The Annals of Statistics
2020, Vol. 48, No. 4, 1898–1901
https://doi.org/10.1214/19-AOS1910
Main article: https://doi.org/10.1214/19-AOS1875

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

BY BEHROOZ GHORBANI[1], SONG MEI[2], THEODOR MISIAKIEWICZ[3] AND ANDREA MONTANARI[4]

[1]*Department of Electrical Engineering, Stanford University, ghorbani@stanford.edu*
[2]*Institute for Computational and Mathematical Engineering, Stanford University, songmei@stanford.edu*
[3]*Department of Statistics, Stanford University, misiakie@stanford.edu*
[4]*Department of Electrical Engineering and Department of Statistics, Stanford University, montanari@stanford.edu*

We congratulate Johannes Schmidt-Hieber for his elegant and thought-provoking results. His article uses deep-learning-inspired methods in the context of nonparametric regression. Schmidt-Hieber defines a rich class of composition-based functions $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ and a class of sparse multilayer neural networks $\mathcal{F}(L, \mathbf{p}, s, F)$. He proves that least squares estimation over the class of sparse neural networks (with suitably chosen architecture $(L, \mathbf{p}, s, F)$) achieves nearly minimax prediction error over $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$.

The modeling and analysis in this paper are both elegant and original. They trigger a natural question: *how much of the empirical success of deep learning can be understood using this model?* As a way to stimulate reflection on this question, we will discuss three challenges: *1. Sparsity and generalization*; *2. Curse of dimensionality*; *3. Computation*.

Throughout, we will denote by $\varepsilon^* = \min_{0\le i\le q}[2\beta_i^*/(2\beta_i^* + t_i)] \in (0, 1)$ the minimax exponent in the class $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$. Also, in our discussion we shall focus on multilayer perceptrons, and, in particular, we exclude convolutional networks. The latter have entirely different structure, and they do not follow within the scope of the present paper.

1. Sparsity and generalization. Modern multilayer neural networks are highly overparametrized. Schmidt-Hieber uses sparsity of the weights as a gauge to control the model's complexity and, hence, to be able to bound the generalization error using tools from empirical process theory.

Is sparsity the right complexity measure in practical deep-learning methods? The present paper requires the number of nonzero weights to be $s \asymp n^{1-\varepsilon^*} \log n$. As an example, consider the VGG-19 architecture [13] which is a state-of-the art deep network trained on ImageNet.[1] This network has approximately $143 \cdot 10^6$ parameters, of which $123 \cdot 10^6$ are in the fully-connected layers. Figure 1 reports the distribution of these weights after training: we are not able to identify any sparsity structure. Notice that—for ImageNet—the sample size is roughly $n \approx 1.2 \cdot 10^6$, hence, much smaller than the number of nonzero coefficients.

The nascent research community in theoretical deep learning is well aware of the fact that some measure of complexity is necessarily controlled by overparametrized neural networks. A popular heuristic explanation uses the notion of of "implicit regularization": model complexity is not controlled by an explicit penalty or procedure but by the dynamics of stochastic gradient descent (SGD) itself [12]. Defining the precise complexity measure that is implicitly controlled by SGD is an open problem, except in some simple examples [7, 9, 14]. A parallel line of work directly analyzes gradient descent and shows that the generalization error can be controlled even in the presence of infinitely overparametrized networks, as long as gradient descent is stopped early [3, 10].

Received September 2019.
[1]The trained parameters were downloaded from Keras 2.2.4.