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.

where α quantifies the Hölder smoothness of h. Comparing Equation (1) and Equation (2) and assuming α and β_i^* are similar, we see that the rate of neural network estimators is superior to wavelet estimators whenever $\max_i t_i \ll d$. As a result, we avoid the curse of dimensionality when the target function is sufficiently sparse.

Although this is an insightful result, I believe that the extent to which it explains the statistical performance of deep neural networks is open to debate. Rewriting Equation (1) and Equation (2) in terms of the sample size required attaining a fixed error level ϵ, and, assuming $\alpha = \beta_i^*$ and $t = t_i$ for all i, we get that neural network estimators require $(1/\epsilon)^{1+t/(2\alpha)}$ observations, whereas wavelet estimators require $(1/\epsilon)^{1+d/(2\alpha)}$ observations. Clearly, the sample size with neural networks is much smaller when $t \ll d$, but the dependence on t is still exponential which means that only extremely sparse models can be accounted for. Essentially, we have replaced a "curse of dimensionality" effect with a "curse of sparsity." To give a quantitative example, suppose[1] $\alpha = 4$, and we wish to attain $\epsilon = 0.05$. Under these assumptions a sample complexity bound of $(1/\epsilon)^{1+t/\alpha}$ is less than a billion (10^9) only for sparsity $t \leq 23$. Since having a billion examples is often too optimistic even for modern large-scale applications, this means that we can only realistically estimate models whose sparsity is up to 23. It is not clear that such extremely sparse models can capture the complex real-world tasks on which neural networks are successfully trained (e.g., object recognition in images where state-of-the-art methods generally depend on more than 23 input pixels in a typical image).

An alternative approach, which has received much attention in the machine learning community in recent years, focuses on parametric models where the target function f_0 is assumed to be a neural network or even on distribution-free models where the underlying data distribution can be arbitrary, and we only attempt to find a predictor whose risk is not much worse than the best neural network from a given class (see, e.g., Anthony and Bartlett (2009), Bartlett, Foster and Telgarsky (2017), Dziugaite and Roy (2017), Golowich, Rakhlin and Shamir (2018), Neyshabur, Tomioka and Srebro (2015)). Such approaches have several motivations: First, since we are attempting to construct neural network predictors of a given architecture, it is natural to assume that the target function is approximately contained in that class (otherwise, it might be better to use a different predictor class to begin with). Second, the rates are generally of order $n^{-\theta}$ for some constant $\theta \in [\frac{1}{2}, 1]$, and the bounds can be meaningful even for very large neural networks, sometimes without explicit dependence on their sizes or any sparsity constraints. Third, distribution-free bounds depend on the learned model rather than the target function, and, hence, they can provide a more fine-grained understanding of what kind of networks lead to good statistical performance. Of course, this does not mean that nonparametric models are not useful. We are still far from a full understanding of the statistical aspects of neural networks, and insights from multiple viewpoints would probably be helpful. However, the fact that extremely large and dense networks are successfully learned in practice hint that factors beyond hidden sparsity may be at play.

Computational considerations. In his paper, Schmidt-Hieber mentions heuristic arguments (e.g., Choromanska et al. (2015)), which indicate that, when training neural networks, local minima in the training objective have a value relatively close to the global minima. It is then stated that "If the heuristic argument can be made rigorous... [t]his would allow us then to study deep learning without an explicit analysis of the algorithm." Although it is often useful to study statistical aspects of a learning problem ignoring computational/algorithmic aspects, I would like to discuss a few potential pitfalls with such an approach, especially in the context of deep learning.

[1]The choice of $\alpha = 4$ here is rather arbitrary, but it is important to note that its choice affects other constants multiplying the bound in Equation (2).

In learning problems which can be reduced to convex optimization problems (such as learning linear predictors with respect to a convex loss function), there is typically a unique minimum which is also global. As a result, one can indeed study the statistical properties of this minimum, ignoring the computational question of how this minimum is found. However, we know that deep-learning problems are typically nonconvex, with the training objective having multiple (often infinitely many) global minima, which differ sharply in their statistical properties. For example, multiple experiments have demonstrated that fitting the weights of large neural networks (using standard algorithms) often achieve zero error on the training data and low error on a validation set, even though there exist other weight configurations for the network with zero training error which strongly overfit (e.g., Zhang et al. (2016)). In such scenarios, even if we assume that a global minimum can always be found, it is impossible to provide meaningful statistical guarantees, unless we take the optimization algorithm and the type of global minimum it returns into account. Indeed, several recent works in the machine learning literature studied the implicit statistical biases which are induced by various algorithms (such as Arora et al. (2019), Li, Ma and Zhang (2018), Soudry et al. (2018) but also going back to, e.g., Schapire et al. (1998) in the context of boosting), whereas other works (such as Bartlett et al. (2019), Belkin et al. (2019), Hastie et al. (2019)) studied how certain parametric estimators which interpolate the training data can avoid overfitting, even in the presence of noise. Of course, purely statistical analyses are still useful, but one needs to be mindful that they probably cannot fully explain the statistical behavior of deep learning systems.

Another more technical issue is the implicit assertion that stochastic gradient descent is capable at all of finding local minima (hence, if the problem is such that all local minima are close to global, the algorithm will find a nearly-optimal solution). Although it is not the focus of this paper, I would like to point out that even this modest assertion is not known to be true: For the type of nonconvex optimization problems which arise in deep learning, we know at best that there is convergence to a stationary point suitably defined (see, e.g., Davis et al. (2020)). Moreover, even when there is convergence to a local minimum, the convergence may be extremely slow, even for generalized linear models with Gaussian inputs and other simple setups (see, e.g., Shalev-Shwartz, Shamir and Shammah (2017), Shamir (2018)). Thus, even if we can show that all local minima have a value close to global ones in deep learning problems—a result which still seems out of reach—that would still be far from sufficient to explain why deep learning systems can be successfully trained.

The importance of adaptiveness. The paper shows that, in contrast to neural network estimators, wavelet estimators based on a uniform random design have a rate deteriorating exponentially with the input dimension, even when the target function satisfies the structural assumptions discussed earlier. In what follows, I would like to point out that this bad dimension dependence also applies to a much larger class of estimators but in a somewhat different sense. Specifically, consider an estimator based on fitting a linear combination of random basis functions, $f_1, \ldots, f_r$, and chosen from some fixed distribution over a function class $\mathcal{F}$. The wavelet estimators discussed in the paper are a very special case, where each f_i is a wavelet function with a center sampled uniformly at random from $[0, 1]^d$. Furthermore, let us ignore all statistical considerations, assume that we know the target function f_0 precisely and find the best approximation using any linear combination of $f_1, \ldots, f_r$. Then, under mild assumptions and in order to achieve some constant accuracy, either the *number* r of basis functions or the magnitude of the weights must be *exponential* in the dimension d. Importantly, this holds even for specific f_0 as simple as a neural network composed of a single neuron. This can be formalized as follows:

THEOREM 1 (Yehudai and Shamir (2019)). *There exists a universal constant $c > 0$ such that the following holds. Let $d > 40$, and let $\mathcal{F}$ be a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})^2] \le \exp(d/40)$. Also, for some $r \in \mathbb{N}$, let D be an arbitrary distribution over tuples $(f_1, \dots f_r)$ of functions from $\mathcal{F}$. Then, there exists a ReLU neuron function $f_0(\mathbf{x}) = \max\{0, \langle \mathbf{w}^*, \mathbf{x}\rangle + b^*\}$ (for some $\mathbf{w}^*, b^*$ with $\|\mathbf{w}^*\| = d^2, |b^*| \le 6d^3 + 1$) such that w.p. at least $1 - r\exp(-cd)$ over sampling $f_1, \dots, f_r$ from D, if $\mathbb{E}_{x\sim\mathcal{N}(0,I)}[(\sum_{i=1}^r u_i f_i(x) - [\langle w^*, x\rangle + b^*]_+)^2] \le \frac{1}{50}$; then,*

$$r \cdot \max_i |u_i| \ge \frac{1}{48d^2}\exp(cd).$$

See also Ghorbani et al. (2019) for related results. Note that in Theorem 1, we crucially assume that the estimator is "nonadaptive," in the sense that the distribution D does not depend on the target function f_0, but, otherwise, only minimal assumptions on the function class $\mathcal{F}$ are made. In particular, it encompasses wavelet estimators, kernel density estimators, algorithms based on random features (e.g., Rahimi and Recht (2008)), kernel methods, etc. In general, it implies that any kind of estimator based on "reasonable" linear combinations of nonadaptive features is inadequate to express even the simplest class of neural networks and even ignoring statistical considerations. Intuitively, the result holds since the class of ReLU neurons occupies a high-dimensional manifold in L^2 function space, and it cannot be well approximated by any small number of basis functions (at least with subexponential weights).

Of course, this result naturally suggests the use of "adaptive" estimators: for example, it is known that single neurons are learnable with an estimator composed of a single neuron (e.g., Mei, Bai and Montanari (2018), Soltanolkotabi (2017)). Thus, it would be interesting to understand which adaptive estimators have good performance under the model proposed in Schmidt-Hieber's paper.

Acknowledgments. I thank Boaz Nadler for very helpful discussions and comments on a draft of this paper.

REFERENCES

ANTHONY, M. and BARTLETT, P. L. (2009). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038 https://doi.org/10.1017/CBO9780511624216

ARORA, S., COHEN, N., HU, W. and LUO, Y. (2019). Implicit regularization in deep matrix factorization. Preprint. Available at arXiv:1905.13655.

BARTLETT, P. L., FOSTER, D. J. and TELGARSKY, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *NIPS*.

BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2019). Benign overfitting in linear regression. Preprint. Available at arXiv:1906.11300.

BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

CHOROMANSKA, A., HENAFF, M., MATHIEU, M., AROUS, G. B. and LECUN, Y. (2015). The loss surfaces of multilayer networks. In *AISTATS*.

DAVIS, D., DRUSVYATSKIY, D., KAKADE, S. and LEE, J. D. (2020). Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* **20** 119–154. MR4056927 https://doi.org/10.1007/s10208-018-09409-5

DZIUGAITE, G. K. and ROY, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. Preprint. Available at arXiv:1703.11008.

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Limitations of lazy training of two-layers neural networks. Preprint. Available at arXiv:1906.08899.

GOLOWICH, N., RAKHLIN, A. and SHAMIR, O. (2018). Size-independent sample complexity of neural networks. In *COLT*.

HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. Preprint. Available at arXiv:1903.08560.

LI, Y., MA, T. and ZHANG, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT*.

MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015). Norm-based capacity control in neural networks. In *COLT*.

RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *NIPS*.

SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686. MR1673273 https://doi.org/10.1214/aos/1024691352

SHALEV-SHWARTZ, S., SHAMIR, O. and SHAMMAH, S. (2017). Failures of gradient-based deep learning. In *ICML*.

SHAMIR, O. (2018). Distribution-specific hardness of learning neural networks. *J. Mach. Learn. Res.* **19** Art. ID 32. MR3862439

SOLTANOLKOTABI, M. (2017). Learning relus via gradient descent. In *NIPS*.

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

YEHUDAI, G. and SHAMIR, O. (2019). On the power and limitations of random features for understanding neural networks. In *NeurIPS*.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. Preprint. Available at arXiv:1611.03530.