The Annals of Statistics
2020, Vol. 48, No. 4, 1916–1921
https://doi.org/10.1214/19-AOS1931
Main article: https://doi.org/10.1214/19-AOS1875

REJOINDER: "NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION"

BY JOHANNES SCHMIDT-HIEBER

Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl

The author is very grateful to the discussants for sharing their viewpoints on the article. The discussant contributions highlight the gaps in the theoretical understanding and outline many possible directions for future research in this area. The rejoinder is structured according to topics. We refer to [GMMM], [K], [KL] and [S] for the discussant contributions by Ghorbani et al., Kutyniok, Kohler & Langer and Shamir, respectively.

1. Overparametrization and implicit regularization. One of the general claims about deep learning is that, even for extreme overfitting, the method still generalizes well. There are numerous experiments showing that running the training error to zero and, therefore, interpolating all data points results in state-of-the-art generalization performance. The rationale behind this is that among all solutions interpolating the data points, of which most result in bad generalization behavior, stochastic gradient descent (SGD) picks a minimum norm interpolant. This is also known as implicit regularization. While this is well known for stochastic gradient descent applied to linear regression, for deep networks some progress has been made recently in finding the norm minimized by (S)GD; see [10, 23].

It is now reasonable to wonder whether the notion of network sparsity could be removed in the article if implicit regularization would have been taken into account. [GMMM] write that "Model complexity is not controlled by an explicit penalty or procedure, but by the dynamics of stochastic gradient descent (SGD) itself." [S] mentions implicit regularization to show that statistical guarantees should involve specific learning methods.

We conjecture that for additive error models, such as the nonparametric regression model considered in the article, implicit regularization in the overfitted regime is insufficient to achieve even consistency. To support our conjecture, we provide the following two-step argument. In the first step we argue that for one-dimensional input and shallow networks with fixed parameters in the first layer, SGD will converge to a variant of the natural cubic spline interpolant. In the second step we show that this reconstruction leads to an inconsistent estimator if additive noise is present.

A shallow ReLU network with one input and one output node can be written as $x \mapsto \sum_{j=1}^m a_j(b_j x - c_j)_+$. We now study an even more simplified setup where b_j is always one. For small $\delta > 0$, $(x - c_j)_+ \approx \int_{c_j}^{c_j+\delta} (x-u)_+ \, du/\delta$. This motivates to study smoothed shallow ReLU networks of the form

$$x \mapsto f_{\mathbf{a}}(x) = \sum_{j=1}^m \frac{a_j}{\sqrt{t_j - t_{j-1}}} \int_{t_{j-1}}^{t_j} (x-u)_+ \, du$$

with parameter vector $\mathbf{a} = (a_1, \dots, a_m)$ and fixed $t_0 < t_1 < \cdots < t_m$. For convenience, we have rescaled the parameters a_j so that the normalization factor becomes $1/\sqrt{t_j - t_{j-1}}$. We consider the overparametrized regime $m \geq n$ assuming that, for any i, there lies at least one t_j in the interval $[X_{(i-1)}, X_{(i)})$ with $X_{(i)}$ the ith order statistic of the sample $X_1, \dots, X_n$ and $X_{(0)} = -\infty$. Under overparametrization this is a rather weak assumption and ensures

Received November 2019.