統計数理研究所 75 周年記念事業（2019）

関数推定の理論に基づく深層学習の原理解析

Analysis for Deep Learning by Function Estimation Theory

統計思考院 今泉 允聡（Masaaki Imaizumi）

1. はじめに

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。

2. 問題設定

非滑らかな関数による回帰問題を考える。$I = [0,1]$ とし、独立同一分布より生成された観測値の集合 $\{(X_i, Y_i) \in I^D \times \mathbb{R}\}_{i\in[n]}$ が与えられ、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：

$$Y_i = f^*(X_i) + \xi_i.$$

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 ξ_i は平均 0 で分散 $\sigma^2 > 0$ のガウスノイズであるとする。また、f^* は区分上でのみ滑らかな関数であるとする。即ち、f^* の定義域 I^D が α-Smooth な境界を持つ複数の区分に分割され、その区分の内部で f^* は β-Smooth であるとする。区分の境界線上では、f^* は非連続になりうる。

観測の集合 $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]}$ による f^* の推定量を考える。DNN によるモデル Ξ_{NN} を用いて、経験リスクを最小化する最小二乗推定量を

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

と定義し、この関数 $\widehat{f}$ を f^* の推定量として用いる。

3. 結果

3.1 DNN による汎化誤差の評価

$\widehat{f}$ による汎化誤差は以下のように評価される。

Theorem 1. （$\widehat{f}$ による汎化誤差）
ある定数 $c_1, C_L > 0$ と DNN のあるネットワークのもとでの推定量 $\widehat{f}$ が

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

を確率 $1 - c_1 n^{-2}$ 以上で満たす。

収束レートのうち、一つ目の項 $n^{-2\beta/(2\beta+D)}$ は、各区分内部の f^* の滑らかさな部分を推定する影響、二つ目の項 $n^{-\alpha/(\alpha+D-1)}$ は各区分そのものを推定する影響を表現している。

3.2 DNN の最適性

定理 1 で得られた結果の最適性を議論するため、区分上でのみ滑らかな関数 f^* を推定する際のミニマックスな収束レートを導出する。

Theorem 2. （区分上でのみ滑らかな関数推定のミニマックスレート）
$\bar{f}$ を $\mathcal{D}_n$ 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：

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

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。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。

3.3 DNN と他手法の比較

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：

$$\widehat{f}^{\mathrm{lin}}(x) = \sum_{i\in[n]} \Upsilon_i(x; X_1, ..., X_n) Y_i. \tag{3.1}$$

なお、Υ_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 の理論的優位性）
$\alpha D/(2\alpha + 2D - 2) \leqslant \beta$ が成立するとする。この時、ある f^* 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：

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

この結果により、線形推定量のクラスに分類さ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が理論的に示されている。

参 考 文 献

Imaizumi, M., & Fukumizu, K. (2018). Deep Neural Networks Learn Non-Smooth Functions Effectively. arXiv preprint arXiv:1802.04474.