The Annals of Statistics
2020, Vol. 48, No. 4, 1911–1915
https://doi.org/10.1214/19-AOS1915
Main article: https://doi.org/10.1214/19-AOS1875

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

BY OHAD SHAMIR

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, ohad.shamir@weizmann.ac.il

I would like to commend Johannes Schmidt-Hieber for a very interesting and timely paper which studies nonparametric regression using deep neural networks. In recent years, the area of deep learning has seen an explosive growth within machine learning, leading to impressive leaps in performance across a wide range of important applications. However, our theoretical understanding of deep learning systems is still very limited, with many unresolved questions about their computational tractability and statistical performance. I believe that the statistics community can play a crucial role in tackling these challenging questions and hope that Schmidt-Hieber's paper will spur additional research. Being a computer scientist rather than a statistician, I am happy for the opportunity to provide an "outsider's" viewpoint on this paper (of course, any opinions expressed are solely my own).

Curse of dimensionality, or curse of sparsity? The paper studies a nonparametric regression model of the form

$$\mathbf{Y}_i = f_0(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i are i.i.d. standard normal, $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^n$ are i.i.d. observations and f_0 is a function with Hölder smoothness properties. Without additional assumptions, this problem suffers from the well-known "curse of dimensionality," with the sample size n required to approximate f_0 scaling exponentially with the dimension d. As a result, such error rates are meaningful only in low-dimensional settings and cannot explain the success of high-dimensional methods such as deep neural networks. To tackle this, Schmidt-Hieber imposes an additional structural constraint on f_0, namely, that it can be written as a composition of smooth vector-valued functions

$$f_0 = g_q \circ g_{q-1} \circ \cdots \circ g_1 \circ g_0,$$

where each $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ is generally sparse and depends on only $t_i \le d_i$ input variables. The main result in the paper is that the convergence rate of this method (using deep neural networks as estimators) is governed by the quantity

$$\phi_n := \max_{i=0,\dots,q} n^{-\frac{2\beta_i^*}{2\beta_i^*+t_i}}, \tag{1}$$

where β_i^* quantifies the smoothness of each g_i. Crucially, the rate is no longer explicitly dependent on the input dimension, as the networks are able to adapt to the internal sparsity in f_0. In contrast, the paper shows that a standard nonparametric estimator, namely wavelet estimators with uniform design, cannot take advantage of this structure. Even for the special case of nonparametric additive models ($f_0(\mathbf{x}) = h(x_1 + \cdots + x_d)$ for a smooth univariate h), the convergence rate of this estimator is no better than

$$n^{-\frac{2\alpha}{2\alpha+d}}, \tag{2}$$

Received July 2019; revised September 2019.