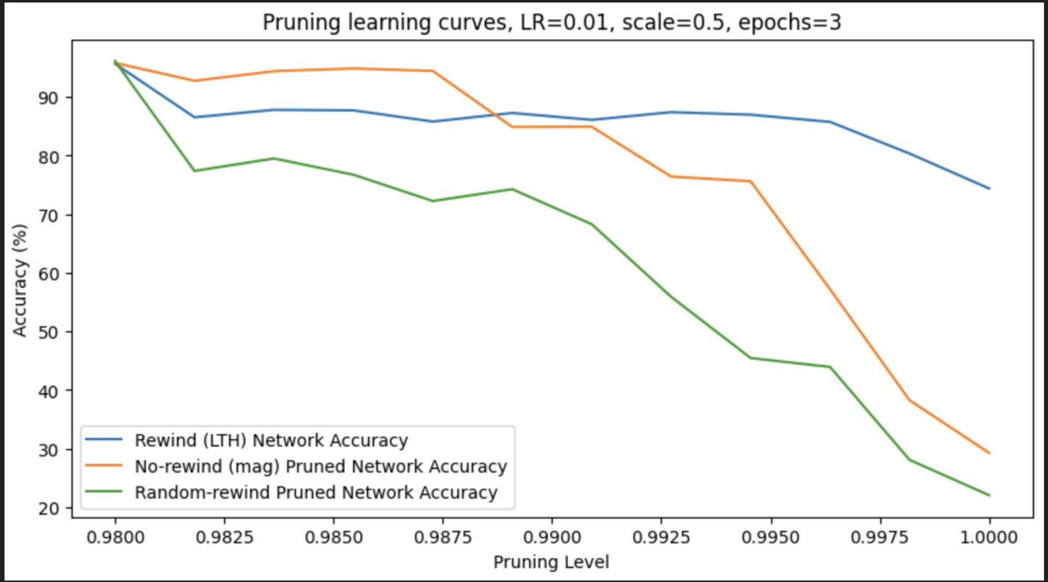
Feature-Aware Pruning in MLPs

Tanishq Kumar

Levels of pruning

1) Before training, at init (SNIP, GraSP, SynFlow) [1, 2, 3]
2) **During training (lottery tickets)** [4]
3) After training, before inference (magnitude-based pruning) [5]

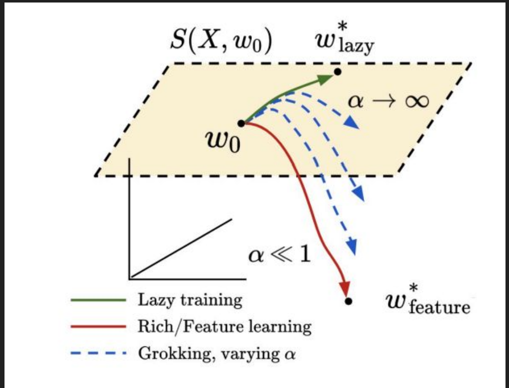
Lottery Ticket Hypothesis



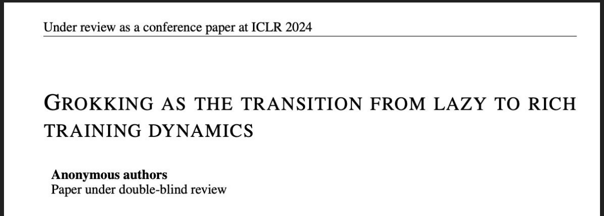
30 seconds of ML theory: lazy and rich training regimes

- Networks are powerful *nonlinear* models.

- ML theorists have discovered ways to modify any neural network to *continuously linearize it* [6, 7].
  - One such way, by tuning a parameter “alpha,” is given on the right
  - Highly nonlinear (small alpha) = “**rich training**,” vs highly linearized (large alpha) = “**lazy training**”
  - **Amount of feature learning = deviation from linearized model (tuned by alpha)**



(c) α Alters Dynamics



Under review as a conference paper at ICLR 2024

GROKKING AS THE TRANSITION FROM LAZY TO RICH TRAINING DYNAMICS

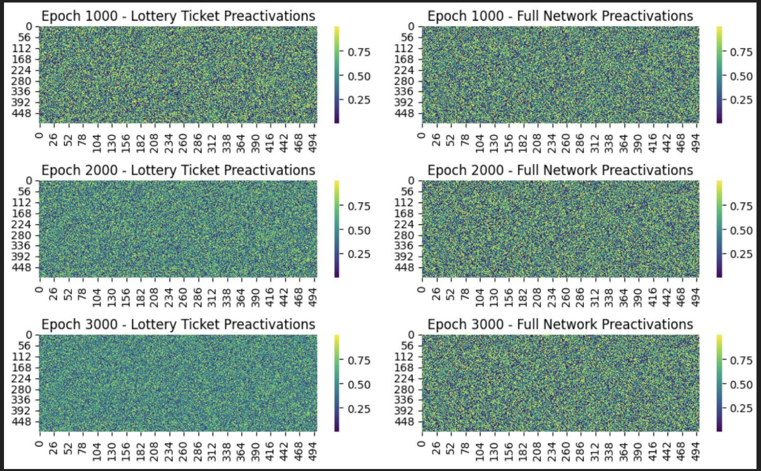
Anonymous authors
Paper under double-blind review

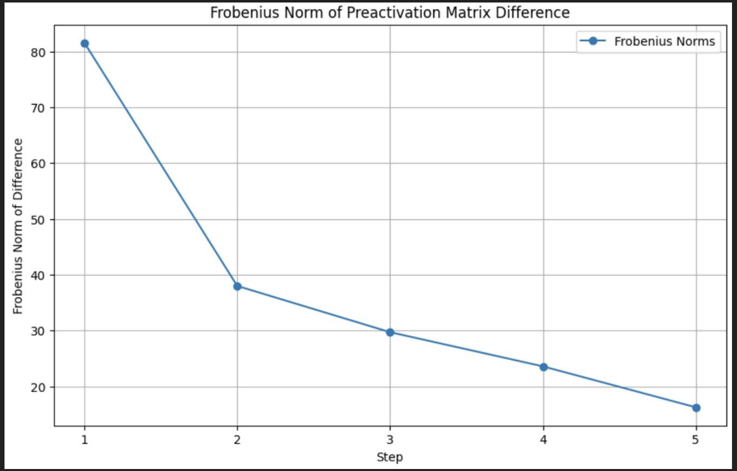
Feature-Learning Networks Are Consistent Across Widths At Realistic Scales

Nikhil Vyas[1*] **Alexander Atanasov**[2,3,4*] **Blake Bordelon**[1,3,4*]
Depen Morwani[1,3] **Sabarish Sainathan**[1,3,4] **Cengiz Pehlevan**[1,3,4]
[1]SEAS [2]Department of Physics [3]Kempner Institute [4]Center for Brain Science
Harvard University
{nikhil,atanasov,blake_bordelon,dmorwani,
sabarish_sainathan,cpehlevan}@g.harvard.edu

Toy model

- One hidden-layer MLP student-teacher task [8, 9]
- Find lottery ticket on this model, then train lottery ticket and full network
- Compare preactivation matrices for lottery ticket and full net during training
  - Visual way to compare “learned features”





Take-away from toy model

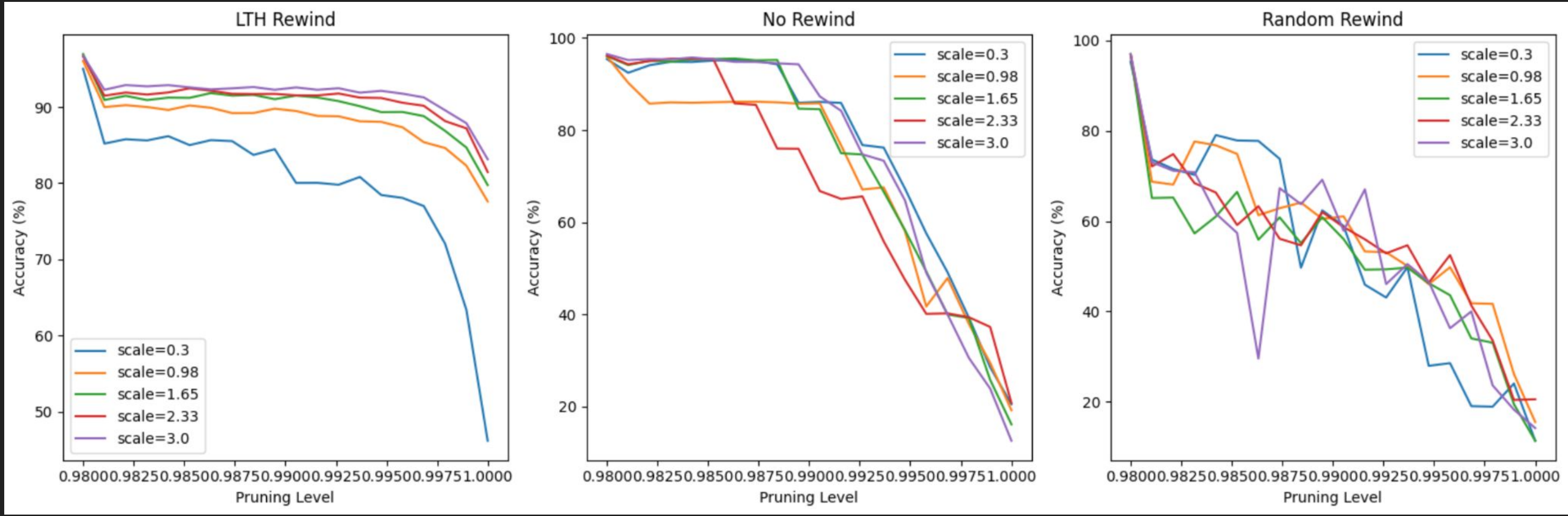
LTH:

1. Pruned networks reach same end-time test error as full network
2. (Stronger) Pruned networks reach same end-time *features* as full network
3. (Strongest) Pruned networks have same *feature dynamics* as full network

Conjecture: LTH paper shows (1). Toy model suggests (3). This property *does not hold for pruning with random rewinding or no rewinding*!

Hypothesis: sweeping over rate of feature learning should change performance of lottery tickets *uniformly*, but performance of random/no rewind in a complicated, *messy* way (features are not necessarily the same over sweep).

Testing conjecture **on MNIST**



Punchline: *we can beat the state of the art in pruning* by using tricks from theory.

Left (orange) is MLP on MNIST from LTH paper. Purple is **(ours).**

References

Before Training, at Initialization:

- [1] Lee, Namhoon, et al. "SNIP: Single-shot network pruning based on connection sensitivity." *International Conference on Learning Representations*. 2018. Link
- [2] Wang, Chaoqi, et al. "Picking Winning Tickets Before Training by Preserving Gradient Flow." *International Conference on Learning Representations*. 2020. Link
- [3] Tanaka, Hidenori, et al. "Pruning neural networks without any data by iteratively conserving synaptic flow." *NeurIPS*. 2020. Link

During Training:

- [4] Frankle, Jonathan, and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." *International Conference on Learning Representations*. 2019. Link

After Training, Before Inference:

- [5] Han, Song, et al. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding." *International Conference on Learning Representations*. 2016. Link

Modifying Neural Networks to Continuously Linearize:

- [6] Jacot, Arthur, et al. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." *NeurIPS*. 2018. Link
- [7] Lee, Jaehoon, et al. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent." *NeurIPS*. 2019. Link

Theory for One Hidden-layer MLP Student-Teacher Task:

- [8] Saxe, Andrew M., et al. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks." *International Conference on Learning Representations*. 2014. Link
- [9] Mei, Song, et al. "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit." *Berkeley*. 2019. Link

Appendix: additional experiments (explained in report)