ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Do We Really Need a New Theory to Understand the Double-Descent?

Luca Oneto, Sandro Ridella, Davide Anguita

University of Genoa - Via Opera Pia 11a, 16145, Genova, Italy

Abstract. This century saw an unprecedented increase of public and private investments in Artificial Intelligence (AI) and especially in Machine Learning (ML). This led to breakthroughs in their practical ability to solve complex real world problems impacting research and society at large. Instead, our ability to understand the fundamental mechanism behind these breakthroughs has slowed down because of their increased complexity. This questioned researchers about the necessity for a new theoretical framework able to help researchers catch up on this lag. One of the still not well understood mechanisms is the so called over-parametrization, namely the ability of certain models to increasing their generalization performance (reduce test error) when the number of parameters is above the interpolating threshold (zero training error), and the associated double-descent curve. In this paper we will show that this phenomena can be better understood using both known theories, i.e., the algorithmic stability theory, and empirical evidence.

1 Introduction

This century, thanks to an unprecedented increase of public and private investments[1], saw Artificial Intelligence (AI), and in particular Machine Learning (ML), deeply impacting the development of science [1, 2] and the society at large [3, 4]. Every branch of science is now empowering human-driven research with AI and, from self-driving cars to smart IoT devices, almost every consumer application now leverages AI-based technologies to make sense of the vast amount of available data collected and stored.

The availability of huge amounts of computing power and data coupled with the work of many researchers and practitioners allowed also to deeply transform AI and ML themselves. In the previous century, the most effective and efficient AI and ML were born from foundational research. In this century, empirical evidence has quickly surpassed our ability to fully understand the fundamental mechanism behind practical and effective algorithms. As a result researchers start questioning themselves and the community about the necessity for new theoretical frameworks able to help researchers catch up on this lag [5–7].

One of the still not well understood, even if studied from a long time [8] mechanism is the so-called over-parametrization and the associated double-descent curve [6, 9]. In fact, classical (shallow) ML mostly relies on the Empirical Risk Minimization (ERM) principle [10]. ERM suggests finding the function that fits (e.g., minimizes the empirical error) on a training set searching in a set of possibly unknown set of functions carefully tuned during the model selection phase [11]. The tuning procedure trades-off error on the training data and complexity of the solution. When the complexity is measured with the capacity of the set of functions this results in the Structural Risk Minimization (SRM) principle, also called bias–variance or under/over-fitting trade-off [6, 10, 12].

[1] https://ai-watch.ec.europa.eu/publications/ai-watch-index-2021_en

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

SRM tells you that you do not have to increase the capacity of you set of function once the error on the training set reached zero, the interpolation threshold, and there is an optimal point below this threshold for which the model will exhibit the optimal generalization capabilities with the so called U-Shaped generalization error curve [10, 12]. Modern (deep) ML models, instead, tends to push themselves after the interpolation threshold, in the so called over-parameterized regime, since empirical evidence show that after this threshold sometimes the generalization error tends to descend again, showing the so-called Double-Descent generalization curve [6, 9]. See Figure 1 to better catch these concepts.

In this paper, we argue that it is possible to better understand and explain the over-parameterized regime measuring the complexity of the algorithms with different measures [11, 13] (Section 2). In particular, using the the Algorithmic Stability [14, 15] (Section 3) and empirical evidence (Section 4) we will show that known theories are powerful enough to give insight and better explain the intriguing properties of the over-parameterized regime and the double-descent curve. Section 5 will conclude the paper.

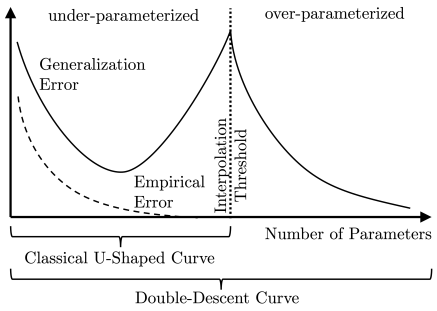

Fig. 1: Under- and Over-Parametrization and the corresponding U-Shaped and Double-Descent Curves.

2 Preliminaries

In this work we focus on supervised learning [10–12]. Based on a random observation of $X \in \mathcal{X}$ one has to estimate $Y \in \mathcal{Y}$ by choosing a suitable function $f : \mathcal{X} \to \hat{\mathcal{Y}}$ (characterized by a number of parameters) in a set of possible ones $\mathcal{F}$. A learning algorithm $\mathscr{A}_{\mathcal{H}}$, characterized by its set of hyperparameters $\mathcal{H}$, selects f from a possibly unknown $\mathcal{F}$ (induced by $\mathscr{A}_{\mathcal{H}}$) by exploiting a set of n labeled samples $\mathcal{D} : \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$. $\mathcal{D}$ consists of a sequence of independent samples distributed according to μ over $\mathcal{X} \times \mathcal{Y}$. The generalization (test) error $\mathsf{L}(f) = \mathbb{E}_{(X,Y)}\ell(f(X), Y)$ associated to function $f \in F$, is defined through a loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, 1]$. As μ is unknown, $\mathsf{L}(f)$ cannot be explicitly computed, but we can compute the empirical error (i.e., the empirical risk) namely the empirical (training) estimator of the generalization error $\hat{\mathsf{L}}(f) = {}^1\!/_n \sum_{(X,Y)\in\mathcal{D}} \ell\left(f\left(X\right), Y\right)$. The purpose of any learning procedure is to find the minimizer f^* of the generalization error $\mathsf{L}(f)$ ($f^* = \arg\min_{f\in\mathcal{F}} \mathsf{L}(f)$) but since $\mathsf{L}(f)$ is unknown we have to estimate h^* exploiting an empirical estimator defined by the algorithm based on the dataset $\hat{f} = \mathscr{A}_{\mathcal{H}}(\mathcal{D})$ (e.g., the empirical risk minimizer $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{\mathsf{L}}(f)$) and then estimate the generalization ability of $\hat{h}$ and then the quality of $\mathscr{A}_{\mathcal{H}}$. In this setting it is possible to prove that [13, 14]

$$\mathbb{P}\{\mathsf{L}(\hat{f}) \leq \hat{\mathsf{L}}(\hat{f}) + \mathsf{C}(\mathscr{A}_{\mathcal{H}}) + \phi(n, \delta)\} \geq 1 - \delta, \tag{1}$$

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

namely, the generalization error of $\hat{f}$ is bounded by the empirical (training) error plus a complexity term, plus a complexity term $\mathsf{C}(\mathscr{A}_{\mathcal{H}})$ which measures the risk due to the choice of the algorithms and its hyperparameters (i.e., the more the algorithm tends to memorize/fit/extract-information and not learn from the data the larger is this term), plus a confidence term[2] $\phi(n, \delta)$ which measures the risk associated to the sample (i.e., the less data we have or the larger confidence we require the larger is this term).

Different approaches from statistical learning theories allow us to obtain these bounds [11, 13]. If $\mathcal{F}$ can be explicitly defined (or estimated [16]) based on $\mathscr{A}_{\mathcal{H}}$ we can use the complexity based theories (e.g., the Vapnik–Chervonenkis or the Rademacher Complexity bounds), if the algorithms tends to deeply compress the original dataset the Compression theories are a good choice (e.g., the Compression or the Minimum Description Length based bounds), when the $\mathcal{F}$ cannot be explicitly defined the Algorithmic Stability Theory (e.g., Uniform or Hypothesis Stability) is a good option, when we have to deal with randomized model PAC-Bayes theory is a very strong approach, when we have to deal with randomized algorithms Differential Privacy theory is a promising research direction, and when the analysis is adaptive (i.e., when the algorithm choice depend on the data itself) information theory based bounds are surely the best option.

3 Theoretical Evidences

The classical approach to study the over-parameterized regime and discuss its ineffectiveness in understanding the double-descent curve is to use the complexity based theories in Bound (1) [6, 9]. In fact, using, e.g., the Rademacher Complexity, what happens is that $\mathsf{C}(\mathscr{A}_{\mathcal{H}}) = \mathsf{R}(\mathcal{F})$ with

$$\mathsf{R}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\sigma_1,\cdots,\sigma_n} \sup_{f\in\mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f(X_i), Y_i), \tag{2}$$

where $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2 \ \forall i \in \{1, \cdots, n\}$, namely the measure of the risk due to the choice of the algorithms and its hyperparameters becomes the capacity of the set of function from which the algorithms chooses the estimator $\hat{F}$. Note that $\mathsf{R}(\mathcal{F})$ can be estimated from the data [11]. Increasing the parameters of the model by changing $\mathcal{H}$ (i.e., increasing the number of neurons or layers in a deep model) will improve our ability to shrink the empirical error $\hat{\mathsf{L}}(\hat{f})$ and is equivalent to enlarging the corresponding $\mathcal{F}$ and consequently $\mathsf{R}(\mathcal{F})$. This means that there will be an optimal value of $\mathcal{F}$ between $\mathcal{F} = \oslash$ (the empty set of functions) and the interpolation threshold, i.e., $\mathcal{F}$ large enough to ensure $\hat{\mathsf{L}}(\hat{f}) = 0$. Above this this threshold, complexity based theories tells you that it is useless to proceed since $\hat{\mathsf{L}}(\hat{f})$ cannot decrees further and $\mathsf{R}(\mathcal{F})$ can only increase. This explanation does not capture the intrinsic mechanism behind the over-parameterized regime and the corresponding Double-Descent curve but it explains just the U-Shaped curve below the interpolation threshold (see Figure 1).

In this work we argue that, in order to understand and better explain the Double-Descent curve, the complexity based theories are not adequate but other theories can give insights and better explain the phenomena. Between the different available theories mentioned in Section 2, Algorithmic Stability (in particular

[2] We will not discuss this term since independent from $\mathscr{A}_{\mathcal{H}}$.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

the Hypothesis Stability) is surely the best option since it is the most general theory (works with deterministic/randomized models/algorithms) and allows to derive tight fully empirical bounds [11, 14, 15]. Contrarily to complexity based theories, Algorithmic Stability does not care about $\mathcal{F}$ from which $\mathscr{A}_{\mathcal{H}}$ selects $\hat{f}$ but it measures how much the model changes by slightly changing $\mathcal{D}$. The idea is simple: if $\hat{f}$ does not change changing $\mathcal{D}$ this means that the algorithm is learning from the data and does not simply fit/memorize them. Consequently, in this case $\mathsf{C}(\mathscr{A}_{\mathcal{H}}) = \mathsf{H}(\mathscr{A}_{\mathcal{H}})$ with

$$\mathsf{H}(\mathscr{A}_{\mathcal{H}}) = \mathbb{E}_{\mathcal{D},(X',Y')} \left| \ell(\mathcal{A}_{\mathcal{H}}(\mathcal{D})(X_i), Y_i) - \ell(\mathcal{A}_{\mathcal{H}}(\mathcal{D}^i)(X_i), Y_i) \right|, \tag{3}$$

where $\mathcal{D}^i = \mathcal{D} \cup (X', Y') \setminus (X_i, Y_i)$ and (X', Y') are sampled according to μ. Note that $\mathsf{H}(\mathscr{A}_{\mathcal{H}})$ can be estimated from the data [14]. In this setting, it is not automatic, as for the complexity based theories, that increasing the parameters of the model by changing $\mathcal{H}$ (or equivalently to enlarging the corresponding $\mathcal{F}$) will increase $\mathsf{H}(\mathscr{A}_{\mathcal{H}})$ but still for sure will improve our ability to shrink the empirical error $\hat{\mathsf{L}}(\hat{f})$. If a model, as empirical evidence show [6, 9], improve its generalization performance when increasing the number of parameters above the interpolation threshold it means that it actually learns more from data and do no simple fit/memorize the training set. And if this is true, it means that it will not simply memorize the data above that threshold but will actually select more smooth (stable) functions, and then $\mathsf{H}(\mathscr{A}_{\mathcal{H}})$ will decrease above the interpolation threshold. We will further support this intuition in the next section with empirical evidence.

4 Empirical Evidences

Let us consider a quite general framework for shallow models[3] where the functional form of the model is $f(X) = \sum_{i=1}^{p} \alpha_i \sigma_i(X)$ where $\alpha_i \in \mathbb{R}$ and $\sigma_i(X) : \mathcal{X} \to \mathbb{R}\ \forall i \in \{1, \cdots, p\}$. As an algorithm we will rely on regularized least squares[4]

$$\hat{\boldsymbol{\alpha}} : \arg\min_{\alpha \in \mathbb{R}^d} \sum_{i=1}^{n} \left[Y_i - \sum_{i=1}^{p} \alpha_i \sigma_i(X) \right]^2 + \lambda \boldsymbol{\alpha}' M \boldsymbol{\alpha}, \tag{4}$$

where $M \in \mathbb{R}^{p \times p}$ allows to models many type or regularization such as Lp-norm, early stopping, and dropout [10, 17] and $\lambda \in [0, \infty)$ is an hyperparameter.

Let us now consider two examples (similarly to what has been done in [6]).

For the first example[5] we consider a toy regression problem where $\mathcal{X} = [0, 1]$ and $\mathcal{Y}$ is induced by our oracle is $Y = ||X - 0.4| - 0.2| + X/2 - 0.1$. From this oracle, we sample $n = 8$ points randomly from $\mathcal{X}$ to construct $\mathcal{D}$. In this case we consider $f(X) = \sum_{i=0}^{p} \alpha_i X^i$ and $M_{i,j} = 0$ if $i, j \leq 2$, and $M_{i,j} = {}^{i(i-1)j(j-1)}/_{(i+j-3)}$ for $i, j > 2$ since we use as regularizer $\int_0^1 [f''(X)]^2 dX = \boldsymbol{\alpha}' M \boldsymbol{\alpha}$, and $\lambda = 10^{-6}$. Then, in Figure 2, we reported[6]: the oracle, the sample, $\hat{f}$ ($\hat{\boldsymbol{\alpha}}$) for the optimal point in the under- and over-parameterized regimes, and, varying p, the train -

[3] This framework can be easily extended to deep model but for space constraints we restricted our analysis to the shallow models.

[4] Other loss functions could be used but this is out of the scope of the paper.

[5] `https://twitter.com/francoisfleuret/status/1269301689095503872`

[6] Note that, as for [6], we reported the results for a single round since results are quite consistent over repetitions and changes in the parameters

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

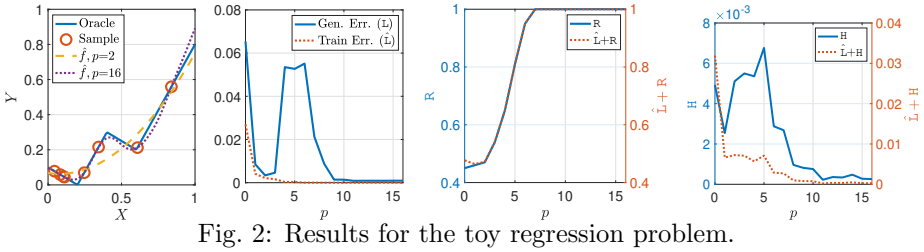

Fig. 2: Results for the toy regression problem.

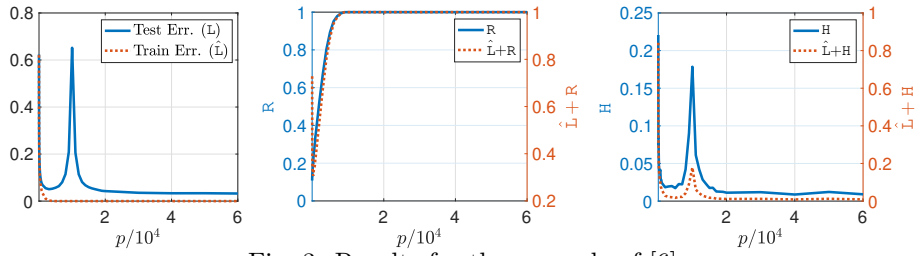

Fig. 3: Results for the example of [6].

$\hat{\mathsf{L}}(\hat{f})$ - and generalization - $\mathsf{L}(\hat{f})$ - errors, the $\mathsf{R}(\mathcal{F})$ [18] with the corresponding Bound (1) ($\hat{\mathsf{L}}(\hat{f})+\mathsf{R}(\mathcal{F})$), and the $\mathsf{H}(\mathscr{A}_{\mathcal{H}})$ [14] with the corresponding Bound (1) ($\hat{\mathsf{L}}(\hat{f})+\mathsf{H}(\mathscr{A}_{\mathcal{H}})$).

From Figure 2 it is possible to confirm the discussion of Section 3. In the second graph one can observe the U-Shaped and Double-Descent curves of the generalization error and the decrease to zero of the train error until the interpolation threshold. In the third graph one can see how the Rademacher Complexity of the class just increases with the increase in the number of parameters while in the fourth graph the Algorithmic Stability is able to detect a change in the behavior of the model which starts to learn again after the interpolation threshold. Using the Rademacher Complexity and the Algorithmic Stability based bound one selects $p = 2$ (the minima in the U-Shaped curve) while $p = 16$ (the second minima after the interpolation threshold) as optimal p respectively (. This result confirms the ability of the Algorithmic Stability of capturing the ability of the over-parameterized model to actually learn and not fit the data and well explain the Double-Descent mechanism.

Let us now consider one of the examples of [6], where the MNIST dataset is considered with $\mathcal{X} \in \mathbb{R}^{28^2}$ and $\mathcal{Y} = \{0, \cdots, 9\}$ sampling $n = 10000$ points randomly. In this case V_i $\forall i \in \{1, \cdots, p\}$ are vectors sampled independently from a uniform distribution over surface of unit hyper-sphere in $\mathbb{R}^{28^2}$, $f(X) = \sum_{i=1}^{p} \alpha_i \max(0, V_i' X)$, $M_{i,j} = 1$ for $i = j$ and $M_{i,j} = 0$ for $i \neq j$, and $\lambda = 10^{-5}$. The multiclass classification problem is mapped into a series of binary classification problems using one-versus-all approach and Problem (4). After the training phase the error on the train and on the test (as estimated of the generalization error) using the percentage of misclassified samples. Figure 3 reports the equivalent of the 2nd, 3rd, and 4th graph of Figure 2 for this new example.

Figure 3 confirms the results and the comments reported for Figure 2 further

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

confirming the statements and the intuitions reported in this work.

5 Conclusions

In this paper we argued that still not well understood mechanisms underlying modern Machine Learning models can be better explained with known theoretical frameworks that can still give insights in the learning process of these new models. In particular, we focused on studying the over-parameterized regime, namely the ability of certain models to increase their generalization performance when the number of parameters is above the interpolating threshold, and the associated double-descent curve. In this context, we show that this mechanism can be better understood using the Algorithmic Stability theory supporting our statement with empirical evidence. This paper is surely just a first step toward rethinking our approaches in studying model Machine Learning going back to milestone theoretical results that can still help better understand modern complex mechanisms but further theoretical analysis and empirical evidence are needed.

References

[1] J. Degrave, F. Felici, J. Buchli, M. Neunert, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[3] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, et al. AI4People - an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.

[4] L. Floridi and J. Cowls. A unified framework of five principles for ai in society. In *Ethics, Governance, and Policies in Artificial Intelligence*, 2021.

[5] T. Poggio, A. Banburski, and Q. Liao. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, 117(48):30039–30045, 2020.

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

[7] A. Wigderson. *Mathematics and computation*. Princeton University Press, 2019.

[8] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. J. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

[9] M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[10] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[11] L. Oneto. *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[12] V. N. Vapnik. *Statistical Learning Theory*. Wiley New York, 1998.

[13] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, 2016.

[14] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE transactions on cybernetics*, 45(9):1913–1926, 2014.

[15] A. Maurer. A second-order look at stability and generalization. In *Conference on learning theory*, 2017.

[16] P. Klesk and M. Korzen. Sets of approximating functions with finite vapnik-chervonenkis dimension for nearest-neighbors algorithms. *Pattern recognition letters*, 32(14):1882–1893, 2011.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[18] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. Structural risk minimization and rademacher complexity for regression. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012.