Manipulating Sparse Double Descent

Ya Shi Zhang[*]
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
ysz23@cam.ac.uk

Abstract

This paper investigates the double descent phenomenon in two-layer neural networks, focusing on the role of L1 regularization and representation dimensions. It explores an alternative double descent phenomenon, named 'sparse double descent'. The study emphasizes the complex relationship between model complexity, sparsity, and generalization, and suggests further research into more diverse models and datasets. The findings contribute to a deeper understanding of neural network training and optimization. The code is available at https://github.com/yashizhang/sparsedoubledescent.

1 Introduction

In the modern era of deep learning, it is commonplace for practitioners to specify models that have *many more* parameters than the number of training data. This is due to a statistically non-intuitive phenomenon known as *double descent* [Belkin et al., 2019]. In a first course on regression, we are often taught the dangers of *overfitting*, when the model fits to the noise in the data and severely deteriorates performance across the entire population. For example, consider polynomial regression, where the degree of the polynomial is significantly higher than the number of training data points. This leads to the notion of a trade-off between the bias and variance of the model, inducing a U-shaped curve (depicted in fig. 1) where the vertical axis depicts the population risk and the horizontal axis depicts the ratio between the number of model parameters and number of data points.

Recently, Curth et al. [2023] has attempted to bridge the gap of our understanding on the prevalence of the double descent phenomenon in non-deep learning models, such as trees, boosting, and linear regression — methods known as *smoothers*. Notably, analysis of these model classes affirmed the wisdom that model complexity and capacity is not necessarily an affine function of the number of model parameters. Rather, there are higher dimensional substructures in the set of model parameters such that increasing model parameters in certain directions do not cause overfitting.

In this paper, we examine the possibility of controlling the double descent phenomenon on simple two layer neural networks with varying representation dimension. We view the first layer as a learned approximation of the 'ground truth kernel,' and the second layer as a linear classifier. Particularly, we extend Curth et al. [2023] by first considering LASSO regressors, then implementing LASSO via L_1 regularization in neural network training as a sparsifier. This is motivated by the perspective of the L_1 regularizer as a convex surrogate for the "L_0" objective, aiming to prune down the number of parameters active in the network. Benefits of pruning and/or sparsification include (1) computational and storage savings, (2) model capacity regularization motivated by Occam's razor [LeCun et al., 1989], and (3) potential lottery ticket subspaces [Frankle and Carbin, 2019].

Furthermore, the prevalence of a 'sparse double descent' first noted by He et al. [2022] — as model sparsity increases, test performance first decreases then increases (before finally decreasing as

[*] https://yashizhang.github.io/

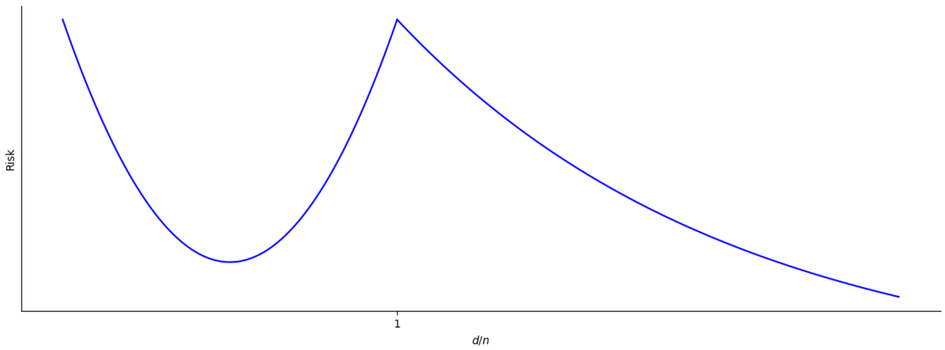


Figure 1: Simple depiction of the double descent phenomenon. d refers to the number of the parameters in the model and n refers to the number of training data points.

sparsity approaches 100%) — opens the possibility of model sparsity also exhibiting many intrinsic complexity axes.

2 Background

Consider the supervised learning problem, we are given training data $(x_i, y_i)_{i=1}^n$. In a family of functions $\mathcal{F} := \{f_\theta(\cdot) \mid \theta \in \mathbb{R}^d\}$ parameterized by $\theta \in \mathbb{R}^d$, we would like to choose a θ^* such that the function induced $f_{\theta^*}(\cdot)$ minimizes some loss function over the training data, usually of the form $\mathcal{L}(\theta) := \sum_{i=1}^n \ell(f_\theta(x_i), y_i)$. Depending on whether the task is classification or regression, common choices for $\ell(\cdot,\cdot)$ would be the cross entropy loss or quadratic loss, respectively.

We now modify our loss function by adding a *regularizer*, typically of the form $\alpha||\theta||_p$ for some $p \in (0, \infty]$ and $\alpha \in \mathbb{R}^+$. In this paper we focus exclusively on the case of $p = 1$. This L_1 penalty — inspired by the LASSO regressor in linear regression — inductively biases the network to learn parametrizations that are sparse and well-generalizing. The former is due to the L_1 penalty being the projection of the L_0 penalty (counting the number of non-zero elements in a vector) to the space of convex functions [Ramirez et al., 2013].

If our family of functions are neural networks with $L > 1$ layers, then we can view the first $L - 1$ layers as a learnable approximation to some unknown, 'ground truth' kernel. Through this perspective, we can view the training of neural networks as first learning an appropriate kernel function, then performing linear regression on the kernelized inputs. This is theoretically supported in the lazy training regime [Atanasov et al., 2022, Geiger et al., 2020, Jacot et al., 2021], and empirically supported via many observed phenomena such as the prevalence of Neural Collapse observed in Papyan et al. [2020].

3 Experiments

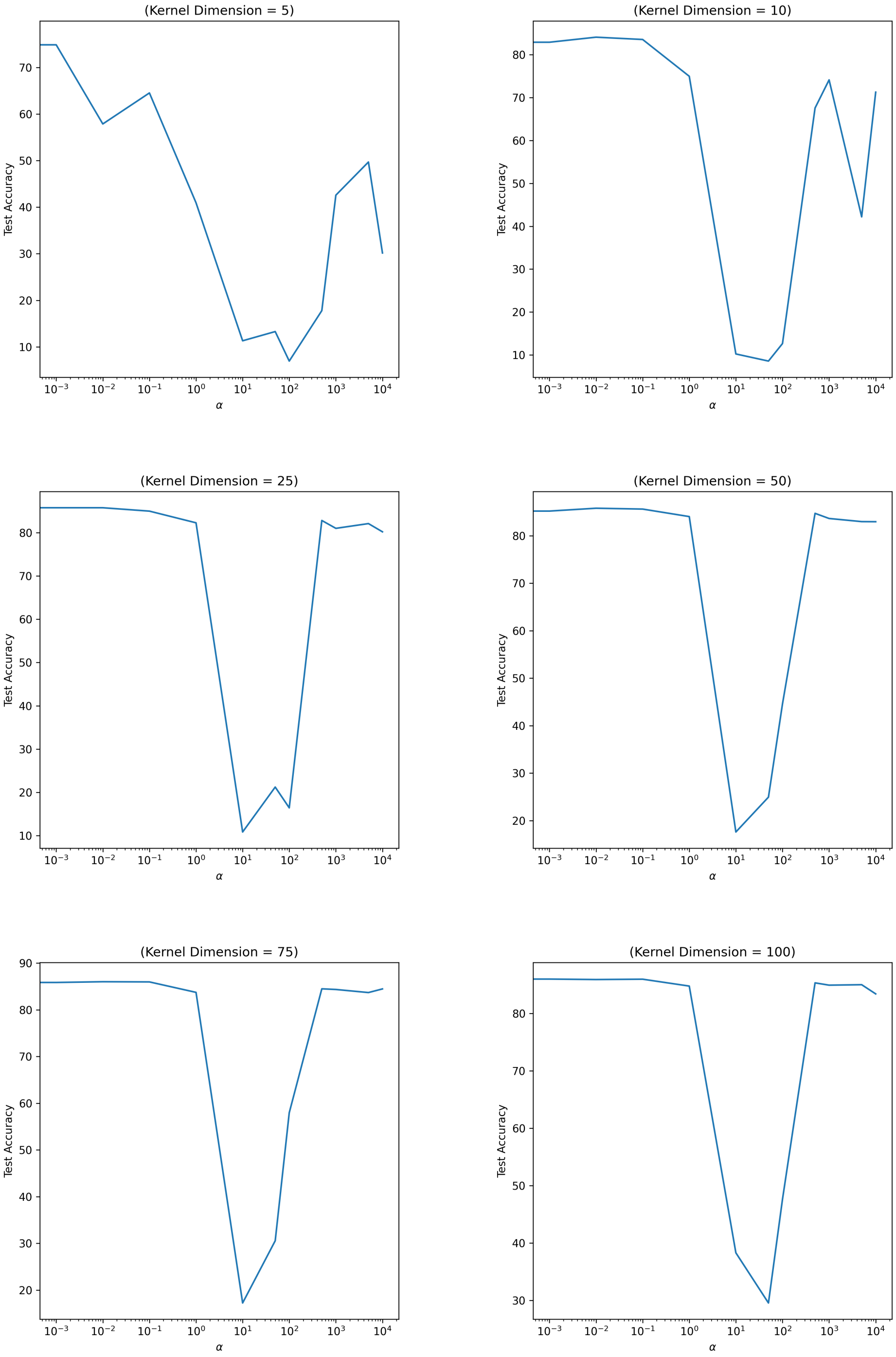
For our experiments, we will be training a two-layer multilayer perceptron with ReLU non-linearity on the MNIST dataset [LeCun et al., 2010] using stochastic gradient descent. The empirical experiments comprise of modifying two components of the training procedure to observe a few interesting phenomena. The first is the regularization coefficient α. As we increase α, we place heavier emphasis on sparsity when training the neural network. The second is the intermediate layers' dimension, which we refer to as the *kernel dimension*. By varying this, we are essentially changing the learned kernel's intrinsic dimension.

3.1 Results

As seen in table 1, for each kernel dimension, we train separate neural networks with varying α. In the top right corner, we have set the middle layer of our two-layer multi-layer perceptron to have 5 neurons. In the top middle, we have 10 neurons. Following this, we also show results for 25, 50, 75, and 100 intermediate neurons.

We vary the coefficient of the L_1 regularizer to first observe, for all choices of number of intermediate neurons, a version of the sparse double descent phenomenon. We also observe the existence of many ascents and descents as we decrease the number of intermediate neurons. Finally, we see that the location of the minima is invariant with respect to α. This seems to suggest the independence of the sparse double descent phenomena with respect to each layer's width in a fixed neural network architecture.

Table 1: The experiment results as detailed in section 3.



4 Discussion

Our exploration of the double descent phenomenon in simple two-layer neural networks has yielded some novel insights into model capacity, sparsity, and double descent phenomenona. The observed behavior in networks under varying representation dimensions and L1 regularization strenghts underscores a nuanced understanding of the balance between model complexity and overfitting.

However, this study's focus on specific network architectures and datasets suggests that broader research is required to generalize these findings. Future work should aim to validate these phenomena in more complex models and diverse datasets. Ultimately, this study contributes to the evolving discourse on neural network training, positioning sparse double descent as a key consideration in the pursuit of optimal model performance. This understanding aligns our work with the broader trajectory of machine learning research, highlighting the dynamic and often counter-intuitive nature of model training and generalization.

4.1 Future Directions

A future direction for this research could be to expand the experiments to include pretrained vision models fine-tuned on the MNIST dataset. This approach would allow for an exploration of the neural network up to the penultimate layer as a learned kernel, with the retraining of the final layer classifier under different objectives or settings providing insights into various forms of kernel regressions, such as ridge and LASSO.

Additionally, exploring other sparsification and pruning methods could be highly valuable. Theoretical underpinnings for these methods could be provided through Neural Tangent Kernel theory or other kernel theories. Integrating intuitive explanations using concepts like Kolmogorov complexity and Minimum Description Length (MDL) principles would further enhance the understanding. Examining the role of data equivariances and symmetries in these contexts would also contribute to a more comprehensive understanding of the dynamics influencing neural network performance and complexity.

References

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=1NvflqAdoom`.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, August 2019. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/doi/10.1073/pnas.1903070116`. Publisher: Proceedings of the National Academy of Sciences.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: Rethinking parameter counting in statistical learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=O0Lz8XZT2b`.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, nov 2020. doi: 10.1088/1742-5468/abc4de. URL `https://dx.doi.org/10.1088/1742-5468/abc4de`.

Zheng He, Zeke Xie, Quanzhi Zhu, and Zengchang Qin. Sparse double descent: Where network pruning aggravates overfitting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8635–8659. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/he22d.html`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks (invited paper). In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 6, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3465355. URL `https://doi.org/10.1145/3406325.3465355`.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL `https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf`.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Carlos Ramirez, Vladik Kreinovich, and Miguel Argaez. Why l1 is a good approximation to l0: A geometric explanation. *Journal of Uncertain Systems*, 2013.