On a mathematical explanation of double descent in stochastic gradient descent

Matilde Dolfato

Abstract

In this talk, I present a research proposal on the double descent phenomenon observed when training neural networks with stochastic gradient descent (SGD). After briefly reviewing the classical bias–variance tradeoff and its apparent contradiction with modern overparameterized models, I focus on the second descent in test error beyond the interpolation threshold. I highlight the absence of a rigorous mathematical explanation for this behavior. I then discuss existing hypotheses on SGD's implicit bias and empirical results reported in the literature. I identify and motivate the most promising research directions, including a formal explanation of why SGD selects smoother models when interpolating and a precise characterization of the notion of smoothness it implicitly optimizes. I conclude by outlining the main technical obstacles and possible strategies to address them, blending theory and empirical results.

The problem I want to work on is **finding the roots of the *double descent* phenomenon**, concerning the stochastic gradient descent algorithm (SGD) for neural networks.

Background The double descent curve was first formalized by [1], and it reconciles the theoretical result of the bias-variance tradeoff, which implies that capacity below the overfitting threshold is preferred, with the empirical findings that large, over-parametrized models perform well. They explain that when training a neural net using SGD, if we increase capacity beyond the overfitting threshold the algorithm will change the shape of the learned model in the regions between data points, while still fitting training data perfectly. According to [1], SGD interpolates in a smooth way, learning a smoother model that generalizes better. Hence, the curve of test error as a function of capacity presents two descents: the classic one, while learning is happening, and a second one after the overfitting threshold, when an overall improved performance is reached. Therefore, a *double descent* curve substitutes the textbook U-shaped curve.

Research question What is still not clear, however, is what determines this second descent, in the sense that it is observed empirically and only some hypotheses on what motivates it are made. My aim is to explain why the second descent happens with rigorous, mathematical tools. This is relevant because it allows us to understand more in depth what is happening and to have a universal formalization of the properties and dynamics of the algorithm. While this field is highly practice-based, a theoretical framework that supports empirical results boosts the development of AI and further discoveries in this field, especially regarding performance and learning dynamics of neural networks. The problem can be broken down into the following questions, where 1,2 are my main goals and 3, 4 are natural follow-ups, if time allows. I would tackle them first in the setting of simple two-layer neural networks, following [1], and only later extend the analysis to more complex architectures.

1. What drives the implicit bias of SGD to choose smoother functions when interpolating? By analyzing the update rule and the loss landscape mathematically, it should be possible to rigorously state what is happening after the overfitting threshold. There are some hypotheses, like the one proposed by [2], which compares the algorithm used in practice with its continuous analog, i.e. SGD with an infinitesimal step size. By doing so, we get that descending the gradient of the loss L with a discrete step size α is equivalent to performing continuous gradient descent on the modified loss function L_{GD},

$$L_{GD}[\phi] := L[\phi] + \frac{\alpha}{4}\left\|\frac{\partial L}{\partial \phi}\right\|^2$$

where a term depending on the norm of the gradient of the loss appears. This motivates the bias for less steep directions in the loss landscape. This result, however, requires to understand the connection between the gradient of the loss and a smoother learned model and it can be enriched with additional explanations and intuition.

2. Which specific notion of *smoothness* does SGD optimize for? I would like to be able to state something on the lines of "SGD has an implicit bias for learning smoother models, in the sense that it chooses functions with a lower gradient norm / lower total variation of the gradient / others". In this direction, [3] tests four different notions of smoothness and finds that SGD optimizes for second-order smoothness. While it is a good starting point, I want to identify this notion of smoothness with necessary mathematical implications, rather than deducing it from empirical results.

3. If the underlying data distribution is not *smooth* (according to this notion), does SGD still interpolate in a smoother way? And do we still see the second descent? The first question bridges previous directions with the following ones, and I believe that the answer is yes, based on the hypothesis that the smooth interpolation is due to an implicit bias that is an intrinsic property of the algorithm. The second question is very interesting to me and aims at further understanding the phenomenon. My hypothesis is that the answer is yes, because the *smoothness* we are talking about only concerns areas of interpolation between data points, hence if we have a reasonable number of points (a not excessively sparse space), then these areas will be *smooth* independently of the smoothness of the underlying function. Hence, we could try to understand what the algorithm does when a small training set is given and we have greater sparsity.

4. Why does a smoother model generalize better? In the literature ([1] and [2] among others), this is motivated by Occam's razor, saying that the simplest explanation compatible with the observations shall be preferred. This is specially suitable for the learning problems we generally consider, based on real-world data, according to [1]. But does this have a mathematical motivation? Although I find this question extremely stimulating, I believe it should not be a priority, as the explanation provided by Occam's razor seems widely accepted.

Obstacles The main obstacles I see as of now are two: the wide variety of the combinations of the many design specifications, like initialization and step size, which hinders a universal explanation of the phenomenon; and working in high-dimensional spaces, where it is not clear how smoothness evolves and what happens in the optimization landscape. As for the former, I believe that starting from empirical results can be useful, as these may help discern which theoretical directions to inquire first. For instance, if the second descent changes with different step sizes, ceteris paribus, that would indicate that the step size plays a role. Regarding dimensionality, I believe that we should study the problem with a mathematical mindset and try to generalize from a two-dimensional loss landscape, where the dynamics can be visualized, to a higher-dimensional space relying on tools from functional analysis and geometry (the specific method is not clear to me yet).

Overall, I find it interesting to inquire the mathematical roots of the *double descent* phenomenon, as the optimization algorithm, the loss landscape, and the learned model are all "just math", hence it should be possible to formalize what is happening. Also, it is relevant, because it would enable us to bolster scientific discovery in the area of neural networks and learning.

References

[1] 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, 2019.

[2] Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

[3] Václav Volhejn and Christoph Lampert. Does sgd implicitly optimize for smoothness? In *DAGM German Conference on Pattern Recognition*, pages 246–259. Springer, 2020.