NO DOUBLE DESCENT IN SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we present evidence that double descent does not exist in self-supervised settings. To our knowledge, this is the first time this class of models' deviation from the known behavior has been identified. This conjecture is based on an extensive experiment with an autoencoder and results from the literature. Instead of the classical double descent curve, the test loss either has a classical U-shape or monotonically decreases, depending on how the model is overparameterized. We briefly discuss a possibility as to why self-supervised models behave differently. We believe that further research in this direction will help elucidate the theoretical underpinnings of deep learning.

1 INTRODUCTION AND BACKGROUND

Double descent (DD) has been a hallmark phenomenon of deep learning, showing the rift between classical theory and deep-learning systems' generalization capabilities (Belkin et al., 2019). This is caused by a model's test loss deviating from the classical U-shape caused by the bias-variance tradeoff, once it becomes overparameterized (Neal et al., 2018; Belkin et al., 2019; Nakkiran et al., 2019). This form of the phenomenon is known as model-wise DD, where a model's number of parameters is increased until it reaches a point of high test loss known as the interpolation peak. This peak occurs when the number of parameters equals the dataset's size times the size of the model's output, such that it can just interpolate the data (Belkin et al., 2019). Beyond it the model continues interpolating, resulting in the training loss going to zero and the test loss monotonically decreasing as model capacity increases.

However, these insights are not necessarily universal as they are almost solely based on supervised models. This begs the question, what happens in the self-supervised setting? The few works that asked this found that DD is not as universal as it seemed. For example, Gedon et al. (2023) showed analytically and empirically that double descent does not occur in PCA. Luzi et al. (2021) found that Generative Adversarial Networks (GANs) do not exhibit it either, except in an artificial pseudo-supervised setting. Dar et al. (2020) contrasted the unsupervised and supervised settings by showing that when going from the former to the latter in subspace fitting tasks DD gradually appears. Dubova (2022) observed double descent in an AE but this result's implications are debatable, as it is unclear whether their task is supervised and the location of their interpolation peak seems to be independent of the model's capacity.

Following these findings, we hypothesize that *double descent does not occur in self-supervised systems*. We supply further evidence for this claim through an empirical analysis of an autoencoder (AE) (Kramer, 1991; Hinton et al., 2011), showing that it too does not exhibit DD. We choose an AE because of the tension with Dubova (2022)'s results and it having not yet been analyzed in the literature. We hope that these findings will help elucidate the theory of overparameterized models.

2 EXPERIMENT

We search for DD when training an AE on artificial data, akin to the analysis done in Gedon et al. (2023). A detailed overview of the experiment can be found in Appendix A.[1] The dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ is generated through the relation

[1]The experiment's code will be given in the camera-ready version.

$$\boldsymbol{x}_i = \boldsymbol{D}\boldsymbol{z}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$, are the latent variables, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{I}_n)$ are noise vectors, n is the number of features, d is the data's latent dimension, and $D_{ij} \sim \mathcal{N}(0, r)$ is a random matrix used to project the latent features to the higher-dimensional space, where r is a constant used to control the signal-to-noise ratio (SNR).

To learn this data's latent distribution we used an AE with fully-connected single-layer decoders and encoders with the same number of hidden units. The latent layer's width was varied independently of the hidden layers' width because of uncertainty regarding where DD would occur. Previous results Dubova (2022) imply that if DD occurs in AEs then it might be independent of the hidden width. Moreover, the interpolation peak is expected to be where the number of model parameters equals the dataset's size times the size of the output, but it is unclear what is the output's size here. For example, one could argue for it being the AE's latent dimension or its number of features. These issues were dealt with by using a wide range of hidden and latent dimensions, which should cover both cases.

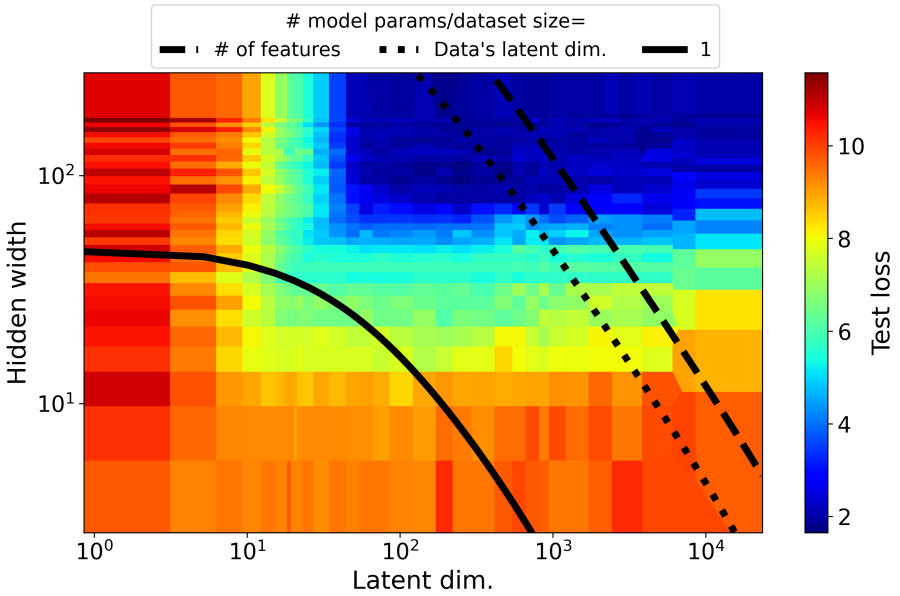

Figure 1: No double descent-like interpolation peak is observed across different latent and hidden widths. When going from a low to high latent dimension the loss has a bias-variance-like U-shape. However, when increasing the hidden dimension the loss only decreases. The black lines signify locations where the interpolation peak would be if assuming different output sizes.

The results are shown in Figure 1. We find that the test loss never exhibits DD, with its specific behavior depending on how the model was overparameterized.

3 Discussion

We have found evidence for double descent not occurring in self-supervised settings. This hypothesis is based on an experiment with an AE that adds to the literature which supports this finding. While our experiment analyzes a relatively simple system, we believe our findings are valid because of the generality that double descent exhibits in similar supervised models (Belkin et al., 2019). This discrepancy between supervised and self-supervised settings may be because of the different problem formulations—while the former assumes noisy data, in the latter case the model tries to learn a distribution. A fundamental analysis and more extensive experiments are necessary to test this claim. As a theory is best understood through its shortcomings, we believe that building on these conjectures will help advance our understanding of overparameterized systems.

ACKNOWLEDGEMENTS

Will be given in the camera-ready version.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

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, jul 2019. doi: 10.1073/pnas.1903070116. URL https://doi.org/10.1073%2Fpnas.1903070116.

Yehuda Dar, Paul Mayer, Lorenzo Luzi, and Richard G. Baraniuk. Subspace fitting meets regression: The effects of supervision and orthonormality constraints on double descent of generalization errors, 2020. URL https://arxiv.org/abs/2002.10614.

Marina Dubova. Generalizing with overly complex representations. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022. URL https://openreview.net/forum?id=qeu7F3VW8qS.

Daniel Gedon, Antonio H. Ribeiro, and Thomas B. Schön. No double descent in PCA: Training and pre-training in high dimensions, 2023. URL https://openreview.net/forum?id=ieWqvOiKgz2.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pp. 44–51. Springer, 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *Aiche Journal*, 37:233–243, 1991.

Lorenzo Luzi, Yehuda Dar, and Richard G. Baraniuk. Double descent and other interpolation phenomena in GANs. *CoRR*, abs/2106.04003, 2021. URL https://arxiv.org/abs/2106.04003.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt, 2019. URL https://arxiv.org/abs/1912.02292.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks, 2018.

A EXPERIMENT DETAILS

A.1 PARAMETER VALUES

The parameters used for the data generation and training procedures are given in Table 1. The optimizer used for training was Adam (Kingma & Ba, 2014) and the reconstruction loss was MSE.

Parameter Name	Value
Learning rate	0.001
Epochs	200
Batch size	10
Data's latent dimension, d	20
Number of features, n	50
Train dataset size, N	5000
Test dataset size	10000
SNR	10

Table 1: Parameters used for the AE experiments.

A.2 SNR

To control the SNR, r is chosen such that $\frac{\mathbb{E}[||\boldsymbol{Dz}||_2]}{\mathbb{E}[||\boldsymbol{\epsilon}||_2]}$ has the desired value. As all the random variables are Gaussian, this amounts to setting $r = \frac{\text{SNR}^2}{d}$.

A.3 Plot details

The parameter space shown in Figure 1 was not sampled in a linear or logarithmic grid because some regions were expected to have a higher chance of exhibiting DD than others. To show all the results at once, the parameter space was colored based on the loss of the closest measured point. The logarithmic distance is used instead of the normal one, so 100 is closer to 1000 than it is to 1.

A.4 Training loss

The training loss corresponding to the results shown in Figure 1 is given in Figure 2. One can see that the AE transitions to the interpolation regime in a large part of the parameter space, which is necessary for DD to occur. This transition happens when the latent dimension is about 20, which is the data's latent dimension, or when the hidden width is about 28. This could imply that interpolation is possible only when the AE does not have an information bottleneck induced by its layers' widths.

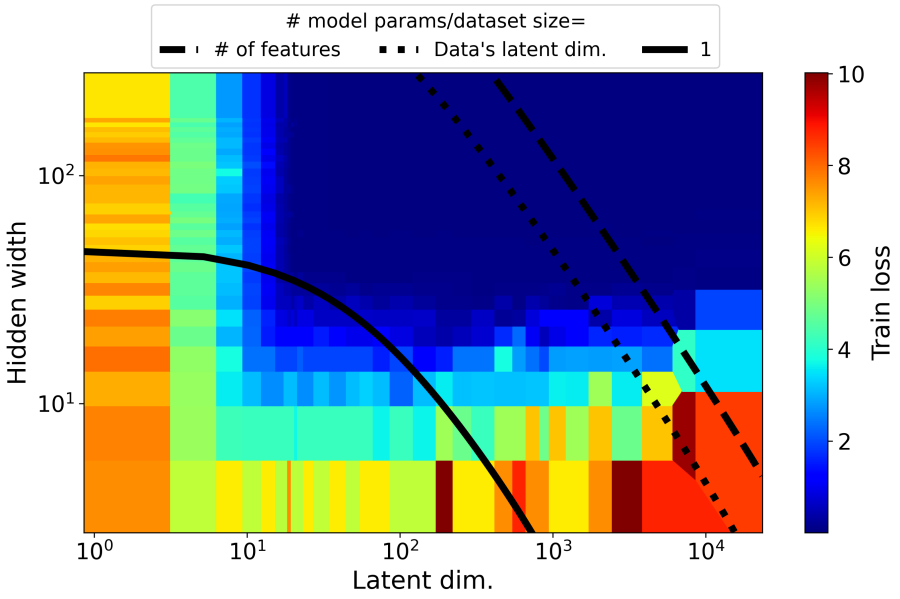

Figure 2: The train loss for the experiment discussed in section 2.