Comprehensive Understanding of Double Descent

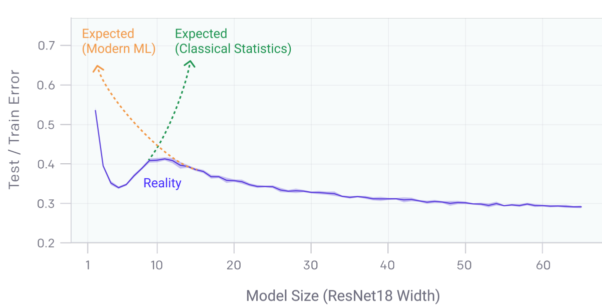
Rudrajit Das Anish Acharya

1. Background

We focus on the phenomenon of double descent in deep learning wherein when we increase model size or the number of epochs, performance on the test set initially improves (as expected), then worsens but again starts to improve and finally saturates, which is against conventional wisdom [2]. This phenomenon has already been demonstrated on traditional machine learning models (Opper, 1995; 2001; Advani & Saxe, 2017; Spigler et al., 2018; Geiger et al., 2019; Belkin et al., 2019a) but more recently was also observed in complex deep learning models (Nakkiran et al., 2019). There have also been attempts at mathematically explaining double descent for simple linear regression settings (Belkin et al., 2019b; Nakkiran, 2019; Dereziński et al., 2019).

2. Understanding effects of Label Noise and Regularization

In this project, we aim to investigate double descent more deeply and try to precisely characterize the phenomenon under different settings. Specifically, we shall focus on a couple of aspects - the impact of **label noise** and **regularization** on double descent. None of the existing works consider these aspects in detail and we hypothesize that these play an integral role in double descent.
Conversely, we pose a particular question: can we mitigate double descent by applying adequate regularization in the case of noisy data?
Further, we also plan to investigate if suitable regularization can bias the trajectory of gradient-based optimization algorithms in such a way that double descent can be mitigated, i.e. we do not observe an intermediate dip at all or observe a very small intermediate dip in the test performance.



3. Datasets

We shall try to reproduce the results of (Nakkiran et al., 2019) which use commonly used datasets such as CIFAR-10, CIFAR-100, etc. on common architectures such as ResNets, VGG, etc. We shall also try to observe this in simple problems such as linear regression. More importantly, we shall try to observe its variation with different kinds and degrees of regularization as well as different noise levels. (Nakkiran et al., 2019) do not consider regularization in too much detail but we think it is a critical factor controlling the extent of double descent.

4. Rough Timeline

- Preliminary experiments and theory on linear regression models - By Mid April.
- Deep learning experiments (and theory if possible) - By Early May.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

Dereziński, M., Liang, F., and Mahoney, M. W. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Nakkiran, P. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Opper, M. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pp. 922–925, 1995.

Opper, M. Learning to generalize. *Frontiers of Life*, 3(part 2):763–775, 2001.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.