Deep Neural Networks Learn Non-Smooth Functions Effectively

Masaaki Imaizumi
The Institute of Statistical Mathematics

Kenji Fukumizu
The Institute of Statistical Mathematics

Abstract

We elucidate a theoretical reason that deep neural networks (DNNs) perform better than other models in some cases from the viewpoint of their statistical properties for non-smooth functions. While DNNs have empirically shown higher performance than other standard methods, understanding its mechanism is still a challenging problem. From an aspect of the statistical theory, it is known many standard methods attain the optimal rate of generalization errors for smooth functions in large sample asymptotics, and thus it has not been straightforward to find theoretical advantages of DNNs. This paper fills this gap by considering learning of a certain class of non-smooth functions, which was not covered by the previous theory. We derive the generalization error of estimators by DNNs with a ReLU activation, and show that convergence rates of the generalization by DNNs are almost optimal to estimate the non-smooth functions, while some of the popular models do not attain the optimal rate. In addition, our theoretical result provides guidelines for selecting an appropriate number of layers and edges of DNNs. We provide numerical experiments to support the theoretical results.

1 Introduction

Deep neural networks (DNNs) have shown outstanding performance on various tasks of data analysis Schmidhuber (2015); LeCun et al. (2015). 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 Hinton et al. (2006); Le et al. (2011); Kingma and Ba (2014). Hence, DNNs are employed in many successful applications, such as image analysis He et al. (2016), medical data analysis Fakoor et al. (2013), natural language processing Collobert and Weston (2008), and others.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

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 Cybenko (1989); Barron (1993); Bengio and Delalleau (2011); Montufar et al. (2014); Yarotsky (2017); Petersen and Voigtlaender (2018); Bölcskei et al. (2017), the statistical learning theory elucidated generalization errors Barron (1994); Neyshabur et al. (2015); Schmidt-Hieber (2017); Zhang et al. (2017); Suzuki (2018), and the optimization theory has discussed the landscape of the objective function and dynamics of learning Baldi and Hornik (1989); Fukumizu and Amari (2000); Dauphin et al. (2014); Kawaguchi (2016); Soudry and Carmon (2016).

Existing statistical analysis does not explain the empirical success of DNNs, since it is already proved that the standard machine learning methods are statistically optimal with a *smoothness assumption* for data generating processes. Specifically, it is usually assumed that data $\{(Y_i, X_i)\}_{i=1}^n$ are given i.i.d. 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 Tsybakov (2009); Wasserman (2006). With this setting, 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 Stone (1982);

Tsybakov (2009); Giné and Nickl (2015), 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.

To break the difficulty, this paper develops a statistical theory for estimation of *non-smooth* functions for the data generating processes. Rigorously, we discuss a nonparametric regression problem with a class of *piecewise smooth functions*, which may be non-smooth, and even discontinuous, 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

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

up to log factors (Theorems 1, 2). Here, α and β denote the smoothness degree of functions on the boundary and interior of the domain, and D is the dimensionality of inputs. We prove also that this rate of generalizations by DNNs is optimal in the minimax sense (Theorem 3). 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 examples supporting our results.

The contributions of this paper are as follows:

- We derive a rate of convergence of the generalzation errros in the estimators by DNNs for the class of piecewise smooth functions. Our convergence results are more general than existing studies, since the class contains the smooth functions.
- We prove that DNNs theoretically outperform other standard methods for data from non-smooth generating processes, as a consequence of the proved convergence rate of generalization error.
- We provide a practical guideline on the structure of DNNs; namely, we show a necessary number of layers and parameters of DNNs to achieve the rate of convergence. It is shown in Table 1.

All proofs are deferred to the supplementary material.

Element	Number
# of layers	$O(1 + \max\{\beta/D, \alpha/2(D-1)\})$
# of parameters	$\Theta(n^{\max\{D/(2\beta+D),(D-1)/(\alpha+D-1)\}})$

Table 1: Architecture for DNNs which are necessary to achieve the optimal rate of generalization errors.

1.1 Notation

We use notations $I := [0,1]$ and $\mathbb{N}$ for natural numbers. The j-th element of vector b is denoted by b_j, and $\|\cdot\|_q := (\sum_j b_j^q)^{1/q}$ is the q-norm ($q \in [0,\infty]$). $\mathrm{vec}(\cdot)$ is a vectorization operator for matrices. For $z \in \mathbb{N}$, $[z] := \{1,2,\ldots,z\}$ is the set of positive integers no more than z. For a measure P on I and a function $f : I \to \mathbb{R}$, $\|f\|_{L^2(P)} := (\int_I |f(x)|^2 dP(x))^{1/2}$ denotes the $L^2(P)$ norm. $\otimes$ denotes a tensor product, and $\bigotimes_{j\in[J]} x_j := x_1 \otimes \cdots \otimes x_J$ for a sequence $\{x_j\}_{j\in[J]}$. For a set $R \subset I^D$, let $\boldsymbol{1}_R : I^D \to \{0,1\}$ denote the indicator function of R; i.e., $\boldsymbol{1}_R(x) = 1$ if $x \in R$, and $\boldsymbol{1}_R(x) = 0$ otherwise. Let $H^\beta(\Omega)$ be the Hölder space on Ω wifh a set Ω, which is the space of functions $f : \Omega \to \mathbb{R}$ such that they are $\lfloor\beta\rfloor$-times continuously differentiableand the derivatives is $\beta - \lfloor\beta\rfloor$-Hölder continuous. For a vector $x \in \mathbb{R}^{D'}$, $x_{-d} := (x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D)$.

2 Preparation for Regression with DNNs

2.1 Regression Problem

Let the D-dimensional cube I^D ($D \geqslant 2$) be a space for input variables X_i. Suppose we have a set of observations $(X_i, Y_i) \in I^D \times \mathbb{R}$ for $i \in [n]$ which is independently and identically distributed with the data generating process

$$Y_i = f^*(X_i) + \xi_i, \tag{1}$$

where $f^* : I^D \to \mathbb{R}$ is an unknown true function and ξ_i is Gaussian noise with mean 0 and variance $\sigma^2 > 0$ for $i \in [n]$. We assume that the marginal distribution of X on I^D has a positive and bounded density function $P_X(x)$.

The goal of the regression problem is to estimate f^* from the set of observations $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]}$. With an estimator $\hat{f}$, its performance is measured by the $L^2(P_X)$ norm: $\|\hat{f} - f^*\|^2_{L^2(P_X)} = \mathbb{E}_{X\sim P_X}[(\hat{f}(X) - f^*(X))^2]$. There are various methods to estimate f^* and their statistical properties are extensively investigated (For summary, see Wasserman (2006) and Tsybakov (2009)).

2.2 Deep Neural Network Models

Let $L \in \mathbb{N}$ be the number of layers in DNNs. For $\ell \in [L+1]$, let $D_\ell \in \mathbb{N}$ be the dimensionality of variables in the ℓ-th layer. For brevity, we set $D_{L+1} = 1$, i.e., the output is one-dimensional. We define $A_\ell \in \mathbb{R}^{D_{\ell+1}\times D_\ell}$ and $b_\ell \in \mathbb{R}^{D_\ell}$ be matrix and vector parameters to give the transform of ℓ-th layer. The *architecture* Θ of DNN is a set of L pairs of (A_ℓ, b_ℓ):

$$\Theta := ((A_1, b_1), ..., (A_L, b_L)).$$

We define $|\Theta| := L$ be a number of layers in Θ, $\|\Theta\|_0 := \sum_{\ell\in[L]} \|\operatorname{vec}(A_\ell)\|_0 + \|b_\ell\|_0$ as a number of non-zero elements in Θ, and $\|\Theta\|_\infty := \max\{\max_{\ell\in[L]} \|\operatorname{vec}(A_\ell)\|_\infty, \max_{\ell\in[L]} \|b_\ell\|_\infty\}$ be the largest absolute value of the parameters in Θ.

For an activation function $\eta : \mathbb{R}^{D'} \to \mathbb{R}^{D'}$ for each $D' \in \mathbb{N}$, this paper considers the ReLU activation $\eta(x) = (\max\{x_d, 0\})_{d\in[D']}$.

The model of neural networks with architecture Θ and activation η is the function $G_\eta[\Theta] : \mathbb{R}^{D_1} \to \mathbb{R}$, which is defined inductively as

$$G_\eta[\Theta](x) = x^{(L+1)},$$

and it is inductively defined as

$$x^{(1)} := x, \quad x^{(\ell+1)} := \eta(A_\ell x^{(\ell)} + b_\ell), \text{ for } \ell \in [L],$$

where $L = |\Theta|$ is the number of layers. The set of model functions by DNNs is thus given by

$$\Xi_{NN,\eta}(S, B, L') := \left\{ G_\eta[\Theta] : I^D \to \mathbb{R} \mid \|\Theta\|_0 \leq S, \|\Theta\|_\infty \leq B, |\Theta| \leq L' \right\},$$

with $S \in \mathbb{N}$, $B > 0$, and $L' \in \mathbb{N}$. Here, S bounds the number of non-zero parameters of DNNs by Θ, namely, the number of edges of an architecture in the networks. This also describes sparseness of DNNs. B is a bound for scales of parameters.

2.3 Two Estimators with DNNs

A Least Square Estimator

We define a least square estimator by empirical risk minimization, using the model of DNNs. Using the observations $\mathcal{D}_n$, we consider the minimization problem with respect to parameters of DNNs as

$$\hat{f}^L \in \underset{\bar{f}: f\in\Xi_{NN,\eta}(S,B,L)}{\operatorname{argmin}} \frac{1}{n} \sum_{i\in[n]} (Y_i - \bar{f}(X_i))^2, \quad (2)$$

where $\bar{f} := \max\{\min\{f, -T\}, T\}$ is a clipping operation for $f \in \Xi_{NN,\eta}(S, B, L)$ with a sufficiently large threshold $T > 0$. We use $\hat{f}^L$ as an estimator of f^*.

Note that the problem (2) has at least one minimizer since the parameter set Θ is compact and η is continuous. If necessary, we can add a regularization term for the problem (2), because it is not difficult to extend our results to an estimator with regularization. Furthermore, we can apply the early stopping techniques, since they play a role as the regularization LeCun et al. (2015). However, for simplicity, we confine our arguments of this paper in the least square.

A Bayes Estimator

We also define a Bayes estimator for DNNs which can avoid the non-convexity problem in optimization. Fix architecture Θ and $\Xi_{NN,\eta}(S, B, L)$ with given S, B and L. Then, a prior distribution for $\Xi_{NN,\eta}(S, B, L)$ is defined through providing distributions for the parameters contained in Θ. Let $\Pi_\ell^{(A)}$ and $\Pi_\ell^{(b)}$ be distributions of A_ℓ and b_ℓ as

$$A_\ell \sim \Pi_\ell^{(A)} \text{ and } b_\ell \sim \Pi_\ell^{(b)},$$

for $\ell \in [L]$. We set $\Pi_\ell^{(A)}$ and $\Pi_\ell^{(b)}$ such that each of the S parameters of Θ is uniformly distributed on $[-B, B]$, and the other parameters degenerate at 0. Using these distributions, we define a prior distribution Π_Θ on Θ by $\Pi_\Theta := \bigotimes_{\ell\in[L]} \Pi_\ell^{(A)} \otimes \Pi_\ell^{(b)}$. Then, a prior distribution for $f \in \Xi_{NN,\eta}(S, B, L)$ is defined by

$$\Pi_f(f) := \Pi_\Theta(\Theta : G_\eta[\Theta] = f).$$

Then, we can obtain the posterior distribution for f. Since the noise ξ_i in (1) is Gaussian with its variance σ^2, the posterior distribution is given by

$$d\Pi_f(f|\mathcal{D}_n) = \frac{\exp(-\sum_{i\in[n]}(Y_i - f(X_i))^2/\sigma^2) d\Pi_f(f)}{\int \exp(-\sum_{i\in[n]}(Y_i - f'(X_i))^2/\sigma^2) d\Pi_f(f')}.$$

Finally, we define a Bayes estimator as a posterior mean

$$\hat{f}^B := \int f d\Pi_f(f|\mathcal{D}_n),$$

by the Bochner integral in $L^\infty(I^D)$.

Note that we do not discuss computational issues of the Bayesian approach since the main focus is a theoretical aspect. To solve the computational problems, see Hernández-Lobato and Adams (2015) and others.

3 Specification for Non-Smooth Functions

We specify a formulation of non-smooth functions to prove a theoretical advantage of DNNs, as motivated in the introduction of this paper. To describe non-smoothness of functions, we introduce a notion of *piecewise smooth functions* which have a support divided into several pieces and smooth only within each of the pieces. On boundaries of the pieces, piecewise smooth functions are non-smooth, i.e. non-differentiable and even discontinuous. Figure 1 shows an example of piecewise smooth functions.

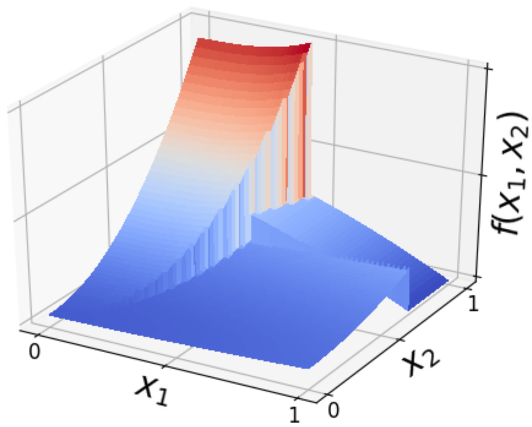

Figure 1: An example of piecewise smooth functions with a 2-dimensional support. The support is divided into three pieces and the function is discontinuous on their boundaries.

3.1 Preparation: Pieces in Supports

Preliminarily, we describe a notion of pieces in the domain I^D. Here, we use the class of *horizon functions* Petersen and Voigtlaender (2018).

Let us consider a smooth function $h \in H^\alpha(I^{D-1})$. Then, we define a horizon function $\Psi_{h,d} : I^D \to \{0,1\}$ as

$$\Psi_{h,d} := \Psi_d(x_1, \ldots, x_{d-1}, x_d \pm h(x_{-d}), x_{d+1}, ..., x_D),$$

for some $d \in [D]$, where $\Psi_d : I^D \to \{0,1\}$ is the Heaviside function such that $\Psi_d(x) = \mathbf{1}_{\{x \in I^D | x_d \geqslant 0\}}$.

For each horizon function, we define a *basis piece* $A \subset I^D$ such that there exist $\Psi_{h,d}$ such that

$$A = \left\{x \in I^D \mid \Psi_{h,d}(x) = 1\right\}.$$

A basis piece is regarded as one side of surfaces by h. Additionally, we introduce a restrict for A as a transformed sphere, namely, we consider $\Psi_{h,d}$ such that there exists an α-smooth embedding $e : \{x \in \mathbb{R}^D \mid \|x\|_2 \leqslant 1\} \to \mathbb{R}^D$ satisfying $A = I^D \cap \mathrm{Image}(e)$ (detail is provided in Appendix A). The notion of basis pieces is an extended version of the boundary fragment class Dudley (1974); Mammen et al. (1999) which is dense in a class of all convex sets in I^D when $\alpha = 2$.

We define a *piece* by the intersection of J basis pieces; namely, the set of pieces is defined by

$$\mathcal{R}_{\alpha,J} := \left\{R \subset [0,1]^D \mid R = \bigcap_{j=1}^J A_j\right\},$$

where $A_1, ..., A_J$ are basic pieces.

Intuitively, $R \in \mathcal{R}_{\alpha,J}$ is a set with piecewise α-smooth boundaries. Also, by considering intersections of J basis pieces, $\mathcal{R}_{\alpha,J}$ contains a set with non-smooth boundaries. In Figure 1, there are three pieces from $\mathcal{R}_{\alpha,J}$ in the support of the function.

3.2 Piecewise Smooth Functions

We define piecewise smooth functions, using $H^\beta(I^D)$ and $\mathcal{R}_{\alpha,J}$. Let $M \in \mathbb{N}$ be a finite number of pieces of the support I^D. We introduce the set of piecewise smooth functions by

$$\mathcal{F}_{M,J,\alpha,\beta} := \left\{\sum_{m=1}^M f_m \otimes \mathbf{1}_{R_m} : f_m \in H^\beta(I^D), R_m \in \mathcal{R}_{\alpha,J}\right\}.$$

Since $f_m(x)$ realizes only when $x \in R_m$, the notion of $\mathcal{F}_{M,J,\alpha,\beta}$ can express a combination of smooth functions on each piece R_m. Hence, functions in $\mathcal{F}_{M,J,\alpha,\beta}$ are non-smooth (and even discontinuous) on boundaries of R_m. Obviously, $H^\beta(I^D) \subset \mathcal{F}_{M,J,\alpha,\beta}$ with $M = 1$ and $R_1 = I^D$, hence the notion of piecewise smooth functions can describe a wider class of functions.

4 Main Results

We provide theoretical results about performances of DNNs for estimating piecewise smooth functions.

4.1 Generalization Errors by DNNs

The Least Square Estimator $\widehat{f}^L$

We investigate theoretical aspects of convergence properties of $\widehat{f}^L$.

Theorem 1. *(Convergence Rate of $\widehat{f}^L$)* *Suppose $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ Then, there exist constants $c_1, c_1', C_L > 0, s \in \mathbb{N}\backslash\{1\}$, $T \geqslant \|f^*\|_{L^\infty}$, and a tuple (S, B, L) satisfying*

(i) $S = c_1' \max\{n^{D/(2\beta+D)}, n^{(D-1)/(\alpha+D-1)}\}$,
(ii) $B \geqslant c_1 n^s$,
(iii) $L \leqslant c_1(1 + \max\{\beta/D, \alpha/2(D-1)\})$,

such that $\widehat{f}^L \in \Xi_{NN,\eta}(S, B, L)$ provides

$$\|\widehat{f}^L - f^*\|^2_{L^2(P_X)} \leqslant C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2, \quad (3)$$

with probability at least $1 - c_1 n^{-2}$.

The rate of convergence in Theorem 1 is simply interpreted as follows. The first term $n^{-2\beta/(2\beta+D)}$ describes an effect of estimating $f_m \in H^\beta(I^D)$ for $m \in [M]$. The rate corresponds to the minimax optimal convergence rate of generalization errors for estimating smooth functions in $H^\beta(I^D)$ (For a summary, see Tsybakov (2009)). The second term $n^{-\alpha/(\alpha+D-1)}$ reveals an effect from estimation of $\boldsymbol{1}_{R_m}$ for $m \in [M]$ through estimating the boundaries of $R_m \in \mathcal{R}_{\alpha,J}$. The same rate of convergence appears in a problem for estimating sets with smooth boundaries Mammen and Tsybakov (1995).

We remark that a larger number of layers decreases B. Considering the result by Bartlett (1998), which shows that large values of parameters make the performance of DNNs worse, the above theoretical result suggests that a deep structure can avoid the performance loss caused by large parameters.

We can consider an error from optimization independent to the statistical generalization. The following proposition provides the statement.

Proposition 1. *(Effect of Optimization)*
If a learning algorithm outputs $\check{f}^L \in \Xi_{NN,\eta}(S,B,L)$ such that

$$n^{-1}\sum_{i\in[n]}(Y_i - \check{f}^L(X_i))^2 - (Y_i - \hat{f}^L(X_i))^2 \leqslant \Delta_n,$$

with a positive parameter Δ_n, then the following holds:

$$\begin{aligned}&\mathbb{E}_{f^*}\left[\|\check{f}^L - f^*\|^2_{L^2(P_X)}\right]\\ &\leqslant C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2 + \Delta_n.\end{aligned}$$

Here, $\mathbb{E}_{f^*}[\cdot]$ denotes an expectation with respect to the true distribution of (X,Y). Applying results on the magnitude of Δ (e.g. Kawaguchi (2016)), we can evaluate generalization including optimization errors.

The Bayes Estimator $\hat{f}^B$

We provide theoretical analysis of the speed of convergence for the Bayes estimator.

Theorem 2. *(Convergence Rate of $\hat{f}^B$)*
Suppose $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Then, there exist constants $c_2, c_2', C_B > 0, s \in \mathbb{N}\backslash\{1\}$, architecture $\Theta : \|\Theta\|_0 \leqslant S, \|\Theta\|_\infty \leqslant B, |\Theta| \leqslant L$ satisfying following conditions:

(i) $S = c_2' \max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}$,
(ii) $B \geqslant c_2 n^s$,
(iii) $L \leqslant c_2(1 + \max\{\beta/D, \alpha/2(D-1)\})$,

and a prior distribution Π_f which provides the Bayes estimator $\hat{f}^B$ such that

$$\begin{aligned}&\mathbb{E}_{f^*}\left[\|\hat{f}^B - f^*\|^2_{L^2(P_X)}\right]\\ &\leqslant C_B \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2.\end{aligned}$$

To provide proof of Theorem 2, we additionally apply studies for statistical analysis for Bayesian nonparametrics van der Vaart and van Zanten (2008, 2011).

This result states that the Bayes estimator can achieve the same rate as the least square estimator shown in Theorem 1. Since the Bayes estimator does not use optimization, we can avoid the non-convex optimization problem, while the computation of the posterior and mean are not straightforward.

4.2 Optimality of the DNN Estimators

We show optimality of the rate of convergence by the DNN estimators in Theorem 1 and 2. We employ a theory of minimax optimal rate which is known in the field of mathematical statistics Giné and Nickl (2015). The theory derives a lower bound of a convergence rate with arbitrary estimators, thus we can obtain a theoretical limitation of convergence rates.

The following theorem shows the minimax optimal rate of convergence for the class of piecewise smooth functions $\mathcal{F}_{M,J,\alpha,\beta}$.

Theorem 3. *(Minimax Rate for $\mathcal{F}_{M,J,\alpha,\beta}$)*
Consider $\bar{f}$ is an arbitrary estimator for $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Then, there exists a constant $C_{mm} > 0$ such that

$$\begin{aligned}&\inf_{\bar{f}} \sup_{f^*\in\mathcal{F}_{M,J,\alpha,\beta}} \mathbb{E}_{f^*}\left[\|\bar{f} - f^*\|^2_{L^2(P_X)}\right]\\ &\geqslant C_{mm}\max\left\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\right\}.\end{aligned}$$

Proof of Theorem 3 is deferred to the appendix, and it employs techniques in the minimax theory developed by Yang and Barron (1999) and Raskutti et al. (2012), and entropy analysis for a family of sets Dudley (1974); Mammen and Tsybakov (1995).

We show that the rate of convergence by the estimators with DNNs are optimal in the minimax sense, since the rates in Theorems 1 and 2 correspond to the lower bound of Theorem 3 up to a log factor. In other words, for estimating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, no other methods could achieve a better rate than the estimators by DNNs.

5 Discussion: Why DNNs work better?

5.1 Non-Optimality of Other Methods

We discuss non-optimality of some of other standard methods to estimate piecewise smooth functions. To this end, we consider a class of *linear estimators*. The class contains any estimators with the following formulation:

$$\widehat{f}^{\text{lin}}(x) = \sum_{i \in [n]} \Upsilon_i(x; X_1, ..., X_n) Y_i, \quad (4)$$

where Υ_i is an arbitrary function which depends on $X_1, ..., X_n$. Various estimators are regarded as linear estimators, for examples, kernel methods, Fourier estimators, splines, Gaussian process, and others.

A study for nonparametric statistics (Section 6 in Korostelev and Tsybakov (2012)) proves inefficiency of linear estimators with non-smooth functions. Based on the result, the following corollary holds:

Corollary 1. *(Theoretical Advantage of DNNs) Suppose $\alpha D/(2\alpha + 2D - 2) \leqslant \beta$ holds. Then, there exist $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ such that $\widehat{f} \in \{\widehat{f}^L, \widehat{f}^B\}$ and any $\widehat{f}^{\text{lin}}$, large n provides*

$$\mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|^2_{L^2(P_X)}\right] < \mathbb{E}_{f^*}\left[\|\widehat{f}^{\text{lin}} - f^*\|^2_{L^2(P_X)}\right].$$

This result shows that a wide range of the other methods has larger generalization errors, hence the estimators by DNNs can overcome the other methods. Some specific methods are analyzed in the supplementary material.

According to the results, we can see that the estimators by DNNs have the theoretical advantage than the others for estimating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, since the estimators by DNNs achieve the optimal convergence rate of generalization errors and the others do not. About the inefficiency of the other methods, we do not claim that every statistical method except DNNs misses the optimality for estimating piecewise smooth functions. Our argument is the advantage of DNNs against linear estimators.

5.2 Intuition for the performance of DNNs

We provide some intuitions on why DNNs are optimal and the others are not.

Firstly, DNNs can easily approximate indicator functions $\mathbf{1}_R, R \in \mathcal{R}_{\alpha,J}$ with a small number of parameters, due to activation functions and a composition structure. A difference of two ReLU functions can approximate step functions, and a composition of the step functions in a combination of other parts of the network can easily express smooth functions restricted to pieces. Rigorously, for $x \in \mathbb{R}$, a step function $\mathbf{1}_{\{x \geqslant 0\}}$ is approximated by

$$\mathbf{1}_{\{x \geqslant 0\}} \approx \eta(ax) - \eta(ax - 1/a) =: \zeta(x), \quad (5)$$

with sufficiently large $a > 0$, and for some $R \in \mathcal{R}_{\alpha,J}$, we can approximate $\mathbf{1}_R$ as

$$\mathbf{1}_R \approx \zeta \circ G,$$

where $G \in \Xi(S_f, B_f, L_f)$ with some network such that $G \approx f \in H^\beta(I)$. Substantially, we need only $S_f + 4$ parameters to approximate $\mathbf{1}_R$, hence DNNs can approximate a non-smooth indicator function with less additional parameters. In contrast, about the other methods without activation functions and composition, they require a larger number of parameters to approximate non-smooth structures, even though the other methods have the universal approximation property. A larger number of parameters increases the variance of estimators and worsens the performance, hence the other methods lose the optimality.

Secondly, composition of functions by multi-layer structures of DNNs can divide the difficulty of estimation for piecewise smooth functions. Namely, each sub-network of DNNs represent elements of piecewise smooth functions such as $f \in H^\beta$ and $\mathbf{1}_R$ with $R \in \mathcal{R}_{\alpha,J}$. Specifically, in the proof of Theorem 1, we provide an explicit DNN which is organized by small sub-networks for approximating piecewise smooth functions. To estimate $f^* = \sum_{m \in [M]} f^*_m \otimes \mathbf{1}_{R^*_m}$, we consider small DNNs $G_{f,m}, G_{r,m}, G_3 \in \Xi(S', B', L')$ with some S', B', L' and all $m \in [M]$, which satisfy $f^*_m \approx G_{f,m}$, $\mathbf{1}_{R^*_m} \approx G_{r,m}$, and $\left(x \mapsto \Sigma_{m \in [M]} x_m x_{M+m}\right) \approx G_3$ for $x \in \mathbb{R}^{2M}$. Then, we construct a specific DNN $\dot{f} \in \Xi_{NN,\eta}(S, B, L)$ such that

$$\dot{f} = G_3(G_{f,1}(\cdot), ..., G_{f,M}(\cdot), G_{r,1}(\cdot), ..., G_{r,M}(\cdot)),$$

and show that $\dot{f}$ can effectively approximate piecewise smooth functions. This result is obtained due to the multi-layer structure of DNNs.

We note that our result for estimating non-smooth functions does not depend on non-smoothness of the ReLU activation function itself. Some smooth activation functions, such as a sigmoid function, may obtain the similar result, since such the activation function can provide the same approximation for a step function as (5).

5.3 Related Studies for Non-Smoothness

Several studies investigate approximation and estimation for non-smooth structures. Harmonic analysis provides several methods for non-smooth structures, such as curvelets Candes and Donoho (2002); Candès and Donoho (2004) and shearlets Kutyniok and Lim (2011). While the studies provide an optimality for piecewise smooth functions on pieces with C^2 boundaries, pieces in the boundary fragment class considered in our study is more general and the harmonic-based methods cannot be optimal with the pieces (studied in Korostelev and Tsybakov (2012)). Also, a convergence rate of generalization error is not known for these methods. Studies from nonparametric statistics investigated non-smooth estimation van Eeden (1985); Wu and Chu (1993); Wolpert et al. (2011); Imaizumi et al. (2018). These works focus on different settings such as density estimation or univariate data analysis, hence their setting does not fit problems discussed here.

6 Experiments

6.1 Non-smooth Realization by DNNs

We show how the estimators by DNNs can estimate non-smooth functions. To this end, we consider the following data generating process with a piecewise linear function. Let $D = 2$, ξ be an independent Gaussian variable with a scale $\sigma = 0.5$, and X be a uniform random variable on I^2. Then, we generate n pairs of (X, Y) from (1) with a true function f^* as piecewise smooth function such that

$$f^*(x) = \mathbf{1}_{R_1}(x)(0.2 + x_1^2 + 0.1x_2) + \mathbf{1}_{R_2}(x)(0.7 + 0.01|4x_1 + 10x_2 - 9|^{1.5}), \quad (6)$$

with a set $R_1 = \{(x_1, x_2) \in I^2 : x_2 \geqslant -0.6x_1 + 0.75\}$ and $R_1 = I^2 \backslash R_1$. A plot of f in figure 2 shows its non-smooth structure.

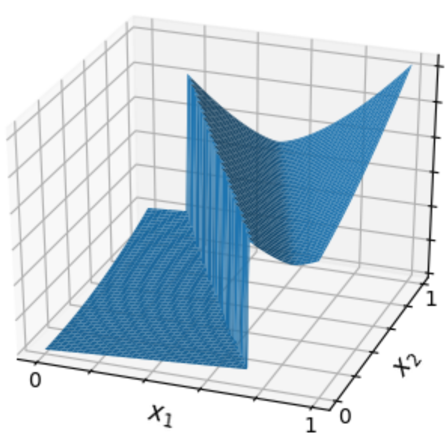

Figure 2: A plot for $f^*(x_1, x_2)$ for $(x_1, x_2) \in I^2$.

About the estimation by DNNs, we employ the least square estimator (2). For the architecture Θ of DNNs, we set $|\Theta| = 4$ and dimensionality of each of the layers as $D_1 = 2, D_\ell = 3$ for $\ell \in \{2, 3, 4\}$, and $D_5 = 1$. We use a ReLU activation. To mitigate an effect of the non-convex optimization problem, we employ 100 initial points which are generated from the Gaussian distribution with an adjusted mean. We employ Adam Kingma and Ba (2014) for optimization.

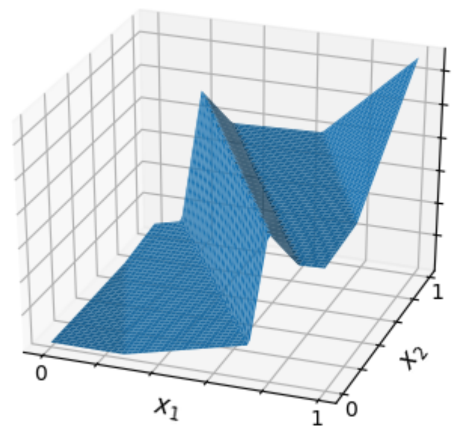

Figure 3: A plot for the estimator $\widehat{f}^L$.

We generate data with a sample size $n = 100$ f and obtain the least square estimator $\widehat{f}^L$ for f^*. Then, we plot $\widehat{f}^L$ in Figure 3 which minimize an error from the 100 trials with different initial points. We can observe that $\widehat{f}^L$ succeeds in approximating the non-smooth structure of f^*.

6.2 Comparison with the Other Methods

We compare performances of the estimator by DNNs, the orthogonal series method, and the kernel methods. About the estimator by DNNs, we inherit the setting in Section 6.1. About the kernel methods, we employ estimators by the Gaussian kernel and the polynomial kernel. A bandwidth of the Gaussian kernel is selected from $\{0.01, 0.1, 0.2, ..., 2.0\}$ and a degree of the polynomial kernel is selected from $[5]$. Regularization coefficients of the estimators are selected from $\{0.01, 0.4, 0.8, ..., 2.0\}$. About the orthogonal series method, we employ the trigonometric basis which is a variation of the Fourier basis. All of the parameters are selected by a cross-validation.

We generate data from the process (1) with (6) with a sample size $n \in \{100, 200, ..., 1500\}$ and measure the expected loss of the methods. In figure 4, we report a mean and standard deviation of a logarithm of the loss by 100 replications. By the result, the estimator by DNNs always outperforms the other estimators. The other methods cannot estimate the non-smooth structure of f^*, although some of the other methods have the universal approximation property.

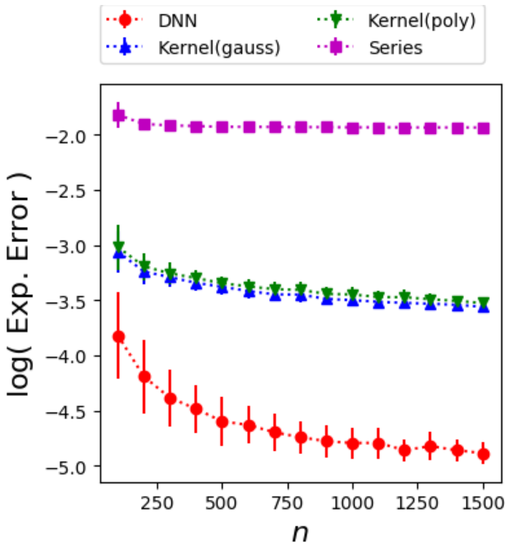

Figure 4: Comparison of errors by the methods.

7 Conclusion and Future Work

In this paper, we have derived theoretical results that explain why DNNs outperform other methods. To this goal, we considered a regression problem under the situation where the true function is piecewise smooth. We focused on the least square and Bayes estimators, and derived convergence rates of the estimators. Notably, we showed that the rates are optimal in the minimax sense. Furthermore, we proved that the commonly used orthogonal series methods and kernel methods are inefficient to estimate piecewise smooth functions, hence we show that the estimators by DNNs work better than the other methods for non-smooth functions. We also provided a guideline for selecting a number of layers and parameters of DNNs based on the theoretical results.

Investigating selection for architecture of DNNs has remained as a future work. While our results show the existence of an architecture of DNNs that achieves the optimal rate, we did not discuss how to learn the optimal architecture from data effectively. Practically and theoretically, this is obviously an important problem for analyzing a mechanism of DNNs.

Acknowledgement

We have greatly benefited from insightful comments and suggestions by Alexandre Tsybakov, Taiji Suzuki, Bharath K Sriperumbudur, Johannes Schmidt-Hieber, and Motonobu Kanagawa. MI was supported by JSPS KAKENHI Grant Number 18K18114, JST Presto Grant Number JPMJPR1852, and the Research Institute for Mathematical Sciences, a Joint Usage/Research Center located in Kyoto University.

References

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536.

Bengio, Y. and Delalleau, O. (2011). On the expressive power of deep architectures. In *Algorithmic Learning Theory*, pages 18–36. Springer.

Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. (2017). Optimal approximation with sparsely connected deep neural networks. *arXiv preprint arXiv:1705.01714*.

Candes, E. J. and Donoho, D. L. (2002). Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Annals of statistics*, pages 784–842.

Candès, E. J. and Donoho, D. L. (2004). New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. *Communications on pure and applied mathematics*, 57(2):219–266.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances* *in neural information processing systems*, pages 2933–2941.

Dudley, R. M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236.

Fakoor, R., Ladhak, F., Nazi, A., and Huber, M. (2013). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning*.

Fukumizu, K. and Amari, S.-i. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327.

Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Imaizumi, M., Maehara, T., and Yoshida, Y. (2018). Statistically efficient estimation for non-smooth probability densities. In *Artificial Intelligence and Statistics*.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Korostelev, A. P. and Tsybakov, A. B. (2012). *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media.

Kutyniok, G. and Lim, W.-Q. (2011). Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163(11):1564–1589.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 265–272. Omnipress.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Mammen, E. and Tsybakov, A. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *The Annals of Statistics*, 23(2):502–524.

Mammen, E., Tsybakov, A. B., et al. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401.

Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*.

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053.

Suzuki, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In *Artificial Intelligence and Statistics*.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation.

van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119.

van der Vaart, A. and van Zanten, J. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Annals of the Institute of Statistical Mathematics*, 37(1):461–472.

Wasserman, L. A. (2006). *All of nonparametric statistics: with 52 illustrations*. Springer.

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011). Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, 39(4):1916–1962.

Wu, J. and Chu, C. (1993). Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statistica Sinica*, pages 557–576.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *ICLR*.