Deep Neural Networks Learn Non-Smooth Functions Effectively

Masaaki Imaizumi (Institute of Statistical Mathematics)
Kenji Fukumizu (Institute of Statistical Mathematics)

1 Deep Neural Network (DNN)

Deep neural networks (DNNs) have shown outstanding performance on various tasks of data analysis. Enjoying their flexible modeling by a multi-layer structure and many elaborate computational and optimization techniques, DNNs empirically achieve higher accuracy than many other machine learning methods such as kernel methods. Hence, DNNs are employed in many successful applications, such as image analysis, medical data analysis, natural language processing, and others.

Despite such outstanding performance of DNNs, little is yet known why DNNs outperform the other methods. Without sufficient understanding, practical use of DNNs could be inefficient or unreliable. To reveal the mechanism, numerous studies have investigated theoretical properties of neural networks from various aspects. The approximation theory has analyzed the expressive power of neural networks, the statistical learning theory elucidated generalization errors, and the optimization theory has discussed the landscape of the objective function and dynamics of learning.

2 Analyze DNN through Regression Framework

One limitation in the existing statistical analysis of DNNs is a *smoothness assumption* for data generating processes, which requires that data $\{(Y_i, X_i)\}$ are given by

$$Y_i = f(X_i) + \xi_i,\ \xi_i \sim \mathcal{N}(0, \sigma^2),$$

where f is a β-times differentiable function with D-dimensional input. With this setting, however, many popular methods such as kernel methods, Gaussian processes, series methods, and so on, as well as DNNs, achieve a bound for generalization errors as

$$O\left(n^{-2\beta/(2\beta+D)}\right), \qquad (n \to \infty).$$

This is known to be a minimax optimal rate of generalization with respect to sample size n, and hence, as long as we employ the smoothness assumption, it is not easy to show a theoretical evidence for the empirical advantage of DNNs.

This paper considers estimation of *non-smooth* functions for the data generating processes to break the difficulty. Specifically, we discuss a nonparametric regression problem with a class of *piecewise smooth functions*, which may be non-smooth on the boundaries of pieces in their domains. Then, we derive a rate of generalization errors with the least square and Bayes estimators by DNNs of the ReLU activation as

$$O\left(\max\left\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\right\}\right), \qquad (n \to \infty)$$

up to log factors. Here, α and β denote a degree of smoothness of piecewise smooth functions, and D is the dimensionality of inputs. We prove also that this rate of generalizations by DNNs is optimal in the minimax sense. In addition, we show that some of other standard methods, such as kernel methods and orthogonal series methods, are not able to achieve this optimal rate. Our results thus show that DNNs certainly have a theoretical advantage under the non-smooth setting. We will provide some numerical results supporting our results.