Submitted to the Annals of Statistics

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

By Michael Kohler and Sophie Langer

Technische Universität Darmstadt

First we would like to congratulate Prof. 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 non-smooth 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 that d is relatively large compared to p this rate suffers from the well–known curse of dimensionality. The only way to circumenvent this phenomenon is to impose additional assumptions on the regression function. One way offer compository assumptions, which were already used by Horowitz

and Mammen (2007), where regression functions have been studied which are of the form

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

for $g, g_{l_1}, \dots, g_{l_1,\dots,l_r} : \mathbb{R} \to \mathbb{R}$ (p, C)-smooth functions and $x^{l_1,\dots,l_r}$ single components of $x \in \mathbb{R}^d$ (not necessarily different for two different indices $(l_1, \dots, l_r)$). 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 so–called generalized hierarchical interaction models introduced as follows:

DEFINITION 2. *Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $m : \mathbb{R}^d \to \mathbb{R}$.*
***a)** We say that m satisfies a **generalized hierarchical interaction model of order d^* and level 0**, if there exist $a_1, \dots, a_{d^*} \in \mathbb{R}^d$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that*

$$m(x) = f(a_1^T x, \dots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d.$$

***b)** We say that m satisfies a **generalized hierarchical interaction model of order d^* and level $l+1$**, if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$ $(k = 1, \dots, K)$ and $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, K)$ such that $f_{1,k}, \dots, f_{d^*,k}$ $(k = 1, \dots, K)$ satisfy a generalized hierarchical interaction model of order d^* and level l and*

$$m(x) = \sum_{k=1}^{K} g_k\left(f_{1,k}(x), \dots, f_{d^*,k}(x)\right) \quad \text{for all } x \in \mathbb{R}^d.$$

***c)** We say that the **generalized hierarchical interaction model** defined above **is (p, C)-smooth**, if all functions f and g_k occurring in its definition are (p, C)–smooth.*

They showed that for such models suitably defined multilayer neural networks (in which the number of hidden layers depends on the level of the generalized interaction model) achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) in case $p \leq 1$. Bauer and Kohler (2019) showed that this result even holds for $p > 1$ provided the sigmoidal function is suitably chosen.

In case that the number of terms in the sum in part b) of the above summation is chosen to be $K = 1$ for all levels and that the vectors a_1, $\dots$, a_{d^*} in part a) are chosen as unit vectors, the corresponding function

is recursively defined as a function of d^* variables, where all variables are either a function of the same kind or one of the components of the input variable (here it is allowed that the same component appears several times). In practice, it is conceivable, that there exist input–output–relationships, which can be described in this way with a small to moderate value of d^*. Particulary, such an assumption is motivated by applications in connection with complex technical systems, which are constructed in a modular form. Here each modular part can be again a complex system, which also explains the recursive construction in the above definition.

The function class studied by Prof. Schmidt–Hieber forms some generalization of Definition 2 in a sense that smoothness and dimension of the g_k in different levels in the recursive construction are allowed to be different. This can be generalized one step further by allowing smoothness and dimension to change within each level:

DEFINITION 3. *Let $d \in \mathbb{N}$ and $m : \mathbb{R}^d \to \mathbb{R}$.*
a) *We say that m satisfies a* ***hierarchical composition model of level*** *0, if there exists a $K \in \{1, \dots, d\}$ such that*

$$m(x) = x^{(K)} \quad \text{for all } x = (x^{(1)}, \dots, x^{(d)})^T \in \mathbb{R}^d.$$

b) *We say that m satisfies a* ***hierarchical composition model of level*** *$l+1$, if there exist $K \in \mathbb{N}$, $g : \mathbb{R}^K \to \mathbb{R}$ and $f_1, \dots, f_K : \mathbb{R}^d \to \mathbb{R}$, such that $f_1, \dots, f_K$ satisfy a hierarchical composition model of level l and*

$$m(x) = g(f_1(x), \dots, f_K(x)) \quad \text{for all } x \in \mathbb{R}^d.$$

c) *We say that a hierarchical composition model satisfies the smoothness and order constraint $\mathcal{P}$, where $\mathcal{P}$ is a subset of $(0, \infty) \times \mathbb{N}$, if in its definition all functions g occuring in part b) satisfy $g : \mathbb{R}^K \to \mathbb{R}$ and g (p, C)–smooth for some $(p, K) \in \mathcal{P}$ and $C > 0$.*

In case $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ a suitably defined least squares neural network regression estimate achieves (up to some logarithmic factor) the rate of convergence

$$\max_{(p,K) \in \mathcal{P}} n^{-\frac{2p}{2p+K}}$$

(cf., e.g., Theorem 1 below). We would like to point out that Definition 3 (which is basically a (slight) generalization of the assumption used in the paper of Prof. Schmidt-Hieber) is a valuable extension of Definition 2 (which was introduced Kohler and Krzyżak (2017)), because it seems to be even more realistic for the applications described above.

2. The necessity of the sparsity of the networks. One of the key features of the neural networks in the paper under discussion is, that the considered neural networks are not fully connected. We would like to point out that this is not necessary required, since similar results also hold for fully connected deep neural network, as the next theorem shows.

THEOREM 1. *Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that supp(X) is bounded and*

$$\mathbf{E}\left\{\exp(c_1 \cdot Y^2)\right\} < \infty \tag{1}$$

for some constant $c_1 > 0$. Let $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ be such that

$$p_{max} = \max_{(p,K)\in\mathcal{P}} p < \infty \quad \text{and} \quad \max_{(p,K)\in\mathcal{P}} K < \infty.$$

Assume that the regression function $m(\cdot) = \mathbf{E}\{Y|X = \cdot\}$ satisfies a hierarchical composition model of finite level l and with smoothness and order constraint $\mathcal{P}$. Set

$$L_n = \lceil c_2 \cdot \max_{(p,K)\in\mathcal{P}} n^{\frac{K}{2\cdot(2p+K)}} \rceil \quad \text{and} \quad r_n = c_3$$

for $c_2, c_3 > 0$ sufficiently large. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the linear rectifier. Let $\mathcal{F}_\sigma(L_n, r_n)$ be the set of all fully connected neural networks with L_n hidden layers, r_n neurons in each hidden layer and σ as actication function. Let $\tilde{m}_n$ be the least squares estimate defined by

$$\tilde{m}_n(\cdot) = \arg \min_{h\in\mathcal{F}_\sigma(L_n,r_n)} \frac{1}{n} \sum_{i=1}^{n} |Y_i - h(X_i)|^2 \tag{2}$$

and define $m_n = T_{c_4\cdot\log(n)}\tilde{m}_n$ for some $c_4 > 0$ sufficiently large, where $T_\beta z = \max\{\min\{z, \beta\}, -\beta\}$ for $z \in \mathbb{R}$ and $\beta > 0$. Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_5 \cdot (\log n)^4 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{2p}{2p+K}}$$

holds for sufficiently large n.

PROOF. See Kohler and Langer (2019). □

A comparison with Theorem 1 in the paper under discussion shows that we can reach the same convergence rate also with simple fully connected networks. Here the topology of our networks is completely specified, which makes an implementation of a corresponding estimate much easier.

3. The theoretical difference between ReLU and sigmoidal functions. The paper of Prof. Schmidt-Hieber focusses on the ReLU activation function, which is nowadays quite popular in applications. One useful characteristic of this kind of function is, that their derivatives are always either 0 or 1. Consequently, the derivative of the neural network can be computed much faster in an application and the backpropagation algorithm can be applied with a much large number of gradient descent steps for the linear rectifier (cf., e.g., Fan, Ma and Zhong (2019)). However, theoretically we cannot see much of a difference in comparison to sigmoidal activation functions, due to the following approximation result:

LEMMA 1. *Let $\sigma : \mathbb{R} \to [0,1]$ be 2-admissible, i.e., assume that σ is nondecreasing and Lipschitz continous and that, in addition, the following three conditions are satisfied:*

(i) The function σ is three times continously differentiable with bounded derivatives.

(ii) A point $t_\sigma \in \mathbb{R}$ exists, where all derivatives up to the order 2 of σ are different from zero.

(iii) If $y > 0$, the relation $|\sigma(y) - 1| \leq \frac{1}{y}$ holds. If $y < 0$, the relation $|\sigma(y)| \leq \frac{1}{|y|}$ holds.

Then for any $\epsilon \in (0,1]$ and $a \geq \max\{1, \frac{3}{\epsilon}\}$ a neural network

$$f_{ReLU}(x) = \sum_{k=1}^{6} d_k \cdot \sigma\left(\sum_{i=1}^{2} b_{k,i} \cdot \sigma(a_i \cdot x + t_\sigma) + b_{k,3} \cdot \sigma(a_3 \cdot x) + t_\sigma\right)$$

exists such that

$$|f_{ReLU}(x) - \max\{x, 0\}| \leq \epsilon$$

holds for all $x \in [-a, a]$. The coefficients of this network satisfy

$$|a_i| \leq \frac{3}{\epsilon}, \quad |b_{k,i}| \leq \frac{c_{20}}{a} \quad \text{and} \quad |d_k| \leq c_{21} \cdot \frac{a^6}{\epsilon^2}$$

for $i \in \{1, \ldots, 3\}$, $k \in \{1, \ldots, 6\}$.

Proof. Let $f_{id}(x)$ and f_{mult} be the networks of Lemma 1 and Lemma 3 in Kohler, Krzyżak and Langer (2019) which satisfy

$$|f_{id}(x) - x| \leq \frac{\epsilon}{3} \quad \text{for } x \in [-a, a]$$

and

$$|f_{mult}(x, y) - x \cdot y| \leq \frac{\epsilon}{3} \quad \text{for } x, y \in [-2a, 2a].$$

Then

$$
\begin{aligned}
&= |f_{mult}\left(f_{id}(x), \sigma\left(\frac{3}{\epsilon}\cdot x\right)\right) - \max\{x, 0\}| \\
&\leq |f_{mult}\left(f_{id}(x), \sigma\left(\frac{3}{\epsilon}\cdot x\right)\right) - f_{id}(x)\cdot\sigma\left(\frac{3}{\epsilon}\cdot x\right)| \\
&\quad + |f_{id}(x)\cdot\sigma\left(\frac{3}{\epsilon}\cdot x\right) - x\cdot\sigma\left(\frac{3}{\epsilon}\cdot x\right)| + |x\cdot\sigma\left(\frac{3}{\epsilon}\cdot x\right) - x\cdot\mathbb{1}_{[0,\infty)}(x)| \\
&\leq \frac{\epsilon}{3} + \frac{\epsilon}{3}\cdot 1 + \frac{\epsilon}{3} = \epsilon. \qquad \square
\end{aligned}
$$

Using this lemma it is be possible to approximate any neural network with ReLU activation function by a neural network with sigmoidal activation function. However, in contrast to the networks in the paper by Prof. Schmidt-Hieber the weights will no longer be bounded in absolute value by one. This might be considered as a drawback, but from a theoretical point of view we do not know any result indicating that least squares neural network regression estimates with small weights achieve a better rate of convergence than neural networks with large weights (as long as the absolute values of the weights do increase at most like a polynomial in the sample size).

References.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* **47** 2261–2285.

Fan, J., Ma, C. and Zhong, Y. (2019). A Selective Overview of Deep Learning. *CoRR* **abs/1904.05526**.

Horowitz, J. L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics* **35** 2589–2619.

Kohler, M. and Krzyżak, A. (2017). Nonparametric Regression Based on Hierarchical Interaction Models. *IEEE Transactions on Information Theory* **63** 1620–1630.

Kohler, M., Krzyżak, A. and Langer, S. (2019). Deep learning and MARS: a connection. arXiv: 1908.11140.

Kohler, M. and Langer, S. (2019). On the rate of convergence of fully connected very deep neural network regression estimates using ReLU activation functions. arXiv: 1908.11133.

Michael Kohler
Fachbereich Mathematik
TU Darmstadt
Schlossgartenstr. 7
64289 Darmstadt, Germany
E-mail: kohler@mathematik.tu-darmstadt.de

Sophie Langer
Fachbereich Mathematik
TU Darmstadt
Schlossgartenstr. 7
64289 Darmstadt, Germany
E-mail: langer@mathematik.tu-darmstadt..de