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.

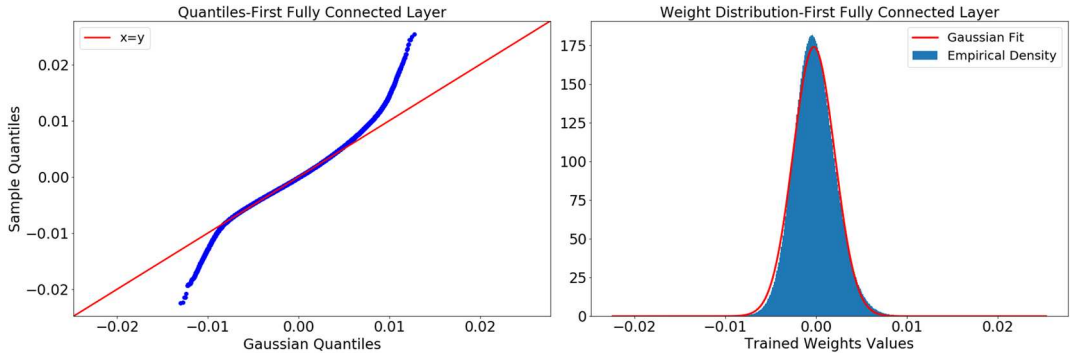

FIG. 1. *First fully-connected layer of a VGG*-19 *network after training. Left frame*: *Sample quantiles plotted against quantiles of a Gaussian with first two moments matching the empirical moments. Right frame*: *Empirical density of the weights plotted alongside the Gaussian density. While the tail of the distribution of the weights is slightly heavier than the Gaussian distribution, the bulk of the distributions are rather similar*: 1*st*, 50*th and* 99*th percentiles of the empirical distribution are* $\{-0.0052, -0.0002, 0.0057\}$ *while the same percentiles for a Gaussian with matching first two moments are* $\{-0.0055, -0.0001, 0.0052\}$. (*A similar plot is obtained for the second layer.*)

2. Curse of dimensionality. An important achievement of the present paper is to establish a dimension-independent error rate $R(\hat{f}_n, f_0) \lesssim n^{-\varepsilon^*} \log^2 n$ (with ε^* independent of d for many cases of interest), holding for any near-minimizer $\hat{f}_n$ of the empirical risk. Indeed, obtaining a dimension-independent rate appears to be an important guiding principle for the construction of the function class $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$. Does this result fully address the curse of dimensionality?

The present analysis establishes an upper bound of the form $R(\hat{f}_n, f_0) \le C(d) n^{-\varepsilon_*} \log^2 n$ without characterizing the d-dependence of the prefactor. Further, it assumes n large enough, that is, $n \ge n_0(d)$ for an unspecified $n_0(d)$.

Consider the example of *additive models* which is discussed in the paper. Applying the current proof strategy requires $n \gtrsim d^d$. Namely, we consider the function

$$f_0(x_1, \dots, x_d) = \sum_{i=1}^{d} f_i(x_i), \tag{1}$$

where $f_i \in C_1^\beta([0,1], K)$. Then, we have $f_0 = g_1 \circ g_0$ where $g_0 : [0,1]^d \to [-K, K]^d$ with $g_0(\boldsymbol{x}) = (f_1(x_1), \dots, f_d(x_d))^\mathsf{T}$ and $g_1 : [-K, K]^d \to [-Kd, Kd]$ with $g_1(\boldsymbol{z}) = \sum_{i=1}^d z_i$. Since for any $\gamma > 1$, $g_1 \in C_d^\gamma([-K, K], (K+1)d)$, we have

$$f_0 \in \mathcal{G}\big(1, (d, d, 1), (1, d), \big(\beta, (\beta \vee 2)d\big), (K+1)d\big).$$

The present paper implies convergence at rate $R(\widehat{f_n}, f_0) \lesssim n^{-\frac{2\beta}{2\beta+1}} \log^3 n$. Examining the proof, we find that this bound holds for $n \ge \max_i (\beta_i + 1)^{2\beta_i^* + t_i}$. Indeed, the proof of Theorem 1 requires $N = c \cdot \max_{i=0,\dots,q} n^{t_i/(2\beta_i^* + t_i)}$ for a c small enough. The statement of Theorem 5 requires $N \ge (\beta_i + 1)^{t_i}$. This gives a lower bound of n for Theorem 1 to hold which is $n \ge \max_i (\beta_i + 1)^{2\beta_i^* + t_i}$. As a consequence, in the case of additive models we need $n \gtrsim d^d$.

While the example (1) can be treated via an ad hoc analysis,[2] it raises the question of whether the present results are strong enough to break the curse of dimensionality.

As a side remark, the condition $n \gtrsim d^d$ is not necessary for learning the model (1); see, for example, [1].

[2]One possible fix of this issue is to treat the function g_1 as the composition of $k = \log_2 d$ functions (assuming k is an integer), $g_1 = g_{1,k} \circ g_{1,k-1} \circ g_{1,1}$, where $g_{1,j} : \mathbb{R}^{d/2^{j-1}} \to \mathbb{R}^{d/2^j}$, $z \mapsto (z_1 + z_2, \dots, z_{d/2^{j-1}-1} + z_{d/2^{j-1}})$.

3. Computation. Classical statistical theory views statistical questions as decoupled from computational ones. Schmidt-Hieber's contribution belongs to this tradition: it postulates an estimator $\hat{f}_n$ that is a near-minimizer of the empirical risk and derives statistical rates for this estimator. In contrast, a broad research effort in modern high-dimensional statistics is emphasizing the fundamental role played by computational bottlenecks. For a large number of problems, there are fundamental computational limitations that are dramatically more stringent than statistical ones. A somewhat arbitrary list of examples include high-dimensional regression (for certain types of prior information) [6], matrix factorization [4], community detection [8], sparse principal component analysis [2], tensor principal component analysis [11] and so on.

In practice, multilayer neural networks are efficiently learnt via SGD or its variants. This seems to us as an important constraint on any statistical theory aiming at explaining the success of deep learning.

We consider the problem of learning a simple ridge function

$$f_\ell(\boldsymbol{x}) = \varphi_\ell\big(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle\big) \tag{2}$$

with the following choices of the nonlinearity φ_ℓ, $\ell \in \{1, 2\}$:

$$\varphi_1(x) = \frac{\tanh(x)}{0.628}, \qquad \varphi_2(x) = \frac{1}{0.1275}\big(\tanh(x) + c_1 \tanh^3(x) + c_2 \tanh^5(x)\big), \tag{3}$$

where $c_1 = -3.422$, $c_2 = 2.551$. These coefficients are chosen in such a way that $\mathbb{E}\{\varphi_i(G)^2\} \approx 1$ (for $G \sim \mathsf{N}(0, 1)$), and φ_2 has vanishing projection (in $L^2(e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi})$) onto the space of polynomials of degree at most four. (See [5] for a related construction in the statistics literature.) Both of these regression functions belong to the class $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ with $d_0 = t_0 = d$, $d_1 = t_1 = 1$, $d_2 = 1$ and $\beta_1 = \beta_2 = \infty$. The theory developed in the present paper suggests that it should be possible to estimate them at the nearly parametric rate, $(\log n)^2/n$, without distinguishing between φ_1 and φ_2.

We choose the true parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ uniformly at random with $\|\boldsymbol{\theta}\|_2 = 1$. We consider data $(y_i, \boldsymbol{x}_i)$, where $\boldsymbol{x}_i \sim \mathsf{Unif}([-a, a]^d)$, $a = \sqrt{3}$ (to fix the normalization $\mathbb{E}\{\|\boldsymbol{x}\|_2^2\} = d$) and $y_i = f_\ell(\boldsymbol{x}_i)$ for either of the two models $f_\ell(\boldsymbol{x}) = \varphi_\ell(\langle \theta_\ell, \boldsymbol{x}\rangle)$, $\ell \in \{1, 2\}$.

Using SGD, we try to learn these functions by fitting fully connected ReLU networks of various depths. Figure 2 reports the results of our experiments. We use $d = 500$, while the number of neurons in each hidden layer is fixed to 100. We vary the number of training data points, n, from $75k$ to $250k$. Our data suggest that these networks, independently of their depth, have difficulty in learning $f_2(\cdot)$. At the same time, $f_1(\cdot)$ is learnt even with a small amount of training data.

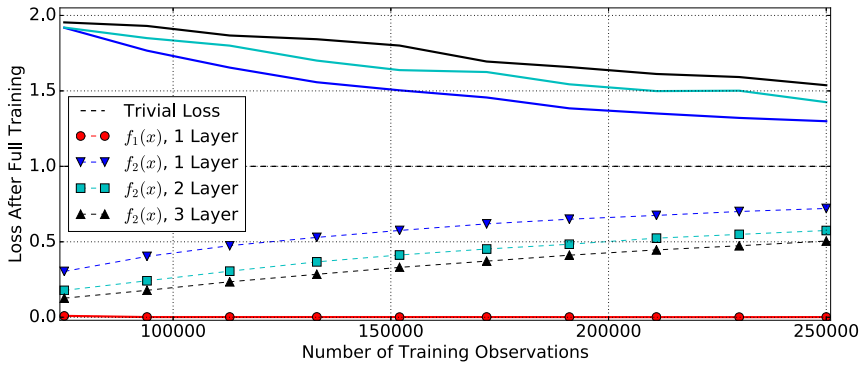

FIG. 2. *Test (solid lines) and training (dashed lines) loss for fully connected networks trained on the two data distributions* $f_\ell(\boldsymbol{x}) = \varphi_\ell(\langle \theta_\ell, \boldsymbol{x}\rangle)$, $\ell \in \{1, 2\}$. *We consider networks with* $\{1, 2, 3\}$ *hidden layers. Each model is trained for* 150 *epochs via SGD. For reference, the loss of the trivial predictor* $\hat{f}(x) = 0$ *is* 1.

Acknowledgments. A. Montanari is supported by NSF Grants DMS-1613091, CCF-1714305, IIS-1741162, and ONR N00014-18-1-2729, NSF DMS-1418362, NSF DMS-1407813.

REFERENCES

[1] BACH, F. (2017). Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res.* **18** Art. ID 19. MR3634886

[2] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.

[3] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* 3036–3046.

[4] MONTANARI, A. and VENKATARAMANAN, R. (2017). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* To appear. https://doi.org/10.1214/20-AOS1958

[5] DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106. MR0981438 https://doi.org/10.1214/aos/1176347004

[6] GAMARNIK, D. and ILIAS, Z. (2017). High dimensional regression with binary coefficients. Estimating squared error and a phase transtition. In *Conference on Learning Theory* 948–953.

[7] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. Preprint. Available at arXiv:1802.08246.

[8] HAJEK, B., WU, Y. and XU, J. (2015). Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory* 899–928.

[9] JI, Z. and TELGARSKY, M. (2018). Risk and parameter convergence of logistic regression. Preprint. Available at arXiv:1803.07300.

[10] MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115

[11] MONTANARI, A. and RICHARD, E. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.

[12] NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. Preprint. Available at arXiv:1412.6614.

[13] SIMONYAN, K. and ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint. Available at arXiv:1409.1556.

[14] SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Art. ID 70. MR3899772