数理解析研究所講究録 第2108巻 2019年 177-186



# 統計理論に基づく深層学習の原理解析

統計数理研究所* 今泉　允聡
Masaaki Imaizumi
Institute of Statistical Mathematics

**概要**

本稿では、深層ニューラルネットワーク（DNN）が他手法より良い性能を発揮する原理を、統計理論を用いて解析した。DNN は既存手法よりも高い性能を発揮することが経験的に知られているが、なぜその性能が発揮されるのかという原理は充分には解明されていない。既存の統計理論では、データが滑らかな関数から生成されている場合、多くの既存の統計・機械学習の手法が理論上の最適精度を達成することが示されており、DNN の相対的優位を説明することは難しい。本稿はその困難さを解決するため、データが非滑らかな関数から生成されている状況で各手法の汎化誤差評価を行った。具体的には、DNN による推定量の汎化誤差の収束レートを導出し、そのレートがミニマックスの意味での最適性を満たすことを示した。加えて、いくつかの既存手法がその収束レートを達成しないことを示し、DNN が他手法に理論的な優越する状況を明らかにした。さらに、DNN が最適な精度を達成するためのネットワークの構成法のガイドラインを与えた。

## 1　イントロダクション

深層ニューラルネットワーク (Deep Neural Network; DNN) によるデータ解析の性能が、近年強い注目を浴びている [26, 20]。DNN の柔軟かつ大規模なモデリングと、それらのパラメータの学習を可能にする計算機・最適化技術の発展により、DNN の分析の性能がカーネル法やサポートベクターマシンといった既存の機械学習手法を超える分析の精度を発揮することが経験的に知られている [14, 19, 17]。これらの性能により、DNN は多様な分野のタスクに活用されている。具体的には、画像解析 [13]、医療データ解析 [10]、自然言語処理 [6] などがある。

DNN の成功は注目されているが、その性能を充分に説明できる理論の構築は未だ発展途上である。そのため、DNN によるデータ解析が成功する場面の特定や、最適なパラメータ選択やネットワークの構成方法といった効率的な運用方法には、まだ改善の余地が残されている。こういった DNN の理論的な課題に対しては、様々な理論的な側面から研究が進められている。例えば、ネットワークの表現能力を調べる近似理論 [7, 2, 4, 22, 34, 24, 5]、汎化性能を調べる統計的学習理論 [3, 23, 27, 35, 30]、効率的な学習手法や学習のダイナミクスを模索する最適化理論 [1, 11, 8, 16, 28] などによる研究がある。

統計理論による DNN の解析は、データ生成過程の滑らかさという仮定に強く依存しており、そのもとでは DNN の性能を理論的に説明することが難しい。統計理論で解析を行う際は、独立同一分布より与えられた $n$

〒 190-0014 東京都立川市緑町 10-3

個の観測 $\{(Y_i, X_i)\}_{i=1}^n$ が

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

という関係に従うとし、この関係を表す関数 $f$ を推定する問題を考える。これを解析する際には、$f$ は $D$ 次元の入力を持つ $\beta$ 回連続微分可能な滑らかな関数であると仮定するのが一般的である。この設定のもとでは、カーネル法やガウス過程法といった多くの主要な推定量による汎化誤差が

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

という収束レートを持つことが知られている [31, 32]。この汎化誤差の収束レートはミニマックスの意味で最適であることが示されており [29, 31, 12]、滑らかな関数を考える限りでは DNN がそれらの手法を優越することを示すことはできない。

この理論的な限界を解決するため、本研究は**非滑らかな関数**を推定する問題を考えた。具体的には、観測されたデータが**区分上でのみ滑らかな関数**から生成されている状況で、DNN および他手法による推定量の汎化誤差を評価する。区分上でのみ滑らかな関数は、複数の区分上の滑らかな関数の組み合わせで表現される関数のクラスで、区分をまたぐ時に関数は微分不可能や非連続となる。この設定のもとで、本稿では DNN による推定量がもたらす汎化誤差が、対数項の影響を無視して

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

であることを示した（定理1)。なお、$\alpha$ と $\beta$ は区分内の関数および区分の境界線の滑らかさを表すパラメータで、$D$ は関数の入力変数の次元である。さらに、我々は上記の汎化誤差の収束レートがミニマックス最適なレートであることを示した（定理2)。加えて、他手法のクラスの一つである線形推定量について、これらが最適性を達成しないことから、DNN がこれらの手法に理論的に優越することを示した（系1)。線形推定量はカーネル法やガウス過程法などを含む広い推定量のクラスで、それらに対する DNN の理論的な優位性が構築された。

なお、本稿の内容は論文 [15] に準じる。本稿の全ての定理の証明は、元論文 [15] に含まれている。

### 1.1 記法

$I := [0, 1]$ を区間とし $\mathbb{N}$ を自然数とする。ベクトルの $b$ の $j$ 番目の要素 $b_j$、$\|\cdot\|_q := (\sum_j b_j^q)^{1/q}$ を $q$-ノルム、$\mathrm{vec}(\cdot)$ を行列のベクトル化作用素とする。$z \in \mathbb{N}$ について、$[z] := \{1, 2, \ldots, z\}$ を $z$ を超えない正の整数の集合とする。$I$ 上の測度 $P$ と関数 $f : I \to \mathbb{R}$ について、$\|f\|_{L^2(P)} := (\int_I |f(x)|^2 dP(x))^{1/2}$ を $L^2(P)$ ノルムとする。$\otimes$ はテンソル積を表す。集合 $R \subset I^D$ について、$\mathbf{1}_R : I^D \to \{0, 1\}$ を $R$ 上の指示関数とする；すなわち $x \in R$ の時 $\mathbf{1}_R(x) = 1$ で、それ以外の場合は $\mathbf{1}_R(x) = 0$ となる。$H^\beta(\Omega)$ を集合 $\Omega$ 上のヘルダー空間とする；関数 $f : \Omega \to \mathbb{R}$ のうち $\lfloor\beta\rfloor$ 回連続微分可能かつその導関数が $\beta - \lfloor\beta\rfloor$-ヘルダー連続であるものの空間である。ベクトル $x \in \mathbb{R}^{D'}$ に対して、$x_{-d} := (x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D)$ は $d$ 番目の要素を抜いたベクトルとする。

## 2 準備

### 2.1 回帰問題

観測された入力変数と出力変数の組より、それらを特徴付ける関数を推定する問題を**回帰問題**という。入力変数 $X_i$ を含む空間として $D$-次元の超立方体 $I^D$ $(D \geqslant 2)$ を考える。今、独立同一分布より生成された観測

値 $(X_i, Y_i) \in I^D \times \mathbb{R}$ が $i \in [n]$ について与えられているとし、またそれらのデータ生成過程は以下の関係を満たしているとする：

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

ここで、$f^* : I^D \to \mathbb{R}$ はデータ生成過程を特徴付ける真の関数（未知）であり、また $\xi_i$ は $i \in [n]$ ごとに平均 0 で分散 $\sigma^2 > 0$ の独立なガウス確率変数であるとする。また、$I^D$ 上の $X$ の周辺分布を $P_X$ とし、これらは有界かつ正の密度を持つとする。

回帰問題の目的は、観測の集合 $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]}$ から未知の関数 $f^*$ を推定することである。推定量を $\widehat{f}$ とし、その性能を $L^2(P_X)$ ノルムを用いて議論する；すなわち汎化誤差 $\|\widehat{f} - f^*\|^2_{L^2(P_X)} = \mathbb{E}_{X\sim P_X}[(\widehat{f}(X) - f^*(X))^2]$ の大きさを評価する。この様な $f^*$ を推定する問題は盛んに研究されており、カーネル法やスプライン法などによる推定量が多数開発されている（概論として [32] や [31] が詳しい）。

## 2.2 深層ニューラルネットワークモデル

**深層ニューラルネットワーク**（DNN）によって表現される統計モデルを定義する。$L \in \mathbb{N}$ を DNN の層の数とし、各 $\ell \in [L+1]$ ごとに $D_\ell \in \mathbb{N}$ を各層の内部の変数の次元とする。なお今回の設定では、モデル全体の出力は一次元を考え、$D_{L+1} = 1$ とする。また、各層ごとに $A_\ell \in \mathbb{R}^{D_{\ell+1}\times D_\ell}$ と $b_\ell \in \mathbb{R}^{D_\ell}$ を行列・ベクトルの形で与えられるパラメータとする。ここで、全層のパラメータの組を合わせたものを

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

とし、これを DNN モデルの**構成**と呼ぶ。ここでは、$|\Theta| := L$ を $\Theta$ の層の数を表し、$\|\Theta\|_0 := \sum_{\ell\in[L]} \|\mathrm{vec}(A_\ell)\|_0 + \|b_\ell\|_0$ を $\Theta$ の非ゼロ要素の数、$\|\Theta\|_\infty := \max\{\max_{\ell\in[L]} \|\mathrm{vec}(A_\ell)\|_\infty, \max_{\ell\in[L]} \|b_\ell\|_\infty\}$ を $\Theta$ のパラメータの最大の絶対値を表すものとする。加えて、DNN の各層での変換に用いる活性化関数 $\eta : \mathbb{R}^{D'} \to \mathbb{R}^{D'}$ を定義する。本稿では、ReLU 活性化関数 $\eta(x) = (\max\{x_d, 0\})_{d\in[D']}$ を考える。

構成 $\Theta$ と活性化関数 $\eta$ を持つ DNN によるモデル $G_\eta[\Theta] : \mathbb{R}^{D_1} \to \mathbb{R}$ を定義する。ある $x \in I^D$ について、$G_\eta[\Theta]$ の出力を

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

とし、それは各層での変換を用いて再帰的に

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

と定義されるものとする。ただし $L = |\Theta|$ である。この DNN によって表現されるモデルの集合を、ハイパーパラメータ $S \in \mathbb{N}, B > 0$ と $L' \in \mathbb{N}$ を用いて

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

と表現する。$S$ は $\Theta$ による非ゼロパラメータの数を制約しており、これは DNN の枝の数を制約しスパースなネットワークを表現していることに等しい。$B$ は各パラメータのスケールを制約している。

### 2.3 DNN による関数の推定量

DNN によるモデルを用いて、経験損失を最小化する推定量を定義する。観測 $\mathcal{D}_n$ 上で二乗誤差を最小化する DNN モデルを

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

とし、$\widehat{f}$ を $f^*$ の推定量として用いる。この最小化問題(2)は、目的関数が連続でかつパラメータの集合 $\Theta$ がコンパクトで $\eta$ が連続であることから、少なくとも一つの最小値を持つ。

## 3 非滑らかな関数の定式化

ここでは非滑らかな関数の具体的な定式化を与える。本稿では**区分上でのみ滑らかな関数**を考える。この関数は、定義域が複数の部分集合（区分）に分割され、各区分の内部でのみ滑らかなであるような構成を持つ。この時、区分の境界線上では関数は微分不可能もしくは非連続になりうる。図1に具体例を示す。

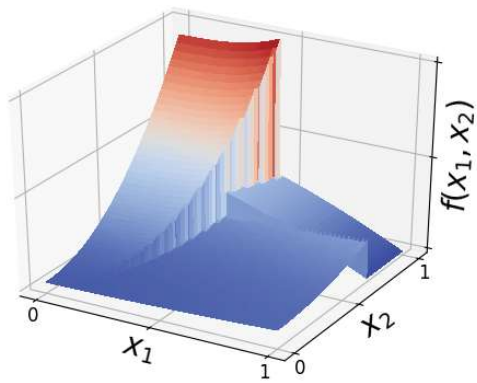


図1 区分上でのみ滑らかな関数の例。定義域は二次元の正方形で、その正方形は三つの区分に分割されている。区分の境界線上で関数が非連続になっている事が確認できる。

### 3.1 準備：区分

準備として、まず定義域 $I^D$ に含まれる区分の定義を与える。ここではホライゾン関数 [24] を用いる。滑らかな関数 $h \in H^\alpha(I^{D-1})$ を考え、その上でホライゾン関数 $\Psi_{h,d} : I^D \to \{0,1\}$ を $d \in [D]$ それぞれについて

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

と定義する。ここで、$\Psi_d : I^D \to \{0,1\}$ は $\Psi_d(x) = \mathbf{1}_{\{x \in I^D | x_d \geqslant 0\}}$ で定義されるヘヴィサイド関数である。

ホライゾン関数を用いて基底区分 $A \subset I^D$ を定義する。$A \subset I^D$ が基底区分であるとは、$\Psi_{h,d}$ が存在して

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

を満たすこととする。基底区分は、超立方体のうち $h$ を境界とした集合の片側であると見なすことができる。なお、基底区分 $A$ は球体の変形で表されるものに限定する。すなわち、$\Psi_{h,d}$ はある $\alpha$-滑らかな埋め込み $e : \{x \in \mathbb{R}^D \mid \|x\|_2 \leqslant 1\} \to \mathbb{R}^D$ を用いて $A = I^D \cap \mathrm{Image}(e)$ と表すことができるものとする（詳細は [15] の Appendix を参照）。

本稿で用いる**区分**を $J$ 個の基底区分の共通部分として定義する。すなわち、区分の集合を基底区分 $A_1, ..., A_J$ を用いて

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

のように定義する。直感的には、区分 $R \in \mathcal{R}_{\alpha,J}$ とは複数の滑らかな境界面によって囲まれる要素の集合である。また、ここでは基底区分の $J$ 個の共通部分を考えているので、$R$ の境界は微分できない部分を含むことができる。

### 3.2 区分上でのみ滑らかな関数

区分上でのみ滑らかな関数を、滑らかな関数の集合 $H^\beta(I^D)$ および区分の集合 $\mathcal{R}_{\alpha,J}$ を用いて定義する。$M \in \mathbb{N}$ を $I^D$ に含まれる区分の数として、区分上でのみ滑らかな関数の集合を

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

のように定義する。ここでは $f_m(x)$ は $x \in R_m$ の場合にのみ実現するようになっており、これらの $M$ 個の和を考えることで $\mathcal{F}_{M,J,\alpha,\beta}$ は区分 $R_m$ 上の滑らかな関数 $f_m$ の組み合わせを表現している。$\mathcal{F}_{M,J,\alpha,\beta}$ に含まれる関数は、区分の境界線上で非滑らか（微分不可能や非連続）になることが確認できる。なお、$M = 1$ かつ $R_1 = I^D$ とすると $H^\beta(I^D) = \mathcal{F}_{M,J,\alpha,\beta}$ となるため、$\mathcal{F}_{M,J,\alpha,\beta}$ の構成は既存の滑らかな関数の集合を含んでいる。

## 4 主結果

推定対象の真の関数が区分上でのみ滑らかな場合の、DNN による推定量の性能を理論的に評価する。

### 4.1 DNN による推定量の汎化誤差

$\widehat{f}$ による汎化誤差は以下のように評価される。

**Theorem 1.** （$\widehat{f}$ の汎化誤差の収束レート）
$f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ とする。この時、ある定数 $c_1, c_1', C_L > 0, s \in \mathbb{N}\backslash\{1\}$ と、*DNN* のある構成 $\Theta$ で

*(i)* $\|\Theta\|_0 = c_1' \max\{n^{D/(2\beta+D)}, n^{(D-1)/(\alpha+D-1)}\}$,
*(ii)* $\|\Theta\|_\infty \geqslant c_1 n^s$,
*(iii)* $|\Theta| \leqslant c_1(1 + \max\{\beta/D, \alpha/2(D-1)\})$,

を満たすものが存在して、この構成のもとでの推定量 $\widehat{f}$ が

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

を確率 $1 - c_1 n^{-2}$ 以上で満たす。

この定理で与えられた汎化誤差の収束レートは以下のように解釈される。一つ目の項 $n^{-2\beta/(2\beta+D)}$ は、$f_m \in H^\beta(I^D)$ を $m \in [M]$ それぞれについて推定する影響を表している。このレートは、$H^\beta(I^D)$ に属する関数を推定する問題の、ミニマックス最適な汎化誤差のレートに等しい（例えば [31] が詳しい）。二つ目の項 $n^{-\alpha/(\alpha+D-1)}$ は、$\mathbf{1}_{R_m}$ を $m \in [M]$ それぞれについて推定する影響を表している。同様のレートは、滑らかな区分を持つ集合を推定する問題で得られることがある [21]。全体の収束レートは、上記の二つの項のうち影響が大きい方で表される。

一般的に、DNN による推定量(2)は非凸最適化問題の解になるため、計算機上での実装では最適化による誤差が発生する場合がある。その影響は、以下の命題で評価される。

**Proposition 1.** （最適化の影響）
定理*1*の設定のもと、最適化の出力 $\check{f} \in \Xi_{NN,\eta}(S, B, L)$ が $\Delta_n > 0$ を用いて

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

を満たすような出力を返したとする。この時、以下が成立する：

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

$\mathbb{E}_{f^*}[\cdot]$ は $(X, Y)$ を生成する分布による期待値である。最適化の誤差を評価するのは本稿の主要な関心ではないが、最適化の影響 $\Delta_n$ は統計的評価とは独立して与えることができるため（例：[16]）、それらの結果を応用することで両方の誤差を統一的に評価することが可能になる。

## 4.2 DNN による推定量の最適性

定理1で得られた結果の最適性について議論する。ここでは、統計理論で用いられる汎化誤差の収束レートのミニマックス最適性に関する議論を用いる（例えば [31] や [12] などが詳しい）。この理論は、最良の推定量を用いる状況での最大の汎化誤差の下限を与えることで、推定量が達成できる汎化誤差の理論的な限界値を評価するものである。

以下の定理は、区分上でのみ滑らかな関数 $\mathcal{F}_{M,J,\alpha,\beta}$ を推定する際のミニマックスな汎化誤差の収束レートを与えている。

**Theorem 2.** （$\mathcal{F}_{M,J,\alpha,\beta}$ 推定のミニマックス収束レート）
$\bar{f}$ を観測 $\mathcal{D}_n$ に依存する任意の推定量とする。この時、ある定数 $C_{mm} > 0$ のもとで以下が成立する：

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

この定理の導出には、ミニマックス収束レートに関係する統計理論 [33, 25] と、区分の集合の性質を扱う理論 [9, 21] を応用している。

定理2の結果より、定理1で得られた汎化誤差の収束レートは、ミニマックスな汎化誤差の収束レートに対数項の影響を除いて一致している。よって、定理1のレートはミニマックスの意味での理論限界を達成していることから、ミニマックス最適な収束レートであると見なすことができる。すなわち、区分上でのみ滑らかな関数の推定問題において、DNN による推定量は理論的な最適性を達成していると言える。

# 5　議論：なぜ DNN は他より良い？

## 5.1　他手法の非最適性

区分上でのみ滑らかな関数を推定する際の、他手法の非最適性について議論する。本稿では、以下の形式で書かれる**線形推定量**と呼ばれる推定量のクラスを考える：

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

なお、$\Upsilon_i$ は $X_1, ..., X_n$ に依存する任意の可測関数である。この推定量のクラスは、カーネル法、フーリエ法、スプライン法、ガウス過程法などの多くの推定量を含んでいる。

非滑らか関数を推定する問題について、過去の研究（[18] の 6 章）が、線形推定量が最適性を達成しないことを示している。それを用いることで、以下の結果を得ることが出来る。

**Corollary 1.**（*DNN* の理論的優位）
$\alpha D/(2\alpha + 2D - 2) \leqslant \beta$ が成立するとする。この時、ある $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ が存在し、そのもとで *DNN* による推定量 $\widehat{f}$ と任意の線形推定量 $\widehat{f}^{\mathrm{lin}}$ に関して、十分大きな $n$ のもと以下が成立する：

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

この結果は、線形推定量に含まれる推定量は最適性を達成しないため、最適性を持つ DNN による推定量を優越できないことを理論的に示している。

## 5.2　DNN の性能の直感的な説明

ここでは、DNN が最適性を得ることへの直感的な説明を与える。

第一の理由として、DNN は区分上の指示関数 $\boldsymbol{1}_R, R \in \mathcal{R}_{\alpha,J}$ を少ない数のパラメータで簡単に表現できることが挙げられる。この性質は、DNN が持つ合成関数の構造と ReLU 活性化関数の性質から導かれる。二つの ReLU 活性化関数の差はステップ関数を効率的に近似し、またステップ関数と滑らかな関数の合成は滑らかな境界を持つ集合上の指示関数を表現できる。すなわち、ステップ関数 $\boldsymbol{1}_{\{x\geqslant 0\}}$ は十分大きなパラメータ $a > 0$ のもとで

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

と近似することが可能である。また適切な区分 $R \in \mathcal{R}_{\alpha,J}$ は、滑らかな関数 $f \in H^\beta(I)$ を近似する DNN による関数 $f \approx G \in \Xi(S_f, B_f, L_f)$ を用いて

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

と近似できる。ここで重要なのは、$\boldsymbol{1}_R$ の近似に必要なパラメータが $S_f + 4$ 個に抑えられていることである。滑らかな関数の近似に必要な $S_f$ 個のパラメータに定数個のパラメータを加えるだけで、指示関数という非滑らかな関数を効率的に近似することが可能になっている。対照的に、線形推定量などの他手法は活性化関数や合成関数の構造を持っていないため、$\boldsymbol{1}_R$ のような関数を近似するにはより多くのパラメータを用いる必要がある。仮に他手法が普遍近似性を持っていたとしても、必要なパラメータが多くなれば過適合を起こしやすくなるため、推定を行う際の汎化誤差は増大する。

第二の理由は、DNN が持つネットワークの構造が、区分上の滑らかな関数に必要な各要素の表現を分業できる点にある。すなわち、DNN の部分ネットワークが $f_m \in H^{\beta}(I^D)$ や $\boldsymbol{1}_R, R \in \mathcal{R}_{\alpha,J}$ といった各要素をそれぞれ近似し、また別の部分ネットワークがそれらの合成や積を表現できる。定理1の証明では、区分上でのみ滑らかな関数を近似するための DNN の具体的なネットワーク構造が与えられているが、それは小さな部分ネットワークの適切な組み合わせで構成されている。具体的には、$f^* = \sum_{m\in[M]} f^*_m \otimes \boldsymbol{1}_{R^*_m}$ を表現するために、小さな部分ネットワークによるモデル $G_{f,m}, G_{r,m}, G_3 \in \Xi(S', B', L')$ を $m \in [M]$ それぞれについて適切なハイパーパラメータ $S', B', L'$ のもとで考え、それらが $f^*_m \approx G_{f,m}$, $\boldsymbol{1}_{R^*_m} \approx G_{r,m}$ や $\left(x \mapsto \Sigma_{m\in[M]} x_m x_{M+m}\right) \approx G_3$ for $x \in \mathbb{R}^{2M}$ を満たすようにする。そして、DNN 全体によるモデルを

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

となるよう構成している。この構成により、$\dot{G}$ は複雑な構造を持つ関数 $f^*$ を効率的に近似することが可能になっている。

## 6 結論

本稿では、DNN による推定量が既存の方法を優越する原理の解明を目指し、非滑らかな関数の推定問題を統計理論の側面から評価した。具体的には、データが区分上でのみ滑らかな関数から生成される状況を考え、その場合の推定量の汎化誤差の評価を行った。結果として、DNN による推定量の汎化誤差は最適性を達成するほどに小さく、最適性を達成しない他手法（線形推定量）よりも良い性能を発揮できることを示した。

## 参考文献


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

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

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

[4] Yoshua Bengio and Olivier Delalleau. On the expressive power of deep architectures. In *Algorithmic Learning Theory*, pages 18–36. Springer, 2011.

[5] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *arXiv preprint arXiv:1705.01714*, 2017.

[6] Ronan Collobert and Jason Weston. 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, 2008.

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

[8] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

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

[10] Rasool Fakoor, Faisal Ladhak, Azade Nazi, and Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning*, 2013.

[11] Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

[12] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[15] Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. *arXiv preprint arXiv:1802.04474*, 2018.

[16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[18] Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012.

[19] Quoc V Le, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 265–272. Omnipress, 2011.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

[22] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

[23] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

[24] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *arXiv preprint arXiv:1709.05289*, 2017.

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

[26] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[27] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

[28] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[29] CJ Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

[30] Taiji Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In *Artificial Intelligence and Statistics*, 2018.

[31] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009.

[32] Larry Alan Wasserman. *All of nonparametric statistics: with 52 illustrations*. Springer, 2006.

[33] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[34] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[35] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.