On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teacher-Student Setting
（教師生徒設定における勾配法による二層 ReLU ニューラルネットワークの学習可能性について）

数理情報学専攻 48-196201　秋山 俊太
指導教員　鈴木 大慈 准教授

1　はじめに

深層学習は近年，画像認識や音声認識を始めとする数多くの応用分野で高い予測精度を発揮している．一方で，その理論的な性質は十分に解明されていないのが現状である．

機械学習における基本的な学習手法として，経験誤差最小化が挙げられる．深層学習では特に仮説集合としてニューラルネットワークを考え，そのパラメータを勾配法を用いて最適化することによって学習を行う．

ニューラルネットワークのパラメータを変数とした経験誤差最小化問題は一般に非凸最適化問題であるため，一次の勾配法では停留点への収束しか保証できない．この問題に対して，近年パラメータ数を過剰に増やす (over-parameterize) ことで大域的最適解が得られやすくなることが実験的・理論的に知られてきている [2, 3]．一般的にこのようにパラメータ数を増やすとモデルが複雑になり，過学習の観点から汎化誤差を増加させる要因になると考えられる．しかし実用上そのような状況下でも汎化性能が上がることが知られていて，直観と反する結果となっている．

実際深層学習の優位性を汎化誤差の観点から解析した研究は最適化を考慮に入れていないものが多く [6]，最適化と汎化誤差理論を結びつけるような理論はまだ発展途上であるといえる．本研究では，教師生徒設定 [5]，すなわちあるニューラルネットワークを別のニューラルネットワークで学習するという枠組みの解析を行った．そして適切な正則化の下で，over-parameterize されたニューラルネットワークに関する経験誤差最小解が元のネットワークに近くなり，かつその最適解が勾配法で学習可能であることを示した．これは学習後のモデルが真のネットワークに近いものに制限されるということを意味しており，その自由度が小さくなりうることを意味する結果となっている．

2　問題設定

2.1　ニューラルネットワークの定義

本研究では，ニューラルネットワークとして二層のモデルを考える．これは入力を $x \in \mathbb{R}^d$ としたとき，パラメータ $\Theta = ((a_1, w_1), \ldots, (a_M, w_M)) \in (\mathbb{R} \times \mathbb{R}^d)^M$ を用いて，

$$f(x; \Theta) = \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle) \tag{1}$$

と書ける関数である．またここでは ReLU と呼ばれる活性化関数 $\sigma(u) = \max\{u, 0\}$ を考える．

2.2　教師生徒設定 (回帰)

ここでは回帰問題，特に入力 $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ が幅 m のニューラルネットワークによって

$$y_i = \sum_{j=1}^{m} a_j^{\circ} \sigma(\langle w_j^{\circ}, x_i \rangle) \tag{2}$$

という関係で生成されているときに，パラメータ $((a_1^{\circ}, w_1^{\circ}), \ldots, (a_m^{\circ}, w_m^{\circ}))$ を推定する問題を考える．このように，元のニューラルネットワークのパラメータを別のネットワークで推定する設定を教師生徒設定といい，元のネットワークを教師ネットワーク，学習するネットワークを生徒ネットワークと呼ぶ．さらに本研究では以下を仮定する．(1) $(x_i)_{i=1}^n$ は $\mathbb{S}^{d-1}$ 上の一様分布から i.i.d. に生成されている．(2) $\forall j, a_j^{\circ} > 0$．(3) $\forall j_1 \neq j_2, \langle w_{j_1}^{\circ}, w_{j_2}^{\circ} \rangle = 0$．この設定の下で，次のような最適化問題を考える:

$$\begin{aligned} &\min_{\Theta} F(\Theta) \\ &:= \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(x_i; \Theta) \right)^2 + \lambda \sum_{j=1}^{M} |a_j| \|w_j\|. \end{aligned} \tag{3}$$

ここで $\lambda > 0$ は正則化パラメータである．同様の正則化は [4] 等で考察されている．

3 大域的最適解における復元性

主張のために, ここではニューラルネットワークの測度表現を導入する. $\nu \in \mathcal{M}(\mathbb{S}^{d-1})$ を用いて

$$f(x;\nu) = \int_{\mathbb{S}^{d-1}} \sigma(\langle\theta, x\rangle)\mathrm{d}\nu \tag{4}$$

と書けるような関数を考える. この関数において特に $\nu = \sum_{j=1}^{M} a_j \|w_j\| \delta_{w_j/\|w_j\|}$ とすると式 (1) の形が得られることから, これは任意の有限幅の二層 ReLU ニューラルネットワークを含んだ表現であるといえる. これを測度表現と呼ぶ. このとき, 最適化問題 (3) は $\mathcal{M}(\mathbb{S}^{d-1})$ 上の最適化問題として

$$\begin{aligned} &\min_{\nu} J(\nu) \\ &:= \frac{1}{2n}\sum_{i=1}^{n} \left(y_i - f(x_i;\nu)\right)^2 + \lambda\|\nu\|_{\mathrm{TV}} \end{aligned} \tag{5}$$

と書ける. これは LASSO の測度空間への拡張とみることができ, 特に BLASSO [1] と呼ばれている.

以上の準備の下, 次のような結果を得た. 以下で $r_j^{\mathrm{o}} = a_j^{\mathrm{o}}\|w_j^{\mathrm{o}}\|, \theta_j^{\mathrm{o}} = w_j^{\mathrm{o}}/\|w_j^{\mathrm{o}}\|$ は教師ネットワークの測度表現に対応するパラメータである.

定理 1. 任意の $0 < \delta < 1$ に対して, サンプルサイズを $n > \mathrm{poly}(m, d, \log 1/\delta)$ としたとき, $1-\delta$ 以上の確率で次が成立する. 任意の $\epsilon > 0$ に対して, ある $\lambda_0 > 0$ が存在し, $\lambda < \lambda_0$ とした下で最適化問題 (5) の解は一意的に

$$\nu^* = \sum_{j=1}^{m} r_j^* \delta_{\theta_j^*} \tag{6}$$

という形で書け, さらに任意の j に対して $|r_j^* - r_j^{\mathrm{o}}| < \epsilon, \|\theta_j^* - \theta_j^{\mathrm{o}}\| < \epsilon$ を満たす.

この定理からさらに次のことが言える.

系 1. $M \geq m$ としたとき, 生徒ネットワークに関する最適化問題 (3) の解の測度表現は定理 1 と同様の性質を満たす.

つまり λ を十分小さくとることで, 教師生徒設定における最適解は教師ネットワークのパラメータに任意の距離で近付きうる.

4 勾配法による学習可能性

前節で最適化問題 (3) の大域的最適解が教師ネットワークに測度表現の意味で近付きうることを示したが, 依然としてその非凸性から勾配法でその最適解が得られるかはわからない. 本研究では次のような勾配法を考え, その大域的最適解への収束を保証した:

$$\begin{aligned} a_{j,k+1} &= a_{j,k} - \eta_{j,k}\partial_a F(\Theta_k), \\ w_{j,k+1} &= w_{j,k} - \eta_{j,k}\nabla_w F(\Theta_k), \\ \eta_{j,k} &= \alpha \frac{|a_{j,k}|\|w_{j,k}\|}{a_{j,k}^2 + \|w_{j,k}\|^2}. \end{aligned}$$

ここで $\Theta_k = ((a_{1,k}, w_{1,k}), \ldots, (a_{M,k}, w_{M,k}))$ であり, $\alpha > 0$ は定数である. 通常の勾配法と異なりステップサイズ $\eta_{j,k}$ を $a_{j,k}$ と $w_{j,k}$ のノルムに応じて定めている.

定理 2. F^* を最適化問題 (3) の最適値として, 反復中の $a_{j,k}$, $w_{j,k}$ のノルムが j,k に依らない定数 $C > 0$ で抑えられていると仮定する. 適当な初期化の下, J から定まる定数 $\alpha_0 > 0$ が存在し, $\alpha < \alpha_0$, $M > M(\alpha)$ としたとき, 定数 $\tau > 0$ があって, 高確率で $k \geq k_0 := \alpha^{-2}$ の下で

$$F(\Theta_k) - F^* \leq (1-\tau)^{k-k_0}(F(\Theta_0) - F^*) \tag{7}$$

が成立する.

この定理は一定の反復後, 関数値と最適値の差が指数的に減少していくことを意味している. M に関する不等式が over-parameterize を意味している. 最適解の一意性より, これは同時に教師ネットワークへと測度表現の意味で収束していることにほかならない.

参考文献

[1] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, Vol. 395, No. 1, pp. 336–354, 2012.

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

[3] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

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

[5] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

[6] Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.