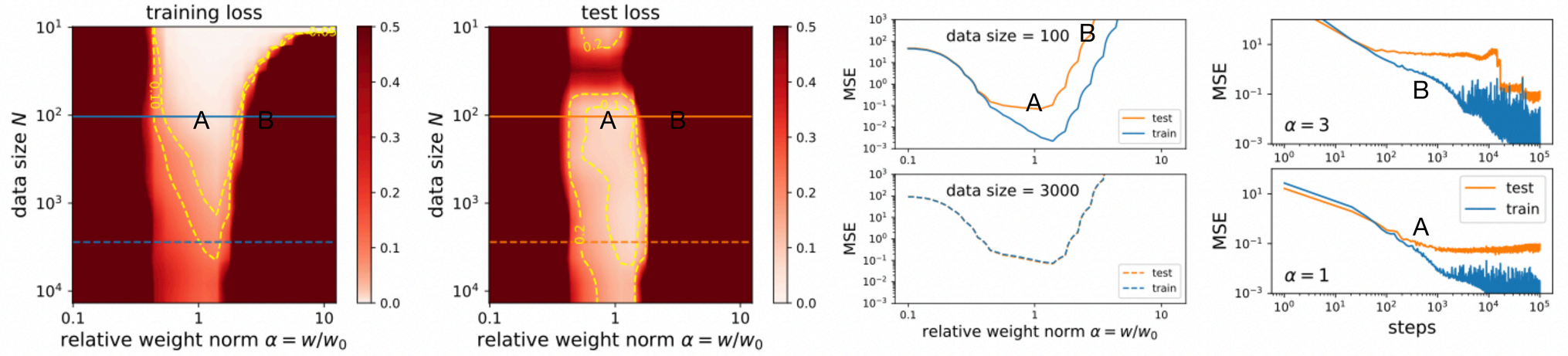
Grokking using the graph convolutional neural network (GCNN) for QM9 dataset.

Grokking and weight norm



Sentiment analysis of text.

Grokking using LSTMs on IMDb dataset

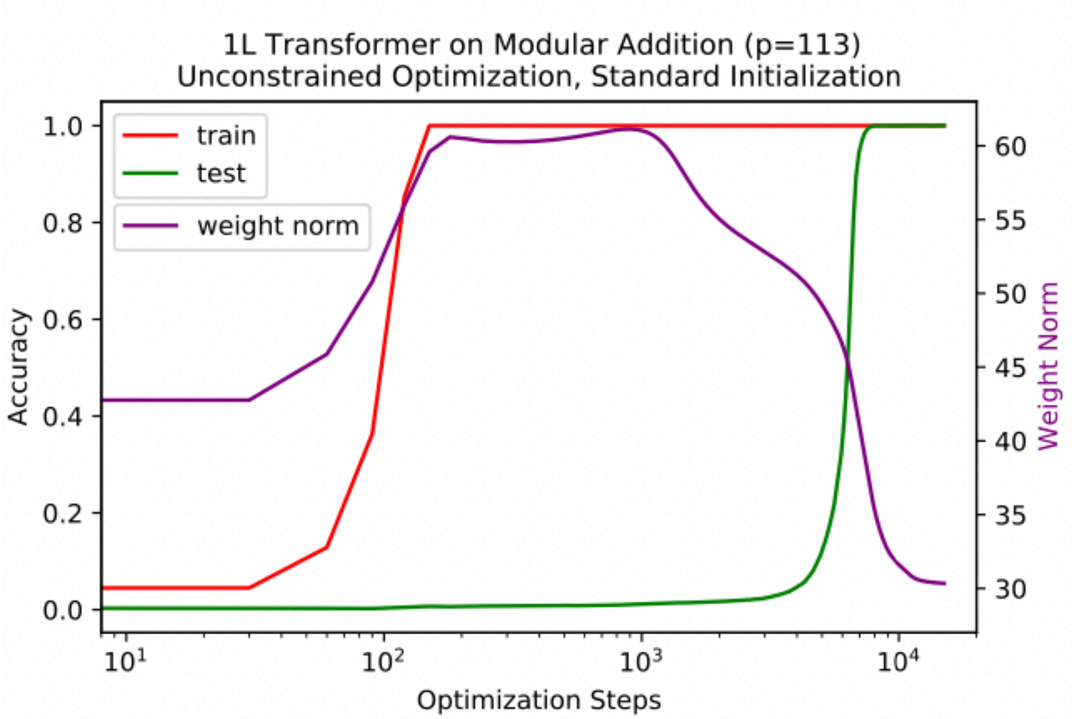


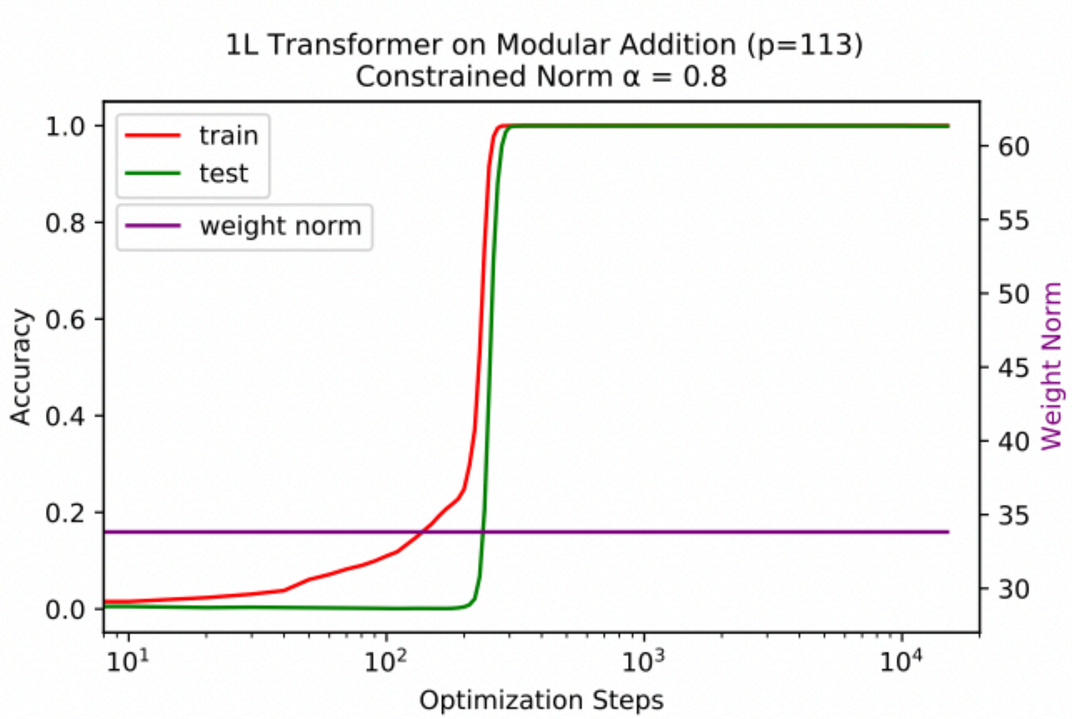
Molecules.

Grokking using the graph convolutional neural network (GCNN) for QM9 dataset.

Grokking and weight norm

Algorithmic datasets





Grokking and weight norm

Algorithmic datasets