スパースなパラメータ空間における深層ニューラルネットワークのミニマックス最適性および優位性について

早川 知志 † 鈴木 大慈 †‡　† 東京大学 ‡ 理研 AIP

深層学習がカーネル法のような他の一般的な手法に対して優位性を示す理由を理論的に解明するために, 本研究では Gauss ノイズを伴うノンパラメトリック回帰問題に対する (ReLU) 深層学習および他の方法の性能を論じる. この方向性の既存の理論研究 [1, 2] は主に Hölder 空間や Besov 空間のようなよく知られた関数クラスの数学的理論に基づいていたが, 本研究では不連続性とスパース性を持つ関数クラスに焦点を合わせる.

問題設定. 真の関数 f° に対して Gauss ノイズ ξ_i の乗った独立同分布の入出力データ $(X_i, Y_i)_{i=1}^n$ が

$$Y_i = f^\circ(X_i) + \xi_i, \qquad i = 1, \dots, n$$

のように生成されるとき, f° を入出力データから推定する. 推定量 $\widehat{f}$ の精度は予測誤差 $\mathbb{E}\left[\|\widehat{f} - f^\circ\|^2_{L^2(\mathbb{P}_X)}\right]$ によって評価する (今回は X_i たちは $[0,1]^d$ 上の一様分布に従うとして解析している).

f° が集合 $\mathcal{F}^\circ$ 上を動くとき, ミニマックス予測誤差 $\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathbb{E}\left[\|\widehat{f} - f^\circ\|^2_{L^2(\mathbb{P}_X)}\right]$ は ($\mathcal{F}^\circ$ が大き過ぎなければ) サンプルサイズ n について何らかのレートで 0 に収束するが, 推定量を特定の形に限定してもこのレートを達成できるかかがその推定手法の性能を測る 1 つの指標となる. $\mathcal{F}^\circ$ を Hölder 空間や Besov 空間の単位球とするのが既存の一般的な設定だが, 本研究では $\mathcal{F}^\circ$ としてスパースな関数集合を採用した場合に (1) 深層学習がほぼミニマックスレートを達成することと (2) スパース性が強ければそれだけ深層学習と線形推定量との差が広がることを示した. 具体的には, スパース性を表す指標 $p > 0$ (p が 0 に近付くほどスパース性が強い) を用いて定義される関数集合 $\mathcal{K}^p$ に対して, 次の結果を得た.

主結果. $\alpha = 1/p - 1/2$ とする. 推定量全体と線形推定量に対して

$$\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{K}^p} \mathbb{E}\left[\|\widehat{f} - f^\circ\|^2_{L^2(\mathbb{P}_X)}\right] \gtrsim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{-\frac{4\alpha^2}{2\alpha+1}}, \quad \inf_{\widehat{f}\,:\,\text{線形}} \sup_{f^\circ \in \mathcal{K}^p} \mathbb{E}\left[\|\widehat{f} - f^\circ\|^2_{L^2(\mathbb{P}_X)}\right] \gtrsim n^{-1/2}$$

が成り立つ一方, 経験誤差最小化による推定量 $\widehat{f}_{\mathrm{NN}}$ が以下をみたす ReLU ニューラルネットが存在する :

$$\sup_{f^\circ \in \mathcal{K}^p} \mathbb{E}\left[\|\widehat{f}_{\mathrm{NN}} - f^\circ\|^2_{L^2(\mathbb{P}_X)}\right] \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^3.$$

$\mathcal{K}^p$ はウェーブレットを用いて定義されているが, 上の結果はそのマザーウェーブレットの近似が容易であるようなもの (例えば Haar ウェーブレットなど) に限定されている. この問題は, ニューラルネットワークの重み共有を導入することにより解消される. そのため, 重み共有が一般的な畳み込みニューラルネットワーク (CNN) への理論の展開も期待される.

また, スパース性が強いときに線形推定量が不利になるという事実は, 今回示した「線形推定量のミニマックス予測誤差は $\mathcal{F}^\circ$ をその凸包で置き換えても変わらない」という一般的な主張によって理解される. これにより, 非凸な $\mathcal{F}^\circ$ に対しては線形推定量が不利になることが直観的に明らかとなる. これは [3] で述べられている結果のより簡明な理解にも繋がっている.

参考文献

[1] Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, to appear (arXiv:1708.06633).

[2] Suzuki, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. ICLR 2019.

[3] Imaizumi, M., & Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. AISTATS 2019, *Proceedings of Machine Learning Research*, 89, 869–878.