Double-Descent Literature Review: Experimental Observations, Practical Uses and Possible Explanations

Joe Down

Abstract

Modern machine learning practices have revealed a relationship between model capacity and generalisation performance which seems to contradict classical statistical understanding. Referred to as double-descent, this relationship extends the well-known bias-variance understanding by showing that, beyond a certain threshold, increases in model capacity result in monotonically increasing performance. This literature review explores where this behaviour has been observed, how it has been induced, and context specific hypotheses for why it may occur. Areas of research which could provide more general insight into the occurrence of double-descent will also be presented. Particular focus will be directed toward the relationship between noise in training data and double descent's occurrence, the usefulness of eliminating double descent using regularisation, and recent findings regarding triple and multiple descent.

1 Introduction

The suitability of supervised machine learning models for regression is typically measured by evaluating an appropriate risk measure comparing predicted results with 'real' results from an unseen validation dataset[13]. Training algorithms seek to find model parameters (and usually also hyper-parameters) resulting in minimal risk. This is considered indicative of a well generalising model[13].

Models of different capacities (e.g. logistic regression with different class of polynomials) can result in differing overall minimal risk under their respective optimal parameters. Classically, it has been understood that functions of very low or very high capacity perform poorly. Optimal capacity exists somewhere between. This is a consequence of the bias-variance trade-off[12]. Models of low capacity usually do not prove rich enough to describe the characteristics of the training data's underlying data generating distribution (under-fit) and models of high capacity tend to over-fit noise in the training data rather than the true underlying distribution. This relationship can be visualised with the generalisation curve in Figure 1A by Belkin et al. [3].

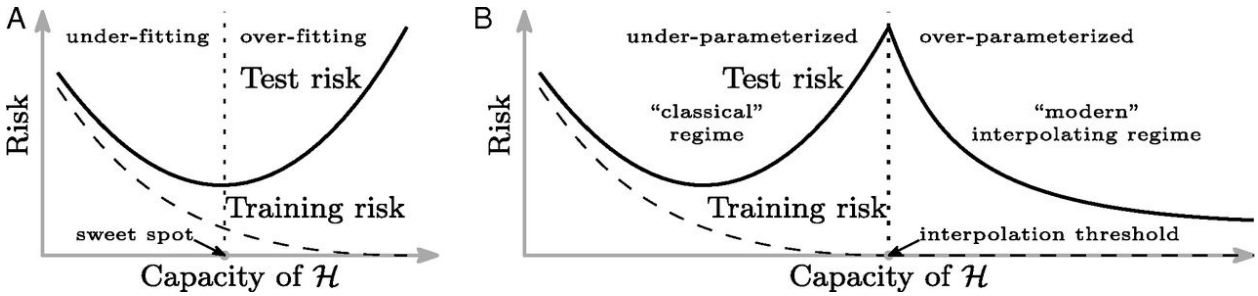

Figure 1: Generalisation curves comparing the classical understanding of model complexity's impact on risk minimisation with modern practical findings[3]

More recently, investigation into training with functions of very large capacity has resulted in an observation seemingly contradicting the implications of the bias-variance trade-off. When a certain interpolation threshold is exceeded, risk begins to decrease again monotonically, usually dropping below the minimal risk which would be found within the classical U-shaped generalisation curve, as can be seen in Figure 1B. This has come to be known as double descent and understanding why this occurs would have numerous implications and uses. Currently however, only domain specific hypotheses exist rather than any widely accepted unifying understanding.

This literature review will begin by justifying belief in the bias-variance trade-off and why this makes double-descent unexpected, before discussing the existence of double-descent in its main forms: model-wise, sample-wise, and epoch-wise. Additionally, some hypotheses for why double-descent occurs within each setting will be provided. Other interesting finings will then discussed, in particular regarding the role of training label noise, the impact of regularisation, and how these may induce or reduce double-descent. This will be followed by discussion of some cutting edge research into triple and multiple descent which provide possible insights into where further understanding could be developed.

2 Justifying Bias-Variance

It seems important to be aware of the prevalence of the bias-variance trade-off in present machine learning knowledge to understand the unexpectedness of double descent. Many common risk measures in machine learning have a bias-variance decomposition which defines risk in terms of bias, variance, and an irreducible error. Classically this is shown for squared error, however unified decompositions such as that of Domingos [10] also exist. The bias-variance decomposition for squared error (between true value y and function result $\hat{f}(x;D)$ for training dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$) can be understood as follows (proofs can be found in Friedman [11] and Brofos et al. [5]):

$$\mathrm{E}_{D,\varepsilon}\Big[\big(y - \hat{f}(x;D)\big)^2\Big] = \Big(\mathrm{Bias}_D\big[\hat{f}(x;D)\big]\Big)^2 + \mathrm{Var}_D\big[\hat{f}(x;D)\big] + \sigma^2$$

In other words, the square error over the training set equals the sum of the squared bias plus the variance plus an irreducible error, σ^2. Bias is characterised as:

$$\mathrm{Bias}_D\big[\hat{f}(x;D)\big] = \mathrm{E}_D\big[\hat{f}(x;D) - f(x)\big]$$

Or the difference between each function result and true function result over the training set. Variance is characterised as:

$$\mathrm{Var}_D\big[\hat{f}(x;D)\big] = \mathrm{E}_D[\big(\mathrm{E}_D[\hat{f}(x;D)] - \hat{f}(x;D)\big)^2]$$

As can be seen, outside of bias and variance the only other factor which can affect our risk is the irreducible error, which is generally understood to result from noise in our training data y values. It therefore seems that the bias-variance understanding is well justified with little room for extension.

3 Settings for Double-Descent

3.1 Introduction

In Nakkiran et al. [21], three key settings for double descent are identified: model-wise, sample-wise, and epoch-wise double descent. These terms seem to have been consistently accepted as terminology in most succeeding literature. The naming of each setting refers to the measure for model capacity against which risk is evaluated (see Section 1 and Figure 1). Important to note are the differences between hypothesised explanations for each setting, making it unclear whether they are all examples of the same phenomenon or just produce similar generalisation curves resulting from different causes.

3.2 Model-wise Double Descent

Model-wise double descent refers to the setting where double descent is a consequence of increasing the "size" of a model.

One measure of size could be the number of parameters per sample[21]. According to Deng et al. [9], Kini et al. [17], and d'Ascoli et al. [8], this model-wise double descent generalisation curve can be characterised around the ratio $\kappa = p/n$. Here, p is the number of parameters per sample and n is the number of test samples. Initially, while $\kappa < \kappa*$, the bias variance generalisation curve is observed. The double descent interpolation threshold is reached at some critical $\kappa*$, generally observed to occur when p and n are close[2] (i.e. $\kappa \approx 1$, though papers such as Deng et al. [9] have also observed values around $\kappa = 0.5$). The second descent follows as the model becomes increasingly "over-parameterised" and $\kappa > \kappa*$. Belkin et al. [3] shows this type of model-wise double descent occurring using both square and zero-one loss measures with a fully connected neural network. Their resulting generalisation curves as parameter count is increased can be seen in Figure 2a.

Model size can also be increased by changing properties of the model (e.g. by adjusting hyper-parameters) which result in increased model complexity. Belkin et al. [3] shows model-wise double descent of this type for random forest models, with increased complexity arising from increasing the maximum allowed leaves per tree, as seen by the generalisation curve in Figure 2b.

Attempts have also been made to provide some model-specific explanations for model-wise double descent. Belkin et al. [2] for example provides a mathematical analysis for the Gaussian model and Fourier series model generalisation curves, both using a least squares risk predictor to justify double descent under these conditions. Cao et al. [6] and Hastie et al. [14] meanwhile characterise benign over-fitting in linear classification with over-parametrisation.

3.3 Sample-wise Double Descent

Sample-wise double descent occurs where double descent is a consequence of increasing the number of samples in the training data set[21].

Min et al. [20] are able to observe double descent in Gaussian mixture classification with linear loss using an adversarial training regime. They also identify that, depending on the strength of the adversary different outcomes occur. This can be observed from the generalisation curves in 2c. Of particular interest is the medium adversary regime, under which the sample-wise double descent is observed. Performance with the strong adversary is also notable however as it was found that increasing sample count beyond a certain threshold produced a monotonically worsening test loss.

Nakkiran et al. [21] provides intuition into why increasing sample count is not always helpful. They identify that while increasing sample count generally reduces loss, it can also result in a shift of the interpolation threshold to the right, creating regions where a lower sample count also has lower risk. This can be observed in Figure 2d from Nakkiran et al. [21], where this shift can be seen to result in greater test loss with 4x the sample count for models of certain dimensionalities.

3.4 Epoch-wise Double Descent

Epoch-wise double descent occurs as iterative training routines reach and exceed a critically large epoch count[21].

Kawaguchi et al. [16] were able to observe epoch-wise double descent while investigating fully-connected neural networks. When tasked to reconstruct an identical duplicate of a simple black and white image of a character, trained only on a single image of a different character using the Adamax algorithm, the behaviour was observed. As can be observed in 2e, first the image produced improves from epoch 1 to 10 and gets progressively worse again from epoch 10 to 1,000. This region resembles the generalising behaviour of

bias-variance. The interpolation threshold is reached however somewhere between epoch 1,000 and 10,000, after which loss was measured to (and can be seen to) monotonically decrease, as would be expected from a second descent.

Heckel et al. [15] attempt to provide an explanation for epoch-wise double descent. They claim that it is caused by the superposition of multiple bias-variance trade-offs originating from different parts of the model which are learned at different epochs. When the minima of the model parts' resulting generalisation curves occurs at different training epochs, the generalisation curve for the overall model takes on the appearance of the double descent generalisation curve. A consequence of this is that they were able to eliminate epoch-wise double descent in two convolutional neural networks by adjusting step sizes for different layers. Pezeshki et al. [23] come to a similar conclusion, believing their occurrence of double descent to result from features being learned at different time scales.

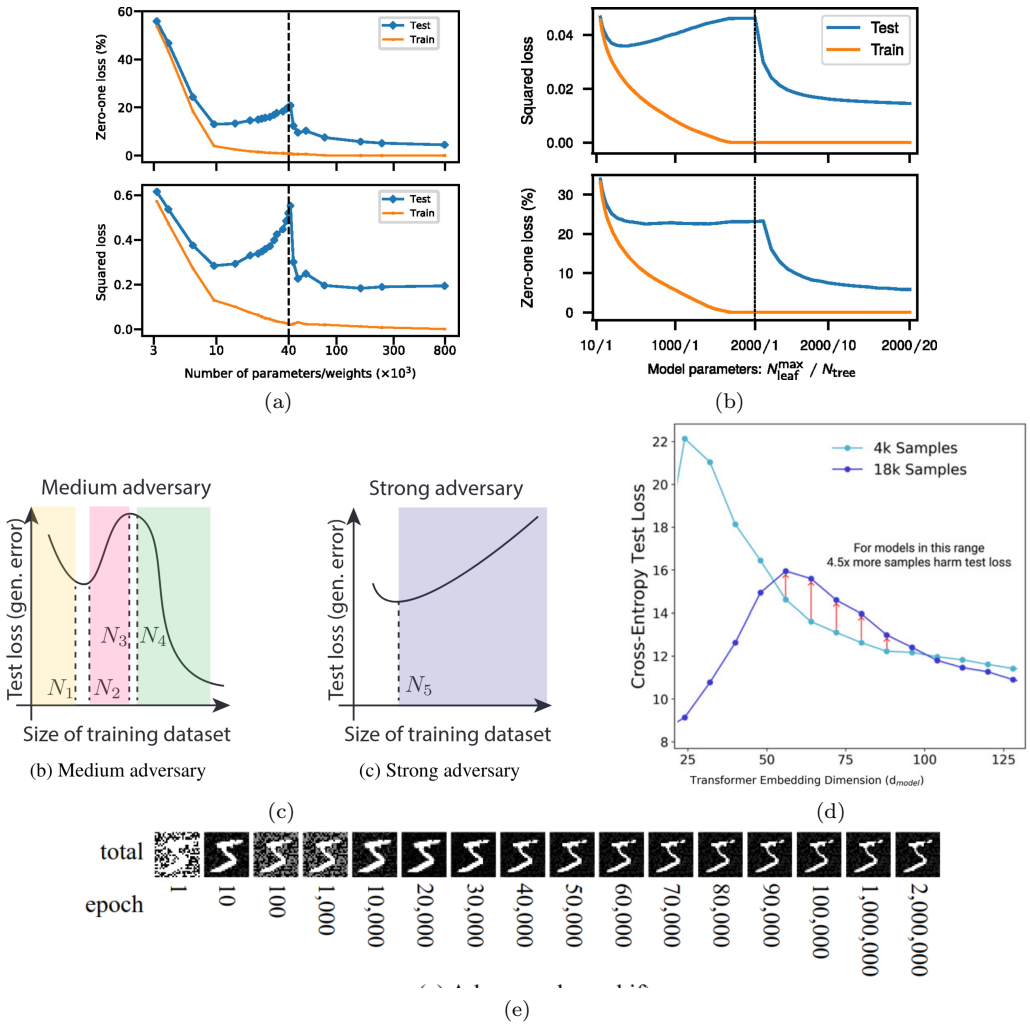

Figure 2: From Belkin et al. [3], Belkin et al. [3], Min et al. [20], Nakkiran et al. [21], and Kawaguchi et al. [16] respectively

4 Noise and Regularisation

4.1 Noise

A recurring feature which seems linked to the concept of double descent is the presence of noise. Luzi et al. [19] and Nakkiran et al. [21] provide some insight into the nature and importance of noise in this setting.

Luzi et al. [19] gives an example of introducing double descent behaviour into an unsupervised setting. They observe that over-parameterised Generative Adversarial Networks (GANs) do not display model-wise double descent. This can be seen from the generalisation curves in Figure 3a where, when the model begins to interpolate (train error becomes 0) at dimensionality n, test error becomes constant.

However, by developing a pseudo-supervised approach involving the introduction of artificial noise inputs in combination with real outputs, they were able to observe model-wise double descent behaviour. By doing this, they were able to achieve similar or superior generalisation performance to the typical approach with faster training times. Figure 3b shows these results. n_{ps} represents the number of artificial samples inserted. We see no double descent without artificial noise present ($n_{ps} = 0$). Equations 4, 5 and 6 are the same experiment but with different model training parameters used, resulting in different double descent patterns (or no double descent at all).

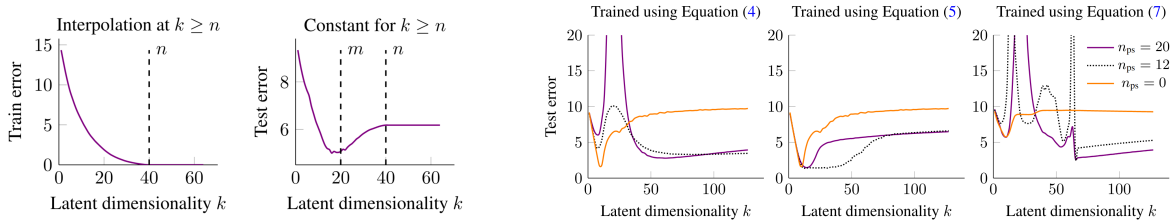

Figure 3: From Luzi et al. [19]

Nakkiran et al. [21] also comment on artificial noise and how they believe double descent can still occur with this in place of truly random label noise. They make the claim that this is because double descent is not a direct consequence of label noise, rather it is a consequence of model mis-specification. Label noise, whether real or artificial, works to make the training data's distribution harder and therefore means more possible valid interpolations exist. This seems consistent with Luzi et al. [19]'s findings.

4.2 Regularisation

Linked to noise in machine learning is the concept of regularisation which aims to avoid over-fitting by constraining model coefficients[4, 25]. It has been shown by Nakkiran et al. [22] that regularisation can be used to eliminate double descent. They explain that this is useful as it ensures monotonic performance increases as model or sample size increase, avoiding early stopping which can occur in the double descent scenario at the first, inferior, minima within the bias-variance curve region.

Yilmaz et al. [24] provide a method for eliminating double descent with regularisation in a two layer neural network. Similar to how Heckel et al. [15] adjusted step size at each layer to eliminate epoch-wise double descent, Yilmaz et al. [24] adjust regularisation strength at each level to ensure risk minima superimpose sufficiently.

5 Triple & Multiple Descent

5.1 Triple Descent

A notable discovery relating to double descent is the observation of triple and more recently multiple descent. An example of triple descent can be seen from a previously discussed result in Figure 3b, under Equation 7 with $n_{ps} = 20$, though Luzi et al. [19] only note its presence and do not provide further insight in their discussion of results. More notable is d'Ascoli et al. [8] which identifies triple descent with the random Fourier feature model and provides an explanation for triple descent as the combined result of a linear and a non-linear peak (with regards to loss) which can occur at different points. They claim that the linear peak results from overfitting label noise, while the non-linear peak results from a combination of label noise and features of the model itself. Double descent is believed to occur in scenarios where these peaks occur at the same point (the interpolation threshold), implying that double descent is not unique but just a special case of triple descent.

5.2 Multiple Descent

Further to this, Chen et al. [7] introduces the concept of multiple descents. They show, for an over-parameterised linear regression model, that the generalisation curve can actually have an arbitrary number of peaks and therefore an arbitrary number of descents. They also show that the number and location of the peaks in the generalisation curve can be controlled, implying that triple descent, double descent and the bias-variance generalisation curves may all be different cases of the same phenomenon.

Adlam et al. [1] was also able to observe multiple descents while investigating over-parameterised kernel regression with the Neural Tangent Kernel. They found an additional peak and then trough occurred as the number of parameters scaled quadratically with the training dataset size (i.e. a new decent started when the number of parameters became equal to the sample count squared) and hypothesised further peaks and troughs could occur further on. It's worth noting that neither of these papers propose clear use cases, advantages or disadvantages to having multiple descents as they are very recent and have had little to no further developments. Intuition from the advantages provided by eliminating double descent through regularisation perhaps implies multiple descent could be disadvantageous. Whatever the case, identifying the existence of multiple descents, and particularly identifying single and double descent as just specific cases, could prove useful to developing further understanding.

6 Conclusions

In summary, double descent exists seemingly in contradiction to the bias variance understanding of the relationship between model complexity and risk minimisation in machine learning. It is however observable, with exceptions, in all of its three main settings: model-wise, sample-wise and epoch-wise. It can be seen to be useful as it informs us of the validity of exploring beyond the seeming bounds of the bias-variance generalisation curve into regions where better generalising models can be found.

Generally it seems that current literature is not especially concerned with finding a single unifying understanding of double descent in all of its settings. Lee et al. [18] attempt to provide an explanation in terms of Vapnik–Chervonenkis theoretical analysis, however results are poorly presented and the paper does not seem to have been used by other more notable authors on the topic. Instead, a wide variety of context specific explanations and practical explanations have been provided. A summary of those discussed follows:

Source	Experiment	Result	Related Figures
Belkin et al. [3]	Random forest models of increasing complexity	Model-wise double descent observed	2a
Belkin et al. [3]	Increasingly parameterised fully connected neural network	Model-wise double descent observed	2b
Min et al. [20]	Adversarial Gaussian mixture classifier training with increasing sample counts	Sample-wise double descent observed in some scenarios, decreasing performance in others	2c
Nakkiran et al. [21]	Comparing generalisation curves for a transformer with different sample counts	Regions of decreased performance with increased sample counts identified	2d
Kawaguchi et al. [16]	Image reproduction generalisation over successive epochs	Epoch-wise double descent observed	2e
Heckel et al. [15]	Adjusting step sizes for neural network layers	Double descent eliminated, leading to monotonically improving generalisation	N/A
Luzi et al. [19]	Semi supervised GAN technique using artificial noise	Double descent introduced into an unsupervised setting	3
Yilmaz et al. [24]	Adjusting regularisation strength at different neural network layers	Double descent eliminated	N/A
d'Ascoli et al. [8]	Over-parametrisation with neural networks	Triple descent observed	N/A (3b (Equation 7) related)
Adlam et al. [1]	Over parameterised kernel regression with the Neural Tangent Kernel	Multiple descents observed	N/A

There are notable similarities in findings across some of the topics discussed which perhaps draw out some key findings. Heckel et al. [15]'s findings related to epoch-wise double descent, Yilmaz et al. [24]'s findings on eliminating double descent through regularisation, d'Ascoli et al. [8]'s hypotheses about triple descent, and Chen et al. [7] and Adlam et al. [1]'s hypotheses about multiple descent all seem to imply multiple descent phenomena exists as a result of the superposition of multiple different generalisation behaviours occurring at different rates. Luzi et al. [19] and Nakkiran et al. [21] both seem to agree that the presence of noise is useful to the occurrence of double descent and that the provenance of it is not important as it can be completely artificial. Nakkiran et al. [22], Yilmaz et al. [24], and Heckel et al. [15] all agree that elimination of double descent by changing model properties (such as by adding and adjusting regularisation) can be advantageous as it creates scenarios involving monotonic improvement in generalisation.

There are also a few contradictions. While Nakkiran et al. [21] claims that noise is not a cause of double descent but simply a cause of increased model complexity which leads to double descent, d'Ascoli et al. [8] identifies noise as a key cause of the linear peak in triple descent (which it also claims double descent is a special case of). It's also key to note that many papers somewhat imply by omission that increased complexity will always cause double descent, when scenarios such as those in Nakkiran et al. [21] and Min et al. [20] have been identified where this can have the opposite effect.

Ultimately it remains difficult to reconcile whether the many similarities imply a single unifying cause for double descent, or the differences, and particularly also this idea of superposition of model parts, imply

multiple different causes producing the same resulting generalisation curve.

References

[1] Ben Adlam and Jeffrey Pennington. "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 74–84.

[2] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features". In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1167–1180.

[3] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[4] Peter J Bickel et al. "Regularization in statistics". In: *Test* 15.2 (2006), pp. 271–344.

[5] James Brofos, Rui Shu, and Roy R Lederman. "A bias-variance decomposition for Bayesian deep learning". In: *NeurIPS 2019 Workshop on Bayesian Deep Learning*. 2019.

[6] Yuan Cao, Quanquan Gu, and Mikhail Belkin. "Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8407–8418.

[7] Lin Chen et al. "Multiple descent: Design your own generalization curve". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8898–8912.

[8] Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. "Triple descent and the two kinds of overfitting: Where & why do they appear?" In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3058–3069.

[9] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification". In: *Information and Inference: A Journal of the IMA* 11.2 (2022), pp. 435–495.

[10] Pedro Domingos. "A unified bias-variance decomposition". In: *Proceedings of 17th international conference on machine learning*. Morgan Kaufmann Stanford. 2000, pp. 231–238.

[11] Jerome H Friedman. "On bias, variance, 0/1—loss, and the curse-of-dimensionality". In: *Data mining and knowledge discovery* 1.1 (1997), pp. 55–77.

[12] Stuart Geman, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma". In: *Neural computation* 4.1 (1992), pp. 1–58.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[14] Trevor Hastie et al. "Surprises in high-dimensional ridgeless least squares interpolation". In: *The Annals of Statistics* 50.2 (2022), pp. 949–986.

[15] Reinhard Heckel and Fatih Furkan Yilmaz. "Early stopping in deep networks: Double descent and how to eliminate it". In: *arXiv preprint arXiv:2007.10099* (2020).

[16] Aoshi Kawaguchi, Hiroshi Kera, and Toshihiko Yamasaki. "Epoch-Wise Double Descent Triggered by Learning a Single Sample". In: ().

[17] Ganesh Ramachandra Kini and Christos Thrampoulidis. "Analytic study of double descent in binary classification: The impact of loss". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020, pp. 2527–2532.

[18] Eng Hock Lee and Vladimir Cherkassky. "VC Theoretical Explanation of Double Descent". In: *arXiv preprint arXiv:2205.15549* (2022).

[19] Lorenzo Luzi, Yehuda Dar, and Richard Baraniuk. "Double descent and other interpolation phenomena in GANs". In: *arXiv preprint arXiv:2106.04003* (2021).

[20] Yifei Min, Lin Chen, and Amin Karbasi. "The curious case of adversarially robust models: More data can help, double descend, or hurt generalization". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 129–139.

[21] Preetum Nakkiran et al. “Deep double descent: Where bigger models and more data hurt”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (2021), p. 124003.

[22] Preetum Nakkiran et al. “Optimal regularization can mitigate double descent”. In: *arXiv preprint arXiv:2003.01897* (2020).

[23] Mohammad Pezeshki et al. “Multi-scale feature learning dynamics: Insights for double descent”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 17669–17690.

[24] Fatih Furkan Yilmaz and Reinhard Heckel. “Regularization-wise double descent: Why it occurs and how to eliminate it”. In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2022, pp. 426–431.

[25] Chiyuan Zhang et al. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115.