識別問題に対する高次元二層ニューラルネットの勾配法による汎化性能解析

二反田 篤史 [†,‡]　鈴木大慈 [†,‡]
[†] 東京大学, [‡] 理研 AIP

近年，深層ニューラルネットが多大な成功を収めているが，その成功を理論的に説明する為には次の二つの大きな問題を解決する必要がある．(I) 非凸最適化問題に帰着されるニューラルネットの学習に対する最適化手法の大域収束性，(II) 訓練データに完全にフィット可能な高次元ニューラルネットに対する汎化誤差保証．本研究では滑らかな活性化関数を持つ高次元二層ニューラルネットを対象に，勾配降下法の大域収束性及び汎化誤差保証を与える．

高次元二層ニューラルネットの勾配降下法の大域収束性は種々の仮定の下で近年示されはじめている．モデルの出力スケール及び証明法に応じ平均場理論に基づくものと Neural Tangent Kernel (NTK) [1] と呼ばれるニューラルネット由来のカーネル理論に基づくものに大別されるが，本研究は特に後者の NTK を用いた理論に注目する．NTK は訓練データが完全フィット可能な場合の勾配降下法の解析に非常に有用であり，事実，ニューラルネットが高次元になる程，勾配降下法が NTK が定める再生核ヒルベルト空間における関数としての勾配降下法の近似となる事が示される．そして，この性質と高次元ニューラルネットの NTK のグラム行列の正定値性から勾配降下法の大域収束性が得られた [2]．更に, [3] ではパラメータの初期値からの移動距離の解析を通じ，勾配降下法の汎化誤差保証も与えた．これら研究は高次元ニューラルネットの最適化を理解する上で重要な貢献をしたが，非現実的な程の高次元性（超高次元ニューラルネット）を前提としていた．

そこで，本研究では識別問題においてはその問題特性から NTK 理論がより広域な二層ニューラルネットに適用可能である事を示す．具体的には NTK のグラム行列の正定値性ではなく，NTK の陽的表現である Neural Tangent Feature (NTF) を用いた線形モデルによるデータのマージン付完全識別可能性を仮定する．この仮定は識別問題では NTK のグラム行列の正定値性よりも自然であり，その結果，現実的なサイズの高次元ニューラルネットに対し勾配降下法の大域収束性と汎化誤差保証が示される．即ち，本研究は既存研究に比べてより広域な二層ニューラルネットの勾配降下法に正当性を与えるものである．以下，本研究の主定理を述べる．NTF は無限次元空間に埋め込まれたデータ特徴であり，パラメータの微分を用いて次のように定義される：$(\partial_\theta \sigma(\theta^{(0)\top} x))_{\theta^{(0)} \sim \mu_0}$．ここで σ は滑らかな活性化関数，μ_0 は二層ニューラルネットの入力層のパラメータの初期化に用いられる分布である．

定理. データ分布が *NTF* を用いた無限次元空間上の線形モデルにより L_∞-制約の下，十分なマージン付きで完全識別可能とする．任意の $\epsilon > 0$ に対し，ハイパーパラメータ *(m:* 中間ノード数，*n:* 訓練データ数，*η:* 勾配降下法の学習率*)* を次のように設定する：$m = \Omega(\epsilon^{-1})$, $\eta = \Theta(m^{-1})$, $n = \tilde{\Omega}(\epsilon^{-4})$. この時，高確率で期待 ϵ-識別誤差を達成するパラメータが勾配降下法の $T = \Theta(\epsilon^{-2})$-反復以内で得られる．

既存研究，例えば最も近い設定での関連研究 [4] でも中間ノード数 $m = \tilde{\Omega}(\epsilon^{-14})$ が必要とされた事に比べて，本定理は中間ノード数が $\Omega(\epsilon^{-1})$ という極めて現実的なサイズの二層ニューラルネットにも理論保証を与えるものであり，この点で非常に重要な結果と言える．

参考文献

[1] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589, 2018.

[2] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations 7*, 2019.

[3] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning-Volume 97*, pages 322–332, 2019.

[4] Yuan Cao and Quanquan Gu. A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384*, 2019.