The Annals of Statistics
2020, Vol. 48, No. 4, 1906–1910
https://doi.org/10.1214/19-AOS1912
Main article: https://doi.org/10.1214/19-AOS1875

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

BY MICHAEL KOHLER* AND SOPHIE LANGER†

*Fachbereich Mathematik, TU Darmstadt, *kohler@mathematik.tu-darmstadt.de; †langer@mathematik.tu-darmstadt.de*

First we would like to congratulate Professor Johannes Schmidt-Hieber for his excellent paper, which shows the surprising result that deep neural networks can achieve good rates of convergence even in case of nonsmooth activation functions.

In the following we divide our discussion into three parts:

1. The importance of compository assumptions.
2. The necessity of the sparsity of the networks.
3. The theoretical difference between ReLU and sigmoidal functions.

1. The importance of compository assumptions. In the sequel we use the following definition of (p, C)-smoothness.

DEFINITION 1. Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth if, for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$, the partial derivative $\partial^q m / (\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d})$ exists and satisfies

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$, where $\| \cdot \|$ denotes the Euclidean norm.

Remark that this assumption on the regression function is similar to the class $\mathcal{C}_r^\beta(D, K)$ of functions mentioned in Section 3 in the paper under discussion. It is well known that the optimal rate of convergence for the estimation of a (p, C)-smooth regression function is

$$n^{-\frac{2p}{2p+d}}.$$

In case d is relatively large compared to p, this rate suffers from the well-known curse of dimensionality. The only way to circumvent this phenomenon is to impose additional assumptions on the regression function. One way is to impose compository assumptions, which were already used by Horowitz and Mammen (2007), where regression functions have been studied which are of the form

$$m(x) = g\left(\sum_{l_1=1}^{L_1} g_{l_1} \left(\sum_{l_2=1}^{L_2} g_{l_1,l_2} \left(\cdots \sum_{l_r=1}^{L_r} g_{l_1,\ldots,l_r}\left(x^{l_1,\ldots,l_r}\right) \right) \right) \right)$$

for $g, g_{l_1}, \ldots, g_{l_1,\ldots,l_r} : \mathbb{R} \to \mathbb{R}$ (p, C)-smooth functions and $x^{l_1,\ldots,l_r}$ single components of $x \in \mathbb{R}^d$ (not necessarily different for two different indices $(l_1, \ldots, l_r)$). With the use of a penalized least squares estimate for smoothing splines, they proved the rate $n^{-2p/(2p+1)}$. Kohler and Krzyżak (2017) extended this function class in form of the so-called generalized hierarchical interaction models introduced as follows:

Received August 2019; revised September 2019.