Latent State Models of Training Dynamics

Michael Y. Hu[1] Angelica Chen[1] Naomi Saphra[1] Kyunghyun Cho[1,2,3]
michael.hu@nyu.edu
[1]New York University [2]Prescient Design, Genentech [3]CIFAR Fellow

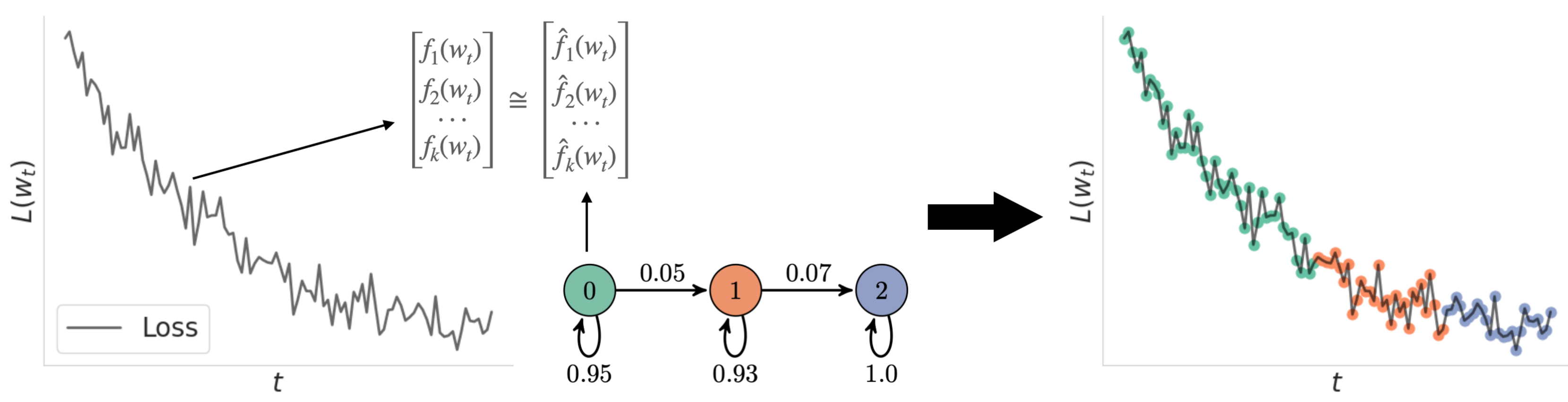
Motivation

Create a method to:

1. Understand random variation during model training.
2. Analyze phase transitions.

Approach

1. Compute summary statistics for model checkpoints.
2. Train a hidden Markov model (HMM) to predict trajectories of statistics. The HMM infers a latent state for each checkpoint.
3. Use the learned HMM to analyze training dynamics.



Finding Detour States

We train linear regression to predict convergence epoch from the empirical distribution over latent states. Let $X_1, ..., X_n$ be the sequence of latent states.

- x: $\hat{P}(X = i) = \frac{\text{number of times } X_j = i}{n}$
- y: The iteration in which evaluation accuracy crosses a threshold.

Dataset	R^2	p-value
Modular addition	0.977	<0.001
Sparse parities	0.961	<0.001
MNIST	0.154	0.315

A learned latent state is a **detour state** if:

- Some training runs do not visit the state.
- Its linear regression coefficient is positive when predicting convergence time.

Detour states are bolded.

Modular addition

State	Coefficient
0	-0.15
1	0.98
2	**1.19**
3	-0.20
4	0.18
5	**0.95**

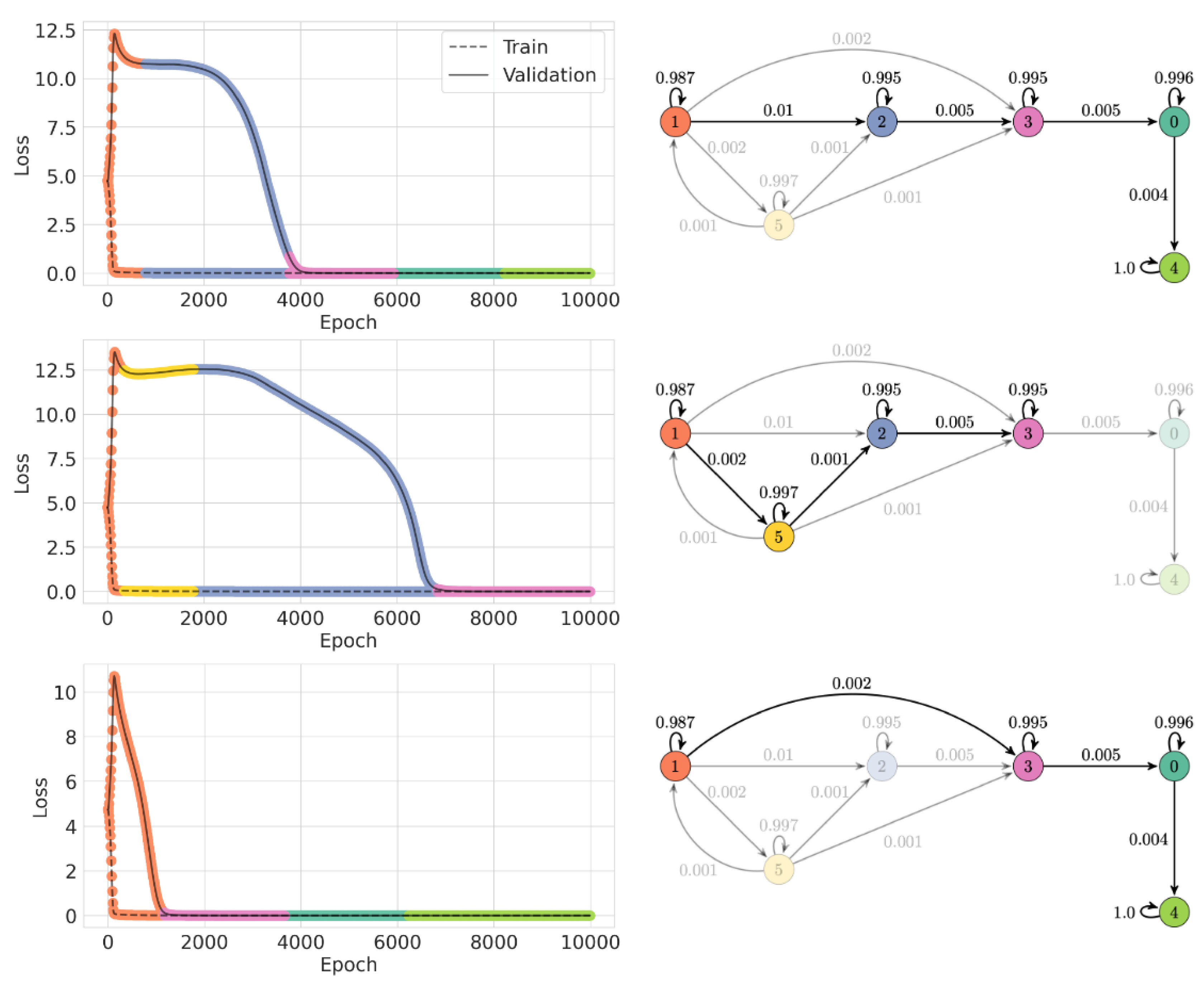
Sparse parities

State	Coefficient
0	**0.77**
1	0.41
2	0.98
3	-0.23
4	0.58
5	1.13

MNIST

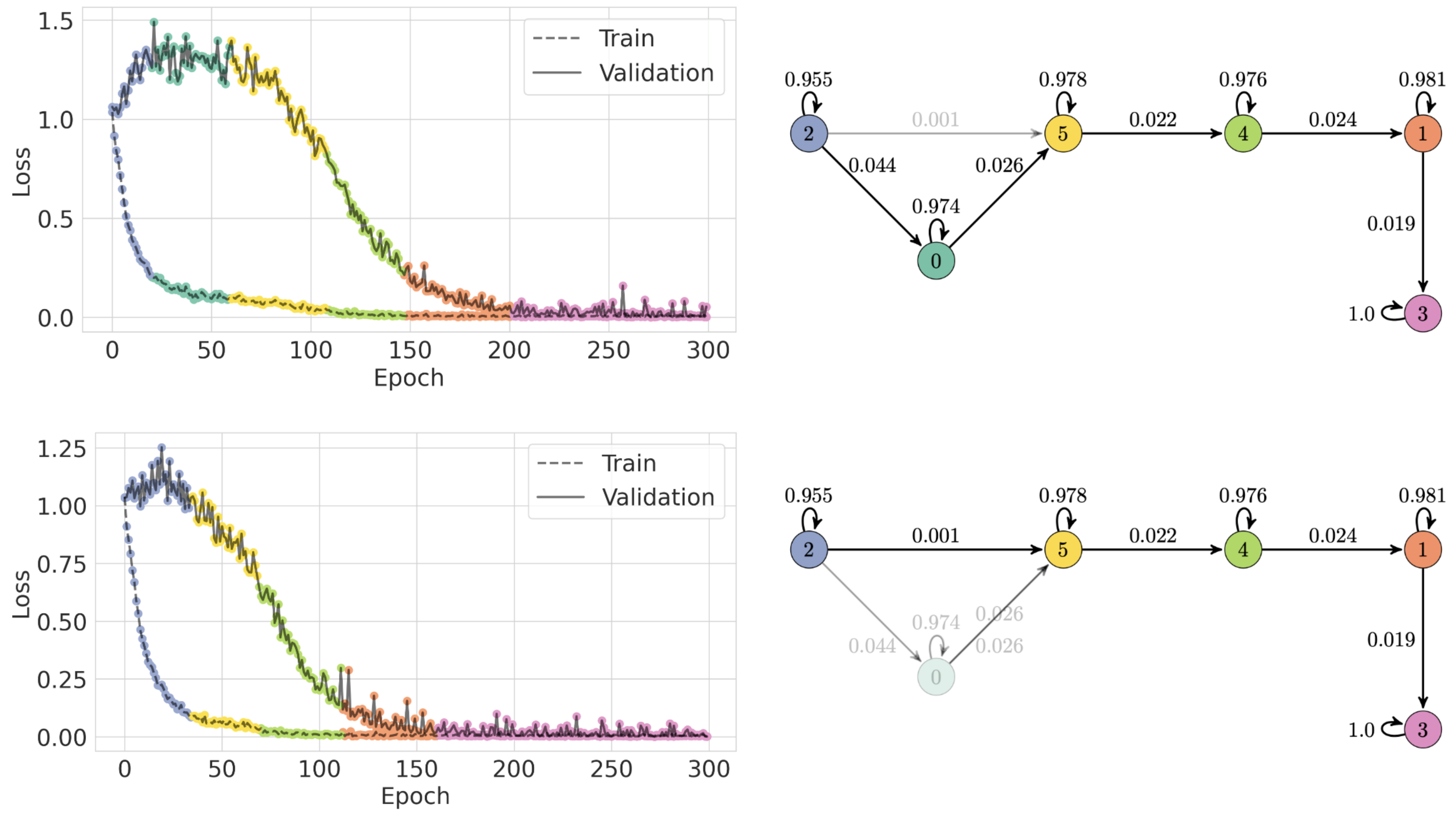
State	Coefficient
0	0.17
1	0.52
2	0.54
3	-0.06
4	-0.33
5	0.46

Grokking: Modular Addition



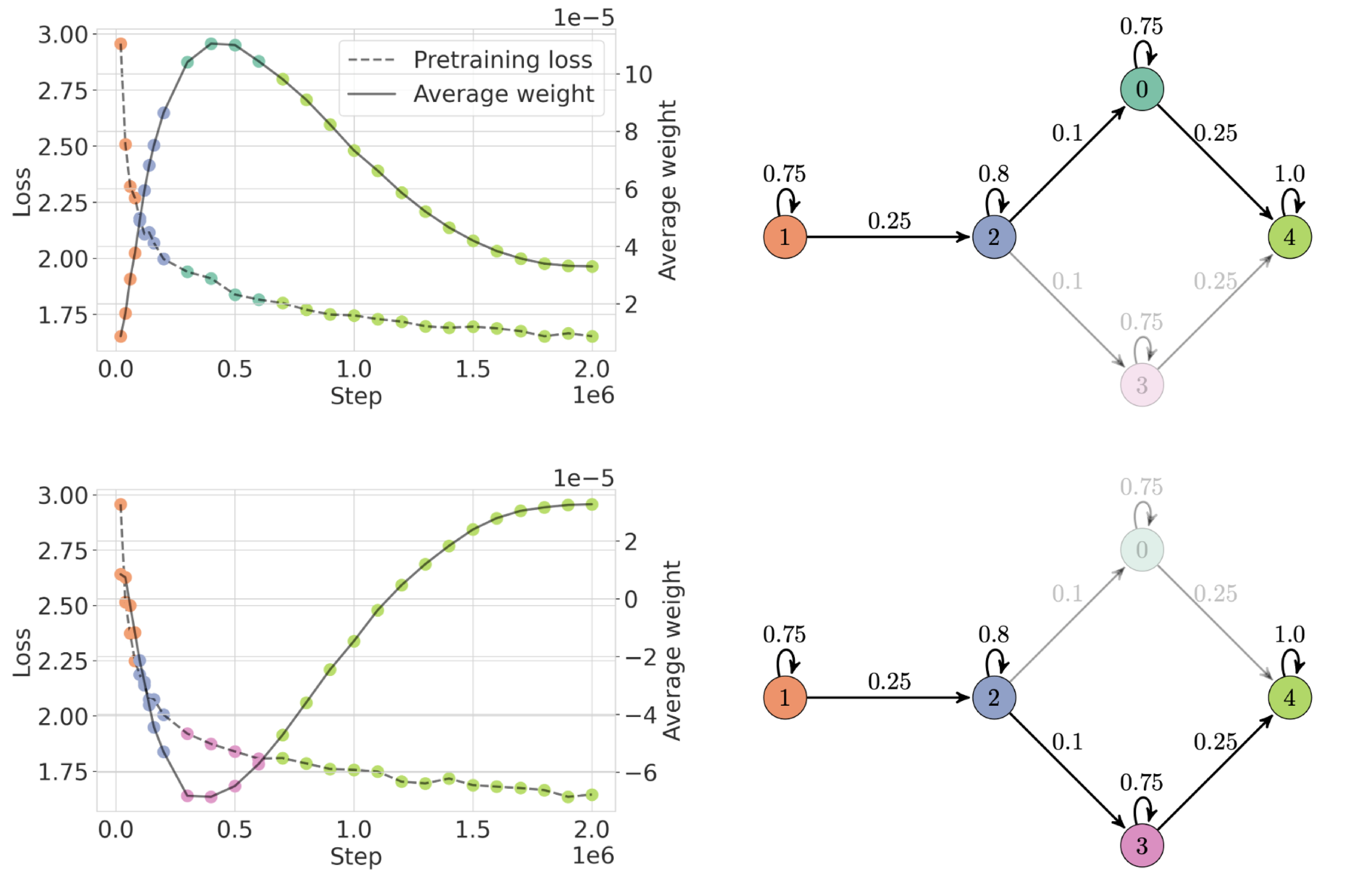
Edge	Top 3 important feature changes, by z-score	# of runs using edge (40 total)
$1 \to 2$	$L_2 \downarrow 0.59$, $L_1 \downarrow 0.88$, $\frac{L_1}{L_2} \downarrow 1.05$	34
$1 \to 5$	$L_2 \downarrow 2.08$, $Var(w) \downarrow 2.24$, $L_1 \downarrow 2.25$	4
$1 \to 3$	$L_2 \downarrow 1.68$, $Var(w) \downarrow 1.99$, $L_1 \downarrow 1.83$	2

Grokking: Sparse Parities



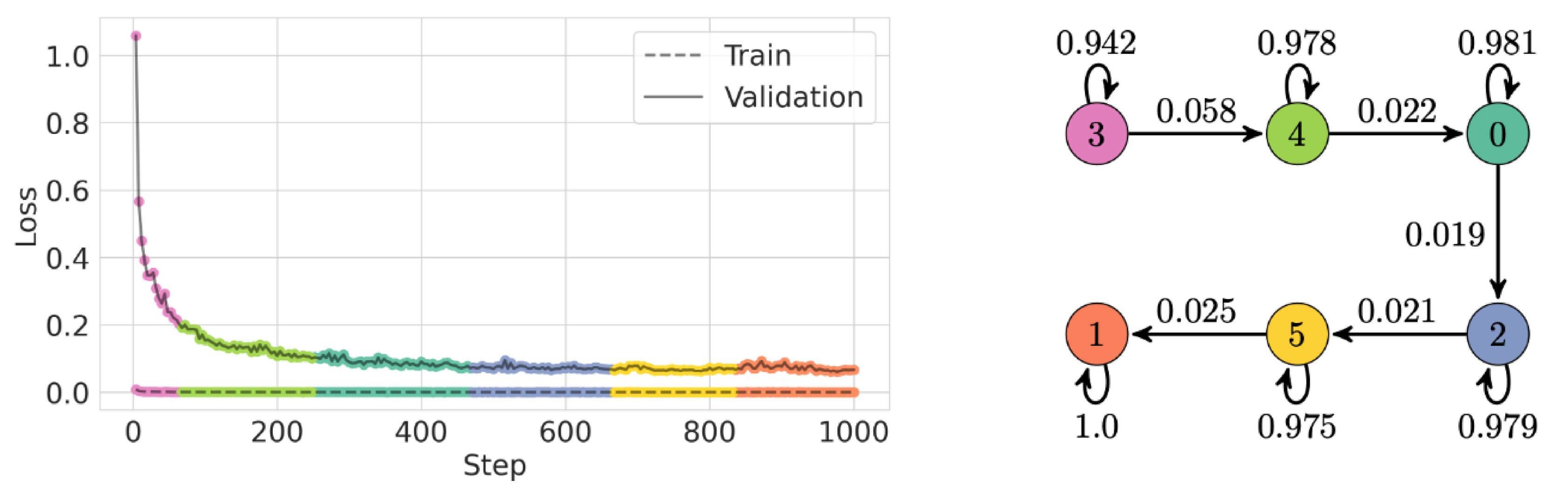
Edge	Top 3 important feature changes, by z-score	# of runs using edge (40 total)
$2 \to 0$	$L_2 \uparrow 0.11$, $L_1 \downarrow 0.61$, $\frac{L_1}{L_2} \downarrow 0.32$	39
$2 \to 5$	$L_2 \downarrow 0.19$, $L_1 \downarrow 1.01$, $\frac{L_1}{L_2} \downarrow 0.54$	1

Masked Language Modeling: MultiBERTs



Edge	Top 3 important feature changes, by z-score	# of runs using edge (5 total)
$2 \to 0$	median(w) $\uparrow 1.69$, mean(w) $\uparrow 1.70$, $\max(\lambda)$ $\uparrow 1.14$	2
$2 \to 3$	median(w) $\downarrow 1.33$, mean(w) $\downarrow 1.30$, $\max(\lambda)$ $\uparrow 1.11$	3

Image Classification: MNIST



Edge	Top 3 important feature changes, by z-score
$3 \to 4$	$L_2 \uparrow 0.62$, $Var(w) \uparrow 0.58$, $L_1 \uparrow 0.61$
$0 \to 2$	$L_2 \uparrow 0.69$, $Var(w) \uparrow 0.70$, $L_1 \uparrow 0.70$
$5 \to 1$	$L_2 \uparrow 0.46$, $Var(w) \uparrow 0.50$, $L_1 \uparrow 0.48$

Contributions

1. The HMM is a principled, automated, and widely applicable method for analyzing variability in model training and phase transitions.
2. Certain latent states are predictive of a training run converging more slowly.
3. Generalization in grokking can be anticipated via changes in the model occurring earlier in training.