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 Pakoor et al. (2013), natural language processing Collobert and Weston (2008), 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 Cybenko (1989); Barron (1993); Bengio and Delalleau (2011); Montufar et al. (2014); Yarotsky (2017); Petersen and Voigtlaender (2018); Bolcskei 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

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

where \( f \) is a \( \beta \)-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

\[
O \left( n^{-\frac{2\beta}{(2\beta+D)}} \right), \quad (n \to \infty).
\]

This is known to be a minimax optimal rate of generalization with respect to sample size \( n \) Stone (1982);
We will provide some numerical examples supporting Table 1: Architecture for DNNs which are necessary to the contributions of this paper are as follows:

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 (Theorem 3). Here, \(\alpha\) and \(\beta\) 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 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 generalization errors 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.

<table>
<thead>
<tr>
<th>ELEMENT</th>
<th>NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td># OF LAYERS</td>
<td>(O(1 + \max{\beta/D, \alpha/2(D-1)}))</td>
</tr>
<tr>
<td># OF PARAMETERS</td>
<td>(\Theta(n^{\max{D/(\beta+D), (D-1)/(\alpha+D-1)}}))</td>
</tr>
</tbody>
</table>

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 (\(q \in [0,\infty]\)). \(\text{vec}(\cdot)\) 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 and a function \(f: I \to \mathbb{R}\), \(\| f \|_{L^2(P)} := (\int_I (f(x))^2dP(x))^{1/2}\) denotes the \(L^2(P)\) norm. \(\otimes\) denotes a tensor product, and \(\otimes_{j\in J} x_j := x_1 \otimes \cdots \otimes x_J\) for a sequence \(\{x_j\}_{j\in J}\).

For a set \(R \subset I^D\), let \(1_R: I^D \to \{0,1\}\) denote the indicator function of \(R\); i.e., \(1_R(x) = 1\) if \(x \in R\), and \(1_R(x) = 0\) otherwise. Let \(H^\beta(\Omega)\) be the Hölder space on \(\Omega\) with a set \(\Omega\), which is the space of functions \(f: \Omega \to \mathbb{R}\) such that they are \([\beta]\)-times continuously differentiable and the derivatives is \([\beta]\)-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 \(I^D (D \geq 2)\) 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

\[Y_i = f^*(X_i) + \xi_i,\]

where \(f^*: I^D \to \mathbb{R}\) is an unknown true function and \(\xi_i\) 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: \(\| \hat{f} - f^* \|_{L^2(P_X)}^2 = \mathbb{E}_{X \sim P_X}[(\hat{f}(X) - f^*(X))^2]\).

There are various methods to estimate \(f^*\) and their statistical properties are extensively investigated (For summary, see Wasserman (2006) and Isybakov (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 \(\ell\)-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 \(\ell\)-th layer. The architecture \(\Theta\) of DNN
is a set of $L$ pairs of $(A_{\ell}, b_{\ell})$:
\[ \Theta := ((A_1, b_1), ..., (A_L, b_L)). \]
We define $|\Theta| := L$ be a number of layers in $\Theta$, $\|\Theta\|_0 := \sum_{\ell \in [L]} \|\text{vec}(A_{\ell})\|_0 + \|b_{\ell}\|_0$ as a number of non-zero elements in $\Theta$, and $\|\Theta\|_\infty := \max\{\max_{\ell \in [L]} \|\text{vec}(A_{\ell})\|_\infty, \max_{\ell \in [L]} \|b_{\ell}\|_\infty\}$ be the largest absolute value of the parameters in $\Theta$.

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

The model of neural networks with architecture $\Theta$ and activation $\eta$ 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
\[ x^{(1)} := x, \quad x^{(\ell+1)} := \eta(A_\ell x^{(\ell)} + b_{\ell}), \quad \text{for } \ell \in [L], \]
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 $\Theta$, 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 $D_n$, we consider the minimization problem with respect to parameters of DNNs as
\[ \hat{f}^L \in \arg\min_{f : f \in \Xi_{NN,\eta}(S, B, L)} \frac{1}{n} \sum_{i \in [n]} (Y_i - f(X_i))^2, \tag{2} \]
where $\hat{f} := \max\{\min\{f, -T\}, T\}$ 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 $f^\ast$.

Note that the problem (2) has at least one minimizer since the parameter set $\Theta$ is compact and $\eta$ 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 $\Theta$ 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 $\Theta$. Let $\Pi^{(A)}_{\ell} \text{ and } \Pi^{(b)}_{\ell}$ be distributions of $A_{\ell}$ and $b_{\ell}$ as
\[ A_{\ell} \sim \Pi^{(A)}_{\ell} \text{ and } b_{\ell} \sim \Pi^{(b)}_{\ell}, \]
for $\ell \in [L]$. We set $\Pi^{(A)}_{\ell} \text{ and } \Pi^{(b)}_{\ell}$ such that each of the $S$ parameters of $\Theta$ is uniformly distributed on $[-B, B]$, and the other parameters degenerate at 0. Using these distributions, we define a prior distribution $\Pi_{\Theta}$ on $\Theta$ by
\[ \Pi_{\Theta} := \bigotimes_{\ell \in [L]} \Pi^{(A)}_{\ell} \otimes \Pi^{(b)}_{\ell}. \]
Then, a prior distribution for $f \in \Xi_{NN,\eta}(S, B, L)$ is defined by
\[ \Pi_f(\theta) := \Pi_{\Theta}(\theta : G_\eta[\Theta] = f). \]

Then, we can obtain the posterior distribution for $f$. Since the noise $\xi_i$ is Gaussian with its variance $\sigma^2$, the posterior distribution is given by
\[ d\Pi_f(f|D_n) = \frac{\exp(-\sum_{i \in [n]} (Y_i - f(X_i))^2/\sigma^2)}{\int \exp(-\sum_{i \in [n]} (Y_i - f'(X_i))^2/\sigma^2) d\Pi_f(f')} d\Pi_f(f|D_n). \]

Finally, we define a Bayes estimator as a posterior mean
\[ \hat{f}^B := \int f d\Pi_f(f|D_n), \]
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.
We define a set of pieces in regard to notions of surfaces by Dudley (1974); Mammen et al. (1999) which is dense is an extended version of the boundary fragment class provided in Appendix A). The notion of basis pieces
\[ P \] 
transformed sphere, namely, we consider
\[ h \]
for some
\[ A \]
and
\[ \alpha \]
Similarly, we introduce a restriction for
\[ A \]
as transformed sphere, namely, we consider
\[ \Psi_{h,d} \]
for some
\[ d \]
\[ \Psi_d : I^D \to \{0,1\} \]
and a tuple
\[ (S,B,L) \]
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
\[ A_1, \ldots, A_J \]
are basic pieces.

Intuitively, \( R \in \mathcal{R}_{\alpha,J} \) is a set with piecewise \( \alpha \)-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
\[ \mathcal{F}_{M,J,\alpha,\beta} := \left\{ f_m \otimes 1_{R_m} : f_m \in H^{\beta}(I^D), R_m \in \mathcal{R}_{\alpha,J} \right\}. \]

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 \( \hat{f} \)

We investigate theoretical aspects of convergence properties of \( \hat{f} \).

**Theorem 1.** (Convergence Rate of \( \hat{f} \))

\[ f^* \in \mathcal{F}_{M,J,\alpha,\beta} \]

Then, there exist constants
\[ c_1, c'_1, C_L > 0, s \in \mathbb{N}\backslash\{1\}, T \geq \|f^*\|_{L^s}, \]
and a tuple
\[ (S,B,L) \]
satisfying
\[ \|f^* - \hat{f}\|_{L^s(P_N)} \leq C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2, \]
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 $H^\beta(D)$ (For a summary, see [41]). The second term $n^{-\alpha/(\alpha+D-1)}$ reveals an effect from estimation of $1_{R_m}$ 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 [42].

We remark that a larger number of layers decreases $B$. Considering the result by [43], 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 $\hat{f}^L \in \Xi_{NN,S}(S,B,L)$ such that

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

with a positive parameter $\Delta_n$, then the following holds:

$$\mathbb{E}_{f*} \left[ \|\hat{f}^L - f\|^2_{L^2(P_X)} \right] \leq C_L \max\{n^{-3\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\} (\log n)^2 + \Delta_n.$$

Here, $\mathbb{E}_{f*} \left[ \right]$ denotes an expectation with respect to the true distribution of $(X,Y)$. Applying results on the magnitude of $\Delta$ (e.g. [44]), 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 $c_2, c_2', C_B > 0, s \in \mathbb{N} \setminus \{1\}$, architecture $\Theta : |\Theta| \leq S$, $|\Theta|x \leq B$, $|\Theta| \leq L$ satisfying following conditions:

1. $S = c_2\max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}$,
2. $B \geq c_2n^s$,
3. $L \leq c_2(1 + \max\{\beta/D, \alpha/(D-1)\})$,

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

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

To provide proof of Theorem 2, we additionally apply studies for statistical analysis for Bayesian nonparametrics [45, 46]. 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 [47]. 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 $f$ is an arbitrary estimator for $f^* \in \mathcal{F}_{M,J.\alpha.\beta}$. Then, there exists a constant $C_{mm} > 0$ such that

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

Proof of Theorem 3 is deferred to the appendix, and it employs techniques in the minimax theory developed by [48] and [49], and entropy analysis for a family of sets [50].

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 formu-
Deep Neural Networks Learn Non-Smooth Functions Effectively

Suppose $e$ are two $R$ functions can approximate $Ad$ff
As 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 $\alpha D/(2\alpha + 2D - 2) \leq \beta$ holds. Then, there exist $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ such that $\hat{f} \in \{\hat{f}_L, \hat{f}^R\}$ and any $\hat{f}^{lin}$, large $n$ provides

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

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 $1_R$, $R \in \mathcal{R}_{\alpha,J}$ 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 $1_{(x \geq 0)}$ is approximated by

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

where $\eta$ is an arbitrary function which depends on $X_1, \ldots, X_n$. Various estimators are regarded as linear estimators, for examples, kernel methods, Fourier estimators, splines, Gaussian process, and others. Studies from nonparametric statistics (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 $\alpha D/(2\alpha + 2D - 2) \leq \beta$ holds. Then, there exist $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ such that $\hat{f} \in \{\hat{f}_L, \hat{f}^R\}$ and any $\hat{f}^{lin}$, large $n$ provides

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

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 $1_R$, $R \in \mathcal{R}_{\alpha,J}$ 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 $1_{(x \geq 0)}$ is approximated by

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

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

$$
1_R \approx \zeta \circ G,
$$

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 $1_R$ with $R \in \mathcal{R}_{\alpha,J}$. Specifically, in the proof of Theorem 4 we provide an explicit DNN which is organized by small sub-networks for approximating piecewise smooth functions. To estimate $f^* = \sum_{n \in [M]} f_m \otimes 1_{R_m}$, 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 $f_m \approx G_{f,m}, 1_{R_m} \approx G_{r,m},$ and $(x \mapsto \sum_{n \in [M]} x_m x_{M+m}) \approx G_3$ for $x \in \mathbb{R}^{2M}$. Then, we construct a specific DNN $\hat{f} \in \Xi_{NN,q}(S, B, L)$ such that

$$
\hat{f} = G_3(G_{f,1}(), \ldots, G_{f,M}(), G_{r,1}(), \ldots, G_{r,M}()),
$$

and show that $\hat{f}$ 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 $\zeta$.

**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 shearelets (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 invest-
tigated 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 \( D = 2 \), \( \xi \) 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

\[
\begin{align*}
   f^*(x) &= \mathbf{1}_{R_1}(x)(0.2 + x_1^2 + 0.1x_2) \\
          &\quad + \mathbf{1}_{R_2}(x)(0.7 + 0.01|x_1| + 10x_2 - 91.5), \quad (6)
\end{align*}
\]

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

We generate data with a sample size \( n = 100 \) and obtain the least square estimator \( \hat{f}_L \) for \( f^* \). Then, we plot \( \hat{f}_L \) in Figure 3 which minimize an error from the 100 trials with different initial points. We can observe that \( \hat{f}_L \) 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 \( \{2\} \) 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.
Deep Neural Networks Learn Non-Smooth Functions Effectively

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


in neural information processing systems, pages 2933–2941.


van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative


