Understanding Surprising Generalization Phenomena in Deep Learning

Wei Hu

University of Michigan
vvh@umich.edu

Deep learning has exhibited a number of surprising generalization phenomena that are not captured by classical statistical learning theory. In my new faculty highlight talk, I will survey some of my work on the theoretical characterizations of several such intriguing phenomena:

- Implicit regularization: A major mystery in deep learning is that deep neural networks can often generalize well despite their excessive expressive capacity. Towards explaining this mystery, it has been suggested that commonly used gradient-based optimization algorithms enforce certain implicit regularization which effectively constrains the model capacity.
- Benign overfitting: In certain scenarios, a model can perfectly fit noisily labeled training data, but still archives near-optimal test error at the same time, which is very different from the classical notion of overfitting.
- Grokking: In certain scenarios, a model initially achieves perfect training accuracy but no generalization (i.e. no better than a random predictor), and upon further training, transitions to almost perfect generalization.

Theoretically establishing these properties often involves making appropriate high-dimensional assumptions on the problem as well as a careful analysis of the training dynamics.

Papers that will be surveyed in this talk are Arora et al. (2019); Frei et al. (2023); Hu et al. (2020); Lyu et al. (2024); Xu et al. (2024).

References

Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32.

Frei, S.; Vardi, G.; Bartlett, P.; Srebro, N.; and Hu, W. 2023. Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data. In *The Eleventh International Conference on Learning Representations*.

Hu, W.; Xiao, L.; Adlam, B.; and Pennington, J. 2020. The surprising simplicity of the early-time learning dynamics of neural networks. *Advances in Neural Information Processing Systems*, 33: 17116–17128.

Lyu, K.; Jin, J.; Li, Z.; Du, S. S.; Lee, J. D.; and Hu, W. 2024. Dichotomy of Early and Late Phase Implicit Biases Can Provably Induce Grokking. In *The Twelfth International Conference on Learning Representations*.

Xu, Z.; Wang, Y.; Frei, S.; Vardi, G.; and Hu, W. 2024. Benign overfitting and grokking in relu networks for xor cluster data. In *The Twelfth International Conference on Learning Representations*.