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 [†‡]

September 9, 2019, 統計関連学会連合大会@滋賀大学

† 東京大学 情報理工学系研究科 ‡ 理研 AIP

概要

問題意識

- 深層学習が広く高い性能を示す原理を理論解析したい
- 回帰問題に対する ReLU 深層学習の理論研究が進展中
  $\rightarrow$ しかし, これまでは各種関数空間 (Hölder, Besov, ...) 個別の性質に依存した解析であった

(Schmidt-Hieber, 2017; Imaizumi and Fukumizu, 2019; Suzuki, 2019)

本研究 (Hayakawa and Suzuki, 2019)

- 従来の理論を一般化し, 自然に現れるスパース・非凸性に着目
- 深層学習と他手法をスパースなモデルでの汎化能力で比較

deep vs. linear / shallow

Hayakawa, S., & Suzuki, T. (2019).
On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. arXiv:1905.09195.

問題設定

ノンパラメトリック回帰問題

- 真の関数 $f^\circ : [0,1]^d \to \mathbb{R}$ とノイズ ξ により

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

  で生成される i.i.d. データ $(X_i, Y_i)_{i=1}^n$ を用いて f° を推定
- X_i は一様分布, ξ_i は入力と独立な Gauss ノイズ
- f° の存在範囲 (真の関数のモデル) $\mathcal{F}^\circ$ が固定

評価基準

- 推定量 $\widehat{f}$ の $\mathcal{F}^\circ$ における最悪予測誤差

$$\sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

ミニマックスレート

- サンプルサイズ n に依存する誤差の**最も速い**収束レート

$$\inf_{\text{データ}\mapsto\widehat{f}} \sup_{f^\circ\in\mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f}-f^\circ\|_{L^2}^2\right] \sim n^{-\gamma}$$

- 例：β-Hölder 連続関数の空間 $(0<\beta\leq 1)$

$$\mathcal{F}^\circ = \left\{ f:[0,1]\to\mathbb{R} \;\middle|\; \|f\|_{C^\beta} = \sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|^\beta} \leq 1 \right\}$$

$\to$ ミニマックスレート $\sim n^{-\frac{2\beta}{2\beta+1}}$ (Tsybakov, 2008)
$\Longrightarrow$ 深層学習：$\mathrm{O}(n^{-\frac{2\beta}{2\beta+1}}(\log n)^3)$ (Schmidt-Hieber, 2017)

主結果

(1) スパース度 $\alpha > 1/2$ をもつモデル $\mathcal{F}^\circ$ に対して

$$\text{(deep)}\ \ n^{-\frac{2\alpha}{2\alpha+1}} \ll n^{-1/2}\ \ \text{(linear)}$$

↑ ミニマックス最適

(2) 線形推定量が「損」する仕組みを解明

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] = \inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F}^\circ)} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

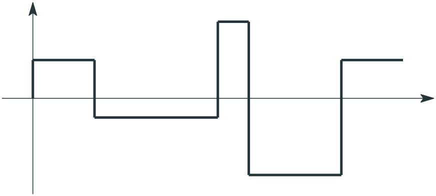
↑ $\mathcal{F}^\circ$ の凸包

$$\mathrm{conv}(\mathcal{F}^\circ) := \left\{ \sum_{i=1}^m t_i f_i \,\middle|\, m > 0,\ f_i \in \mathcal{F}^\circ,\ t_i \geq 0,\ \sum_{i=1}^m t_i = 1 \right\}$$

簡単な例

ジャンプが k 回までの空間 $\rightarrow$ スパース $(\alpha = \infty)$

$$J_k := \left\{ a_0 + \sum_{i=1}^{k} a_i 1_{[t_i,1]} \,\middle|\, t_i \in (0,1],\ |a_0|, \sum_{i=1}^{k} |a_i| \le 1 \right\}$$



▷ 凸包：有界変動関数

(deep) $\boxed{n^{-1}} \ll \boxed{n^{-1/2}}$ (linear)

推定量

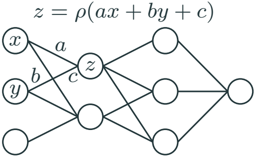
- ニューラルネットワークによる推定量
- 線形推定量

ニューラルネットワーク：非線形な活性化関数 ρ を用いて

$$\varphi_{\ell+1}(x) = \rho(W_\ell \varphi_\ell(x) - v_\ell)$$

$W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell+1}}, \ v_\ell \in \mathbb{R}^{d_\ell}$

$\rho(t) = \max\{t, 0\}$：ReLU



の形の多重合成で定まる $\to$ W_ℓ, v_ℓ の要素を学習 (勾配法 etc)

経験誤差最小化

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}_n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

$\to$ 理論解析では W, v の非ゼロ要素数を制限 (スパース正則化)

線形推定量：出力 Y_i に対して線形

$$\widehat{f}(x) = \sum_{i=1}^{n} Y_i \varphi_i(x; X_1, \dots, X_n)$$

- 例：カーネル法 (Kernel Ridge Regression) ▷ shallow

$$\widehat{f}(x) := [k(x, X_1) \cdots k(x, X_n)] \, (K_{XX} + \lambda I_n)^{-1} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

  k は半正定値カーネル, $K_{XX} = (k(X_i, X_j))_{ij}$
  (他にも Series estimator, Nadaraya-Watson estimator 等)
- f° が滑らかならミニマックス最適
  (Donoho and Johnstone, 1998; Tsybakov, 2008)

スパースな $\mathcal{F}^\circ$ の導入

弱 ℓ^p ノルム

- 数列に対する弱 ℓ^p ノルム (Donoho, 1993)

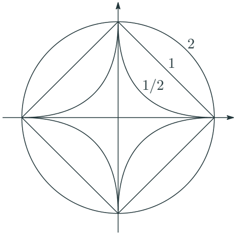
$$\|a\|_{w\ell^p} := \sup_{i=1,2,\ldots} i^{1/p}|a|_{(i)}$$

$(|a|_{(i)}$：$|a_1|, |a_2|, \ldots$ のうち i 番目に大きい値)

- 普通の ℓ^p ノルムより少し広い (が大体同じ)：

$$\|a\|_{\ell^p} := \left(\sum_{i=1}^{\infty} |a_i|^p\right)^{1/p} \leq C$$

$$\Longrightarrow \|a\|_{w\ell^p} \leq C$$



<u>スパースなモデルの導入</u> ($d=1$ で説明)

$$\mathcal{J}_\psi^{p,\beta} := \left\{ \sum_{(k,\ell)} a_{k,\ell}\psi_{k,\ell} \;\middle|\; \begin{array}{c} \|a\|_{w\ell^p} \le 1 \\ \underline{\sum_{k>m} |a_{k,\ell}|^2 \le 2^{-\beta m}} \end{array} \right\}$$

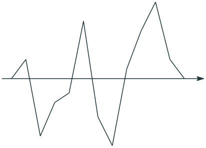
↑ 空間をコンパクトにする条件

ただし ψ は正規直交ウェーブレット ($\psi_{k,\ell}(x) = \psi(2^k x - \ell)$)

$$\mathcal{K}_\Psi^{p,\beta} := \left\{ \sum_{j=1}^J f_j(a_j x - b_j) \;\middle|\; \begin{array}{c} f_j \in \mathcal{J}_{\psi_j}^{p,\beta},\ \psi_j \in \Psi\ (\leftarrow \psi\ \text{の集合}) \\ \frac{1}{2} \le |a_j| \le 2,\ |b_j| \le 1 \end{array} \right\}$$

イメージ

基本波形 ψ



$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$\sum_{(k,\ell)} a_{k,\ell}\psi_{k,\ell} \qquad \mathcal{J}^{p,\beta}_{\psi_1} \qquad \mathcal{J}^{p,\beta}_{\psi_2}$$

$\searrow$ アフィン変換・結合 $\swarrow$

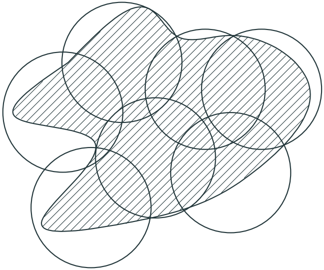
$$\Psi = \{\psi_1, \psi_2\} \qquad \mathcal{K}^{p,\beta}_{\Psi}$$

実際, 現実の画像などはウェーブレット展開すると係数がスパースになるということが知られている (e.g., Candès and Wakin, 2008)

主結果

カバリングナンバー：ある集合をいくつの ε-ball で覆えるか

$$N(\varepsilon; \mathcal{F}^\circ) = \min\{K \mid K \text{ 個の } \varepsilon\text{-ball } (L^2 \text{ 距離}) \text{ で } \mathcal{F}^\circ \text{ を被覆可能 }\}$$



▷ $\log N(\varepsilon; \mathcal{F}^\circ)$ の $\varepsilon \to 0$ でのオーダーが重要

ミニマックスレートの評価

カバリングナンバー $\leftrightarrow$ ミニマックスレート

Theorem (Yang and Barron (1999))

$\log N(\varepsilon; \mathcal{F}^\circ) \sim \varepsilon^{-1/\alpha}$ $(\varepsilon \to 0)$ ならば

$$\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] \sim n^{-\frac{2\alpha}{2\alpha+1}} \quad (n \to \infty)$$

$\mathcal{K}_\Psi^{p,\beta}$ の場合：$\alpha = 1/p - 1/2$ がスパース度

($\beta \leq 2\alpha$ ならば, $\log$ 項を除いて$\cdots$)

$\log N(\varepsilon; K_\Psi^{p,\beta}) \sim \varepsilon^{-1/\alpha}$ $\Longrightarrow \inf_{\widehat{f}} \sup_{f^\circ \in K_\Psi^{p,\beta}} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] \sim n^{-\frac{2\alpha}{2\alpha+1}}$

主結果 1：deep の最適性

Theorem

$\mathcal{K}_{\Psi}^{p,\beta}$ $(0 < p < 1,\ \beta > 1)$ に対して $\alpha = 1/p - 1/2$ とすると
($\uparrow$ スパース度 $\alpha > 1/2$ をもつパラメータ空間 $\mathcal{F}^{\circ}$)

深層学習はレート $n^{-\frac{2\alpha}{2\alpha+1}}(\log n)^3$ を達成

これはミニマックス最適, $\underbrace{\text{スパース性が強い}}_{(\alpha\text{ が大きい})}$ほど速い収束を達成

∗ 上の結果は「軽い」Ψ (e.g., Haar ウェーブレット, 多項式) に対してしか成り立たない

→ 重み共有を許せば解消・CNN の理論へ広がる可能性

∗ 線形推定量は $n^{-1/2}$ しか達成しない

$$\text{(deep)}\ \ n^{-\frac{2\alpha}{2\alpha+1}} \ll n^{-1/2}\ \ \text{(linear)}$$

主結果 2：linear が何故損をするのか

線形推定量は $\mathcal{F}^\circ$ とその凸包を区別できない：

Theorem

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] = \inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F}^\circ)} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

スパースな空間は凸包を取るとカバリングナンバーが増大

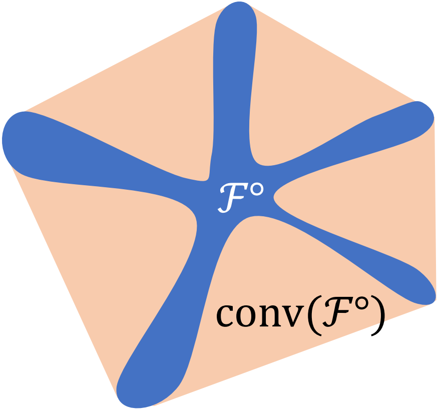
▷ deep が linear に勝つ！

ちなみに $0 < \beta \leq 1$ の範囲では…

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{K}_\Psi^{p,\beta}} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] \gtrsim n^{-\frac{\beta}{1+\beta}}$$

↑いくらでも遅くなる！

主結果 2：linear が何故損をするのか

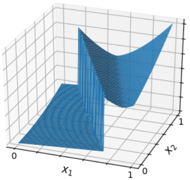


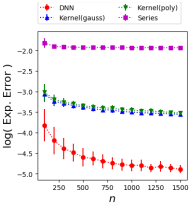
$\ast$ 主結果 1 より deep は青い部分だけを効率的に近似できる

$\leftrightarrow$ linear/shallow は凸包全体を近似しなければならない

deep が linear に勝つ理論研究

- Imaizumi and Fukumizu (2019)：区分的に滑らかな関数





- Suzuki (2019)：あるパラメータの範囲での Besov 空間
- Schmidt-Hieber (2017)：$f^\circ(x) = g(w^\top x)$ と表されるクラス

▷ これらをモデルの非凸性で統一的に説明可能！

まとめ

主結果

(1) $\mathcal{K}_{\Psi}^{p,\beta}$ $(0 < p < 1, \beta > 1)$, $\alpha = 1/p - 1/2$ に対して

$$\text{(deep)}\ \ n^{-\frac{2\alpha}{2\alpha+1}} \ll n^{-1/2}\ \ \text{(linear)}$$

↑ ミニマックス最適

(2) 線形推定量が「損」する仕組みを解明

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] = \inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F}^\circ)} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

展望

- CNN (畳み込みニューラルネットワーク) への展開
- 「正規直交ウェーブレット」の緩和

References

Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE signal processing magazine*, 25(2):21–30.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.

Hayakawa, S. and Suzuki, T. (2019). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *arXiv preprint arXiv:1905.09195*.

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

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

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

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer.

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