Appendix

Additional tables & figures

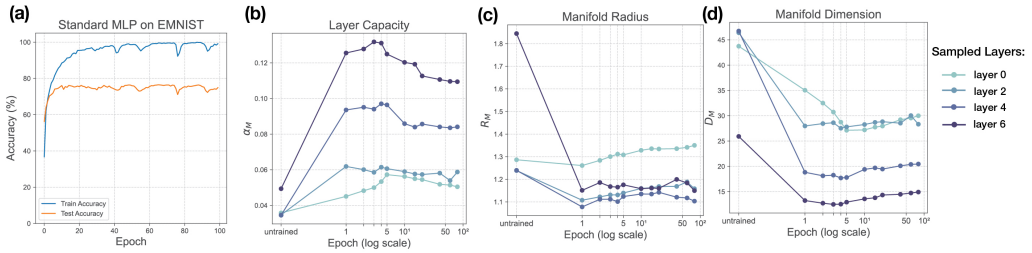

Figure 5: **Manifold capacity measurements on a standard initialized MLP network (no weight scaling, no output scaling, and trained with sufficiently large dataset).** Inspecting across layers, the later layers also show higher object manifold capacity as expected, since the later layers should learn high-level features that more directly contribute to classification [8].

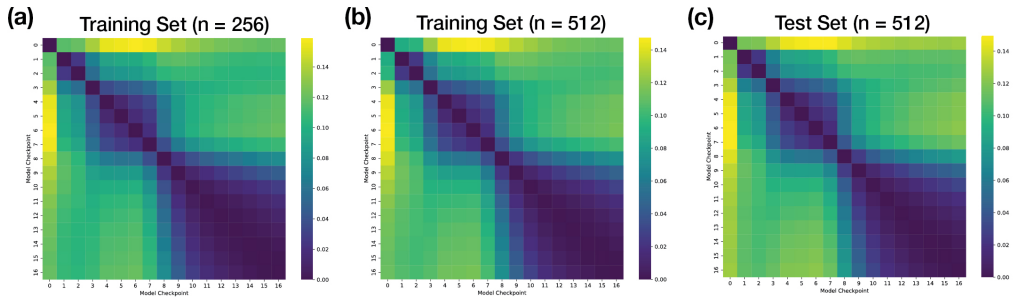

Figure 6: **Example pairwise kernel distance heatmap calculated at different samples.** The MNIST network with output scaling $\alpha = 0.5$ is shown here. **(a)** Using 256 training samples. **(b)** Using 512 training samples. **(c)** Using 256 test samples.

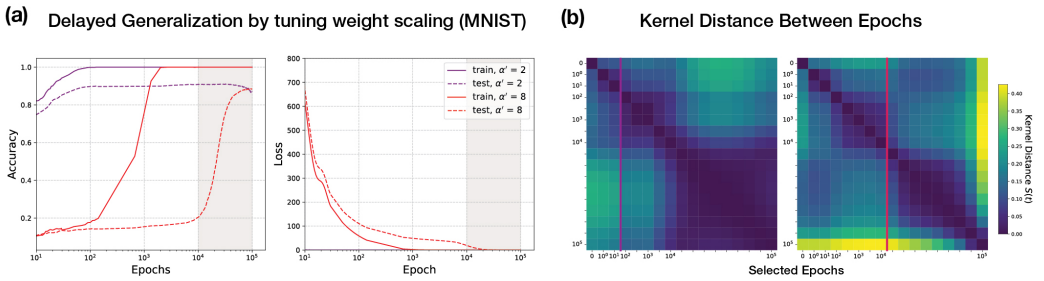

Figure 7: **Delayed generalization in MNIST by tuning weight scaling parameter α'. (a)** Performance of example grokked and no-grok networks. **(b)** Kernel distance heatmaps. Left is for $\alpha' = 2$, right is $\alpha' = 8$.

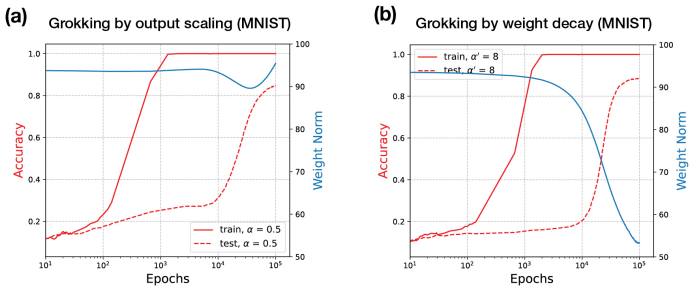

Figure 8: **Weight norm changes during grokking**. Weight norm decrease precedes improved test performance, but does not reflect all changes in the kernel (see Fig. 1), 7).

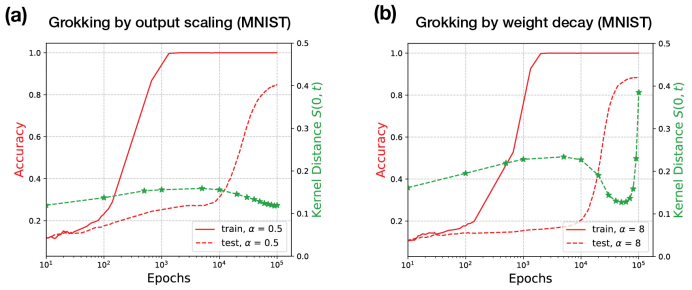

Figure 9: **Kernel Distance to network at initialization over the course of training.**

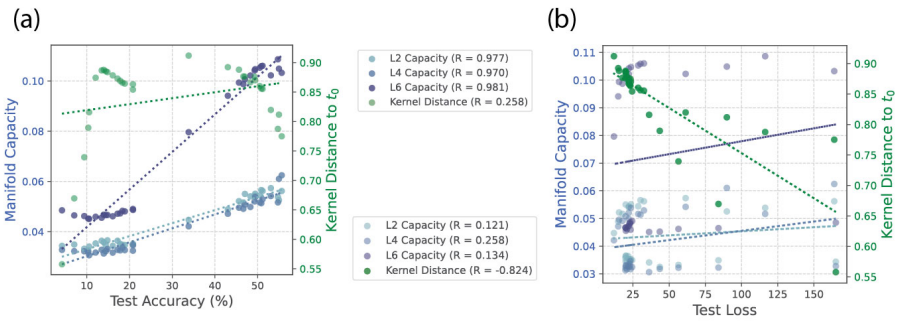

Figure 10: **Correlation between test performances and Manifold capacity measure and kernel distance to initialization. (a)** Correlation to test accuracy. **(b)** Correlation to test loss. We sampled manifold capacity of Layer 2, 4 and 6 for the visualization.

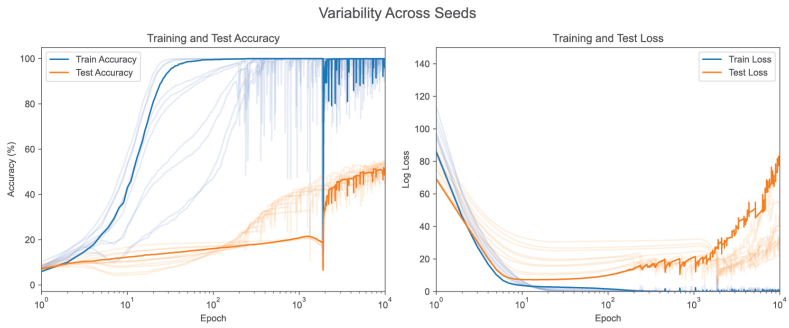

Figure 11: **Variability in induced grokking in EMNIST.** In this network, $\alpha = .5$, training samples $n = 2000$, with no weight decay.

	sample size	weight scaling	output scaling	weight decay
varying initialized weight norm	2000	10	1	0.001
	2000	5	1	0.001
	2000	1	1	0.001
varying sample size	1000	10	1	0.001
	2000	10	1	0.001
	5000	10	1	0.001
varying output scale	2000	10	0.5	0
	2000	10	0.1	0
	2000	10	0.001	0

Table 1: **Three setups to induce/eliminate grokking.**

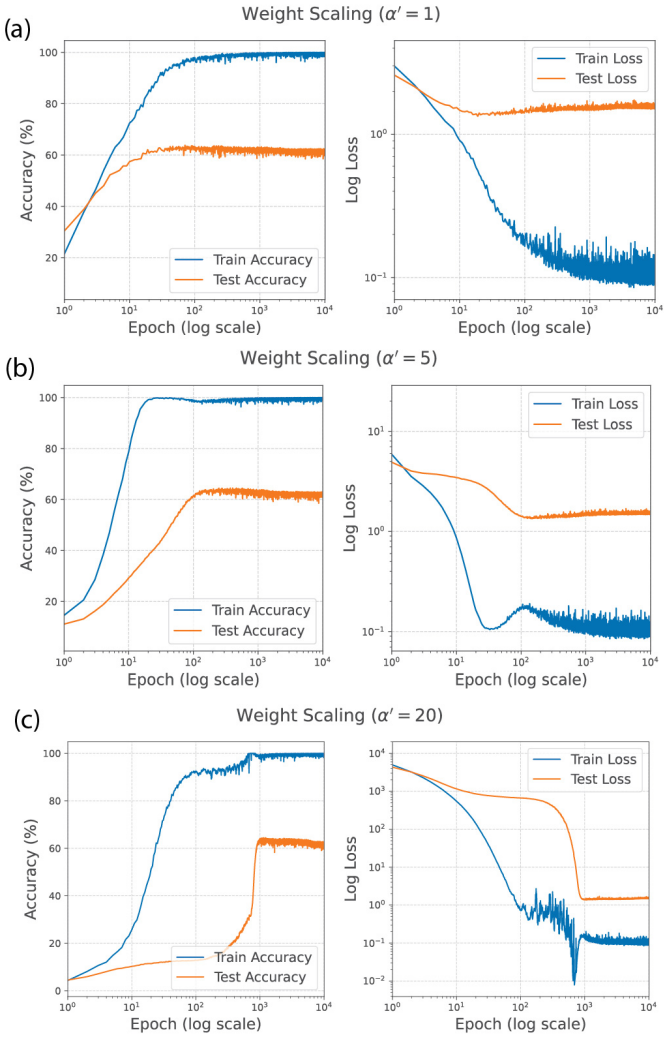

Figure 12: **Grokking induced with Adam optimizer with weight decay by tuning weight scaling parameter α'.** We use the same model architecture as described in Section 3; and other setting the same with weight decay using AdamW optimizer (Table 1).

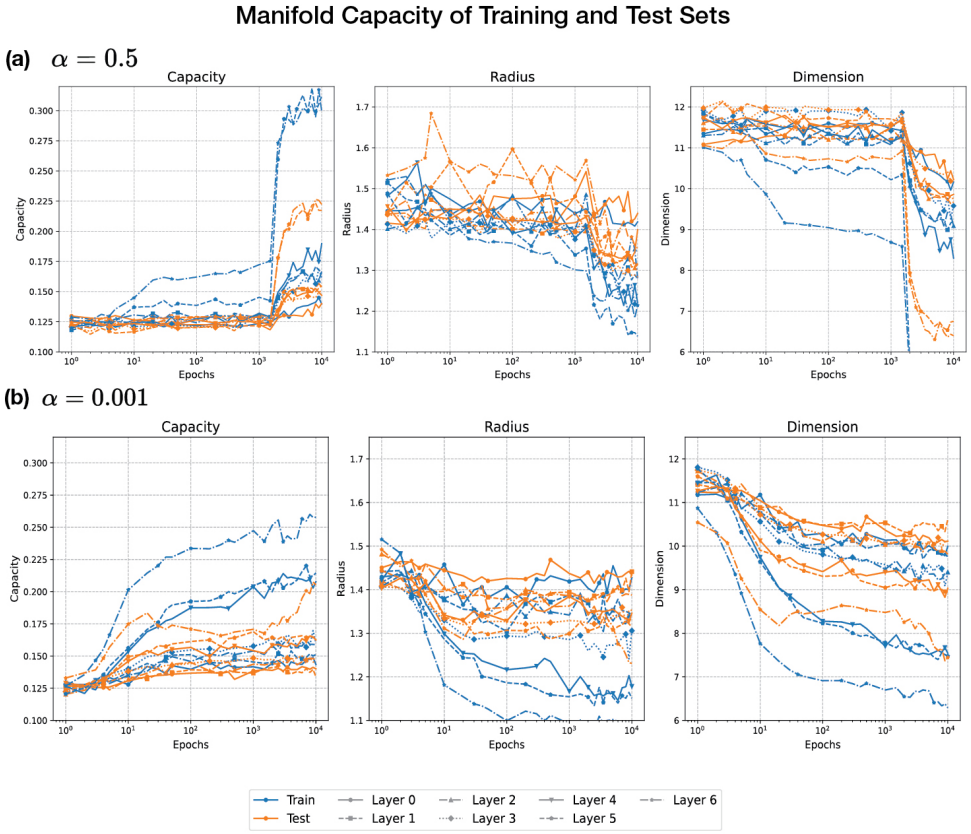

Figure 13: **Manifold capacity and geometry measures for both training and test set for the two networks in.** About 20 samples per class was used from training/ test dataset. The overall trends of changes are consistent for training and test samples.

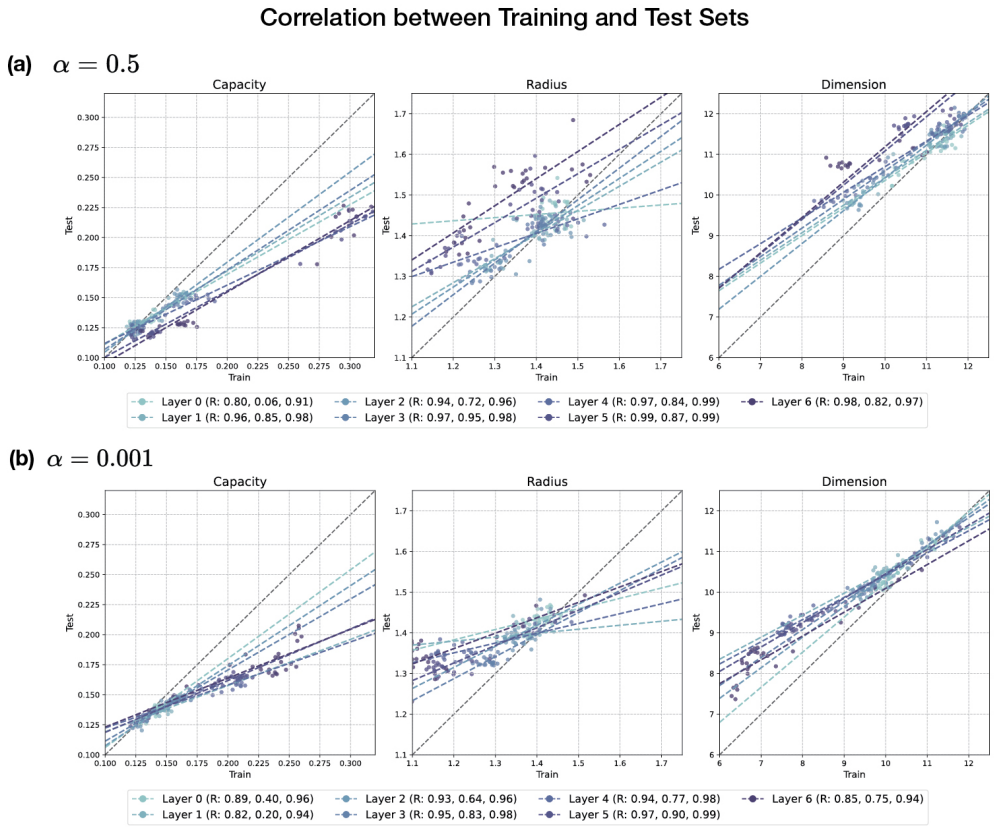

Figure 14: **Correlation of α_C, R_M, D_M between train samples and test samples.** Same two networks is shown as Fig. 13. The R-value (correlation coefficient) is shown in the legend, and overall correlation is high for all measures, and there is no big difference between the two different network setups.

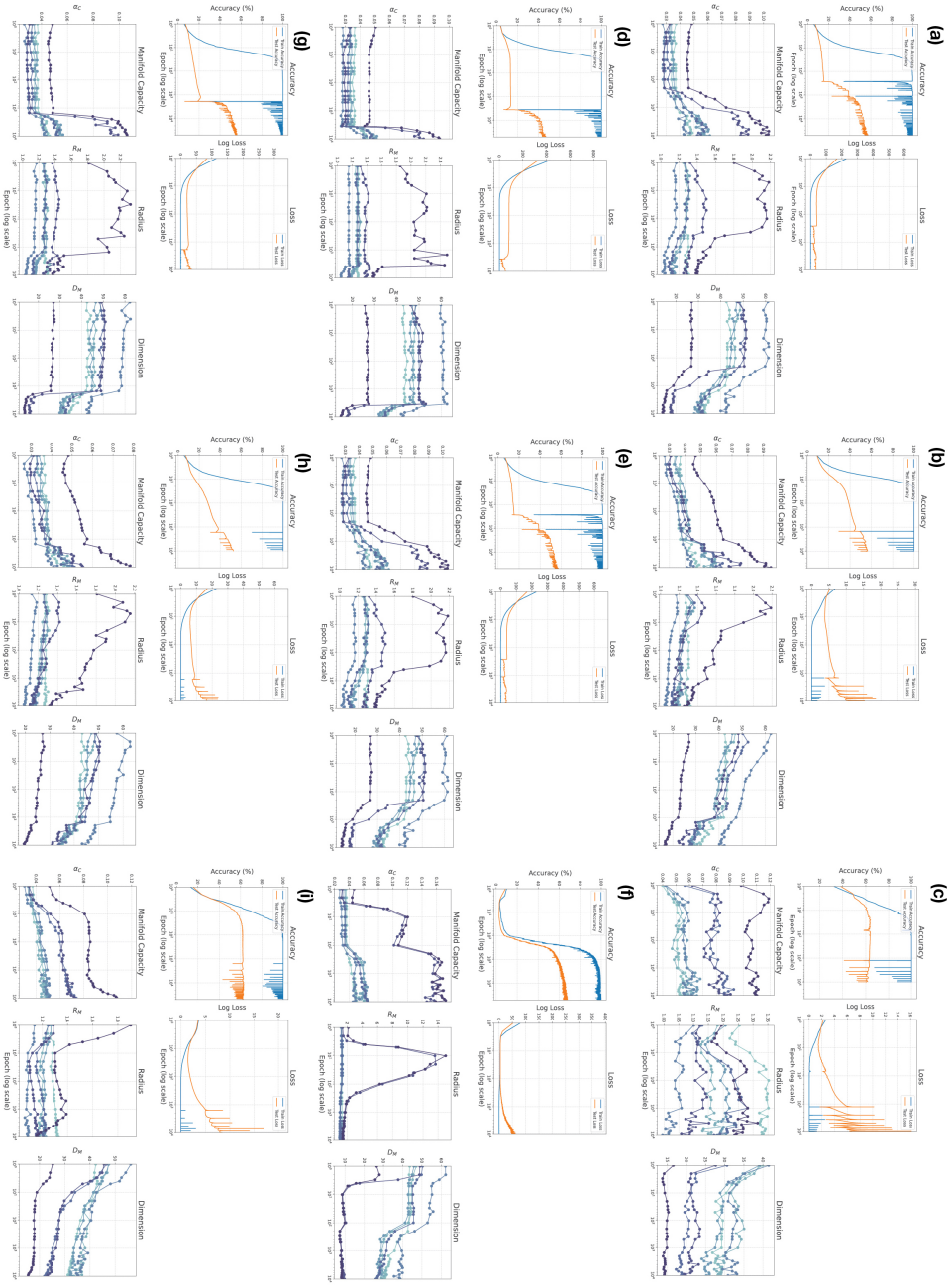

Figure 15: **Performance metrics (upper row) and manifold capacity and geometry measures (lower row), ordered the same as in Table 1. (a-c)**. Changed weight norm. **(d-f)**. Changed tranining set size. **(g-i)**. changed output scaling.