The Annals of Statistics
2020, Vol. 48, No. 4, 1902–1905
https://doi.org/10.1214/19-AOS1911
Main article: https://doi.org/10.1214/19-AOS1875

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

BY GITTA KUTYNIOK

Institute of Mathematics, Technische Universität Berlin, kutyniok@math.tu-berlin.de

I would like to congratulate Johannes Schmidt–Hieber on a very interesting paper in which he considers regression functions belonging to the class of so-called compositional functions and analyzes the ability of estimators based on the multivariate nonparametric regression model of deep neural networks to achieve minimax rates of convergence.

In my discussion, I will first regard such a type of result from the general viewpoint of the theoretical foundations of deep neural networks. This will be followed by a discussion from the viewpoint of expressivity, optimization and generalization. Finally, I will consider some specific aspects of the main result.

I will start my discussion by setting the results by Schmidt–Hieber into context. Neural networks have lately shown tremendous performance in a variety of real-world applications and are already starting to impact public life in various aspects such as in the health care sector. However, despite the outstanding success of deep neural networks, most of the related research is empirically driven, and a deep theoretical understanding is of great demand, in particular, for sensitive applications.

Deep learning can be regarded as a statistical learning problem, whose empirical risk consists of three parts, namely: the approximation error related to the hypothesis class, the optimization error from the optimization procedure as well as the out-of-sample error related to the capability of predicting the outcome of unseen samples correctly. Consequently, also the theoretical analysis of deep neural networks can be separated—certainly with various interconnections—into three parts as well: expressivity, optimization and generalization.

The work by Schmidt–Hieber focusses on expressivity and generalization, and excludes the optimization aspects such as convergence of the training algorithm or occurring errors due to a wrong choice of the initial values. In his work he focusses on the regression problem, assuming the regression function to belong to the class of compositional functions, and as hypothesis class he considers sparsely connected deep neural networks.

The area of expressivity is, currently, perhaps the furthest developed of the three directions. By combining expressivity and generalization of deep neural networks, the work by Schmidt–Hieber can be considered an important contribution. Now, due to the rapid evolvement of the theory of deep learning, other intriguing approaches aiming to provide a comprehensive framework combining all three error components were also developed, such as [6], in which the authors regard neural networks as interacting particle systems.

In the sequel, I will now discuss several aspects of the paper in more detail.

1. Expressivity: The choice of models. As a model for the regression function, Schmidt–Hieber considers the class of compositorial functions, which are characterized by the fact that they are a composition of functions, each of which just depends on a very small

Received September 2019.
MSC2020 subject classifications. 62G08.
Key words and phrases. Deep neural networks, generalization, nonparametric regression.

number of variables. One should positively stress that this model is mathematically independent of the choice of models for neural networks, namely, sparsely connected ones, in the sense that the characteristics such as the number of compositions q or the number t_i of variables of the functions $g_{i,j}$ are not linked to characteristics of the network class such a depth L, sparsity bound s, etc. This provides a quite general framework.

However, Schmidt–Hieber states that one argument for choosing the class of compositorial functions is that it is "natural for neural networks." I agree that choosing a proper function class is highly difficult, and compositorial functions were used before for approximation results of neural networks, for example, in [4]. From my viewpoint, the choice of the model should, however, not depend on (or be "natural for") the approximation system, here the neural network but vice versa. In fact, one key question to my mind is to identify realistic function classes for which neural networks do still perform very well.

This is closely related to the realm of the curse of dimensionality. The fact that deep neural networks are that effective in high-dimensional data settings is usually attributed to their ability to beat the curse of dimensionality. This has been already proven in several instances, in particular, in the situation of partial differential equations (see, e.g., [1–3]). It would be interesting to discuss this aspect also in the setting considered by Schmidt–Hieber. From this viewpoint, the considered model of the class of compositorial functions might, however, not be the best choice due to the fact that the low-dimensional structure is already easily accessible, not allowing for a deep insight in the ability of deep neural networks to circumvent the curse of dimensionality.

At this point, let me refer to the useful overview by Schmidt–Hieber of the current state of the art of approximation theory for neural networks in Section 6 of his paper, which also includes a focus on ReLU neural networks. I would like to add [5] to this list. To my mind, it would have been nice to include and discuss this reference, since it provides, in fact, optimal approximation results for piecewise smooth functions by ReLU neural networks, analyzing also their ability to avoid the curse of dimensionality.

2. Optimization: The role of training. For the practitioner it is of key importance to derive a profound understanding of the algorithmic aspects of the training, namely, the optimization part. Schmidt–Hieber does not analyze optimization in detail but instead uses the term $\Delta_n(\hat{f}_n, f_0)$, thereby also allowing that the optimization problem is not solved exactly, but sufficiently precisely. Although it would certainly be highly desirable to ultimately have a theory available which encompasses all three components, that is, expressivity, optimization and generalization, to my mind the results by Schmidt–Hieber are already a fundamental contribution and an excellent basis for continuation.

However, I would still like to allow myself to raise constructive criticism and provide comments on two aspects. For this, I will first briefly review the key components of training a neural network. A neural network in its vanilla form is a highly structured function $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form

$$f(\mathbf{x}) = W_L \sigma_{\mathbf{v}_L} W_{L-1} \sigma_{\mathbf{v}_{L-1}} \cdot \ldots \cdot W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x},$$

where W_i, $i = 0, \ldots, L$ are $p_{i+1} \times p_i$ matrices and $\sigma_{\mathbf{v}_i}$, $i = 0, \ldots, L$ is the ReLU function shifted by the vector $\mathbf{v}_i \in \mathbb{R}^{p_i}$ componentwise. Given training data $(\mathbf{X}_i, Y_i)_{i=1}^n$, optimization strategies such as stochastic gradient descent (SGD) intend to minimize the empirical risk, that is, solve

$$\min_{(W_j, \mathbf{v}_j)_{j=0}^L} \sum_{i=1}^n \mathcal{L}\big(f_{(W_j, \mathbf{v}_j)_{j=0}^L}(\mathbf{X}_i), Y_i\big) + \lambda \mathcal{R}\big((W_j, \mathbf{v}_j)_{j=0}^L\big), \tag{2.1}$$

where $f_{(W_j,\mathbf{v}_j)_{j=0}^L}$ is a deep neural network, $\mathcal{L}$ a loss function, $\mathcal{R}$ a regularizer and λ the regularization parameter. SGD does not compute the gradient for each pair of data $(\mathbf{X}_i, Y_i)$ but, in each iteration, selects one of those indices randomly to reduce the computational complexity. A variant of SGD is to select a so-called batch instead of selecting just one sample.

Closely analyzing the definition of $\Delta_n(\hat{f}_n, f_0)$ in equation (5), one realizes that the term $\hat{f}_n$ is, in fact, a stochastic term. This is due to the fact that the common strategy of SGD consists of a random selection step in each iteration. Therefore, it would be interesting—and from a practical viewpoint necessary—to understand the expected value with respect to a suitable distribution incorporating also the selections of the batches which is not the case in the definition in (5).

The second point I would like to raise is the fact that, in practise, the empirical risk minimization typically contains a regularization term as in (2.1). Moreover, the quadratic loss as the loss function is presumably the most often selected one, but other loss functions, such as the hinge loss, are also frequently used. It might be that the results are not too difficult to transfer to more general loss functions as well as to regularization which would be closer to the training typically performed in applications.

3. Generalization: Role of neural networks. Schmidt–Hieber considers the prediction error in his analysis, for which he provides upper and/or lower bounds under some conditions. One key question in the analysis of deep neural networks is why this particular model generalizes this well. In fact, although the optimization problem does not possess a unique global minimizer in general and instead there might be (infinitely) many solutions and spurious local minima, neural networks trained by SGD generalize extremely well in the over-parameterized regime.

The fact that classical methods do not explain the success concerning the generalization ability of deep neural networks sufficiently well, has and still is the point of an intense discussion; see, for example, [7]. Schmidt–Hieber's work contributes to that extent that he, in particular, shows an upper bound for the generalization error depending on the depth L of a neural network, the number of training samples as $\log^2 n$ and the term ϕ_n, depending on the chosen class of compositorial functions. In fact, this connection gives very useful indications on the impact of those components on the generalization ability of a neural network. In light of a deeper understanding of generalization, it would be very interesting to shed more light on the term $\Delta_n(\hat{f}_n, f_0)$ which depends on the optimization scheme/algorithm as well as the function class.

4. Theorem 1. Let me end with two comments on Theorem 1 which is the main theorem of Schmidt–Hieber. Condition (iii) requires the minimal width of the layers to be an upper bound of $n\phi_n$. This, in particular, implies that, if the smoothnesses β_i^* are small, which one usually encounters in applications, the number t_i of variables of the functions $g_{i,j}$ has to be very small to compensate for the factor n, which could be in the range of millions, whereas the number of neurons per layer is usually less or in the order of the input dimension. Thus, the key question is to which extent this condition can be relaxed to encompass compositorial functions composed of higher dimensional components in typical network settings.

The second comment relates to the conditions on $\Delta_n(\hat{f}_n, f_0)$, depending on which different estimates for $\mathcal{R}(\hat{f}_n, f_0)$ are derived. It would be interesting to analyze numerically which situation occurs and when. This would also provide additional insight into the generalization question, as discussed in Section 3.

Acknowledgments. Also affiliated with Machine Learning Group, University of Tromsø.

REFERENCES

[1] BECK, C., HORNUNG, F., HUTZENTHALER, M., JENTZEN, A. and KRUSE, T. (2019). Overcoming the curse of dimensionality in the numerical approximation of Allen–Cahn partial differential equations via truncated full-history recursive multilevel Picard approximations. arXiv:1907.06729.

[2] GROHS, P., HORNUNG, F., JENTZEN, A. and VON WURSTEMBERGER, P. (2018). A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. arXiv:1809.02362.

[3] KUTYNIOK, G., PETERSEN, P., RASLAN, M. and SCHNEIDER, R. (2019). A theoretical analysis of deep neural networks and parametric PDEs. arXiv:1904.00377.

[4] MHASKAR, H. N. and POGGIO, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Anal. Appl.* (*Singap.*) **14** 829–848. MR3564936 https://doi.org/10.1142/S0219530516400042

[5] PETERSEN, P. and VOIGTLAENDER, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.* **108** 296–330. https://doi.org/10.1016/j.neunet.2018.08.019

[6] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. arXiv:1805.00915.

[7] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2017). Understanding deep learning requires rethinking generalization. In *ICLR* 2017.