The Role of Over-Parameterization in Machine Learning - the Good, the Bad, the Ugly

Fanghui Liu

Department of Computer Science, University of Warwick, Coventry, UK
fanghui.liu@warwick.ac.uk

Statement - New Faculty Highlights

The conventional wisdom of simple models in machine learning misses the bigger picture, especially over-parameterized neural networks (NNs), where the number of parameters are much larger than the number of training data. The target of my research is to explore the mystery behind over-parameterized NNs from a theoretical side by addressing the following three questions:

- **Q1:** Over-parameterization helps or hurts robustness in NNs?
- **Q2:** Why over-parameterized NNs generalize well under SGD training?
- **Q3:** How does deep RL work well for function approximation beyond "linear" regime?

My research is able to address the above three questions as below.

- **A1:** A large numbers of literature in this community have a contradicting conclusion on the fundamental question: over-parameterization helps or hurts robustness? Our work (Zhu et al. 2022) aims to investigate this apparent contradiction in theory, and to close the gap as much as possible. We demonstrate the *good* in width, the *bad* depth, the *ugly* in initialization regarding the average robustness of DNNs: in the over-parameterized regime, width helps robustness (*good*); depth (*bad*) helps robustness under LeCun initialization but hurts the robustness He-initialization and NTK initialization (*ugly*).
- **A2:** Regarding generalization of over-parameterized models, our work (Liu, Suykens, and Cevher 2022) aims to understand over-parameterized two-layer neural networks trained by stochastic gradient descent (SGD), which coincides with practical neural networks training, and accordingly bridges the theoretical gap of previous work depending on the closed-form solution. Our results are able to characterize the double descent behavior by the unimodality of variance and monotonic decrease of bias, which provides a theoretical justification to understand why over-parameterized model can generalize well under SGD training. Furthermore, our finding shows that the constant step-size SGD setting incurs no loss in convergence rate when compared to the exact minimum-norm interpolator, as a theoretical justification of using SGD in practice. Besides, extension our result from two-layer NNs to shallow Transformers can be found in (Wu et al. 2023).
- **A3:** Based on our generalization results, we are able to analyse the function approximation in deep RL beyond "linear" regime, e.g., NTK, Eluder dimension. This scheme is powerful in practice, e.g., deep Q-network (DQN) using DNNs for function approximation. Our work (Liu, Viano, and Cevher 2022) transforms the estimation of temporal difference error to a *generalization guarantees* problem under the non-iid setting. We demonstrate that the sublinear regret can be achieved for deep neural function approximation with reasonably finite width and depth in practice. This result is achieved under Besov spaces and Barron spaces, which is beyond the "linear" regime for better understanding nonlinearlity in deep RL. These results could also motivate practitioners to consider different architectures of implementations of deep RL.

Acknowledgments

My talk for AAAI New Faculty Highlights 2004 was supported by University of Warwick. The presented work is theoretical and generally will have no negative societal impacts.

References

Liu, F.; Suykens, J. A.; and Cevher, V. 2022. On the Double Descent of Random Features Models Trained with SGD. In *Advances in Neural Information Processing Systems*.

Liu, F.; Viano, L.; and Cevher, V. 2022. Understanding Deep Neural Function Approximation in Reinforcement Learning via ϵ-Greedy Exploration. In *Advances in Neural Information Processing Systems*.

Wu, Y.; Liu, F.; Chrysos, G. G.; and Cevher, V. 2023. On the convergence of shallow Transformers. In *Advances in Neural Information Processing Systems*.

Zhu, Z.; Liu, F.; Chrysos, G. G.; and Cevher, V. 2022. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in Neural Information Processing Systems*.