2024 年度情報通信工学プログラム卒業研究発表要旨

深層ニューラルネットワークにおける学習ダイナミクスの初等的解析

60210412 宇都宮 幸大

指導教員：伊達 章 准教授

1　はじめに

膨大な数の学習可能パラメータを持つ深層ニューラルネットワークは，幅広いタスクで驚異的な性能を発揮している．こうした背景の下，中間層のニューロン数が無限大の極限における学習ダイナミクスを決定論的に記述する *Neural Tangent Kernel* (NTK) [1] や，パラメータの最適化時の更新量を保証する *Maximal Update Parametrization* (μP) [2] などの理論が発展してきた．一方，これらはすべての中間層のニューロン数が同じオーダーで増加する状況を想定しており，現実のモデルで頻繁に用いられるボトルネック構造（特定の中間層のニューロン数が相対的に著しく少ないアーキテクチャ）における学習能力については未解明の点が多い．本研究では，中間層のニューロン数が層によって異なるオーダーで増加する状況に焦点を当て，出力のダイナミクスがニューロン数に依存して消失や発散しない安定な学習を実現するパラメータ設定を提案した．また，画像分類タスクにおいて，μP の拡張である Spectral Parametrization [3] と比較し，提案手法が訓練性と汎化性能の両面で優位性があることを検証した．$*$ 卒業論文への参照を [2.2 節] のように示す．

2　学習が安定に進む条件の導出

$\boldsymbol{x} \in \mathbb{R}^{n_0}$ を入力とする L 層の深層ニューラルネットワークを以下のように再帰的に定義する：

$$\boldsymbol{f}(\boldsymbol{x}) := \widetilde{\boldsymbol{h}}^{(L)} \tag{1}$$

$$\widetilde{\boldsymbol{h}}^{(\ell)} = g_\ell \left[\boldsymbol{W}^{(\ell)} \boldsymbol{h}^{(\ell-1)} + \sqrt{n_{\ell-1}}\, \boldsymbol{b}^{(\ell)} \right] \tag{2}$$

$$\boldsymbol{h}^{(\ell)} = \psi(\widetilde{\boldsymbol{h}}^{(\ell)}) \tag{3}$$

$$\boldsymbol{h}^{(0)} = \boldsymbol{x}. \tag{4}$$

ただし，$\ell \in [L]$ は層番号である．また，重み行列 $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ とバイアス項 $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{n_\ell}$ の各要素が学習可能パラメータであり，$W_{ij}^{(\ell)}$，$b_i^{(\ell)} \sim \mathcal{N}\left(0,\ \gamma_\ell^2 \sigma_\ell^2\right)$ で独立に初期化する．$\psi(\cdot)$ は活性化関数であり，ReLU 関数を用いた場合は大きさが約 $1/\sqrt{2}$ になるため，$\gamma_\ell = \sqrt{2}$ ($\gamma_L = 1$) のようにするのが良い．損失を $\mathcal{L}$ とする勾配降下法による学習において，g_ℓ，σ_ℓ の適切なオーダーを知りたい．まず，順伝播時に $|\widetilde{h}_i^{(\ell)}| = \Theta(1)$ であるための条件として $g_\ell \sqrt{n_{\ell-1}} \sigma_\ell \overset{\text{req.}}{=} \Theta(1)$，$g_L \sqrt{n_{L-1}}\, \sigma_L \overset{\text{req.}}{=} O(1)$ を得る [2.2 節]．勾配降下法によってパラメータが更新されると中間表現 $\widetilde{\boldsymbol{h}}^{(\ell)}$ が変化する．この変化 $\Delta \widetilde{\boldsymbol{h}}^{(\ell)}$ が損失 $\mathcal{L}$ の減少にニューロン数に依らず寄与するためには

$$\left| \left\langle \frac{\partial \mathcal{L}}{\partial \widetilde{\boldsymbol{h}}^{(\ell)}},\ \Delta \widetilde{\boldsymbol{h}}^{(\ell)} \right\rangle \right| \overset{\text{req.}}{=} \Theta(1) \tag{5}$$

である必要がある．$g_\ell \sqrt{n_{\ell-1}}\ \sigma_\ell \overset{\text{req.}}{=} \Theta(1)$ 等を踏まえると，$\{\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2\}_{\ell=1}^{L-1}$ が共通のオーダーである必要性が導かれる [2.3.1 項] [4]．この条件や $g_\ell \sqrt{n_{\ell-1}} \sigma_\ell \overset{\text{req.}}{=} \Theta(1)$，$g_L \sqrt{n_{L-1}}\, \sigma_L \overset{\text{req.}}{=} O(1)$ を用いると，中間層 $1 \le \ell \le L-1$ において

$$g_\ell = \Theta\left(\frac{\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2}{\sqrt{n_{\ell-1}}} \right),\quad \sigma_\ell = \Theta\left(\frac{1}{\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2} \right).$$

出力層 $\ell = L$ において

$$g_L = \Theta\left(\frac{1}{\sqrt{n_{L-1}}} \right),\quad \sigma_L = \Theta\left(\frac{1}{\|\Delta \widetilde{\boldsymbol{h}}^{(L-1)}\|_2} \right).$$

が得られる [2.3.1 項]．$\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2$ の上限はどのくらいだろうか．$t \ge 2$ 回目の最適化時の中間表現に関する勾配を計算し，最適化の進行に伴う勾配の拡大を防ぐための条件を考えると $\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2 \overset{\text{req.}}{=} O\left(\sqrt{n_{\ell-1}}\right)$ を得る [2.3.2 項]．ここで，$\{\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2\}_{\ell=1}^{L-1}$ が共通のオーダーである必要性を踏まえると，$\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2 \overset{\text{req.}}{=} O(\sqrt{n_{\min}})$ のように，中間層における最小のニューロン数 $n_{\min} := \min\left(n_1, n_2, \ldots, n_{L-1}\right)$ が上限として現れることが示唆される [2.3.2 項]．また，$g_L \sqrt{n_{L-1}}\, \sigma_L \overset{\text{req.}}{=} O(1)$ より，$\|\Delta \widetilde{\boldsymbol{h}}^{(L-1)}\|_2 \overset{\text{req.}}{=} \Omega(1)$ を得る．以上の議論をまとめると，$r \in [0,\ 1/2]$ に対して

$$\|\Delta \widetilde{\boldsymbol{h}}^{(\ell)}\|_2 = \Theta\left(n_{\min}^r\right) \tag{6}$$

で学習が安定に進むと考えられる．便宜のため，上記の $\|\Delta\widetilde{\boldsymbol{h}}^{(\ell)}\|_2$，$g_\ell$，$\sigma_\ell$ の基準を満たす設定を *Dynamical Parametrization* (DP) と呼ぶことにする．$\{n_\ell\}_{\ell=1}^{L-1}$ が同じオーダーで $\Theta(n)$ のとき，中間層で $\|\Delta\widetilde{\boldsymbol{h}}^{(\ell)}\|_2 = \Theta(n^r)$ すなわち $|\Delta\widetilde{h}_i^{(\ell)}| = \Theta(\sqrt{n^{2r}/n})$ となる．従って $n \to \infty$ の極限では $r = 1/2$ のときにのみ $|\Delta\widetilde{h}_i^{(\ell)}| = \Theta(1)$ となり，ニューロン数に依らず更新量を保つことができる．この $r = 1/2$ のときが Maximal Update Parametrization (μP) であり，$r = 0$ のときが NTK Parametrization (NTP) と呼ばれる設定である．

3　数値実験による理論の検証

Yang *et al.* (2023) は，ニューロン数が層によって異なる設定において $|\Delta h_i^{(\ell)}| = \Theta(1)$ を実現する *Spectral Parametrization* を，本研究とは別の観点から提案した [3]．Spectral Parametrization では，ニューロン数が同じオーダーでないとき，$\{\|\Delta\widetilde{\boldsymbol{h}}^{(\ell)}\|_2\}_{\ell=1}^{L-1}$ もそれぞれ異なるオーダーとなる [3.1 節]．従って，式 5 が満たされない．これについて実験で検証する．CIFAR-10 [5] から 2 つのクラスを抽出し，2 クラス分類を行う．モデルは $L = 6$，$\psi(z) = \max(0, z)$，$\gamma_\ell = \sqrt{2}$ ($\gamma_L = 1$)，$n_0 = 3072$，$n_6 = 2$，$n := n_1 = n_3 = n_5$，$n_{\min} := n_2 = n_4$ とする．実験として，n を 10^3 から 10^4 まで log-scale で 10 段階用意し，$n_{\min} = 150\, n^{1/5}$ とする．すなわち $n_{\min}$ は $\Theta(n)$ でなく $\Theta(n^{1/5})$ である．そして，勾配降下法による最適化を行い，式 5 の左辺の値を各層で記録する [3.2 節]．実験結果の抜粋を図 1 に示す．Spectral Parametization では，ニューロン数の差異が非常に大きい状況で値が低下し，式 5 が成り立たないことが確認できる．一方，DP では成り立ち，各層における最適化能力を保持することができている．

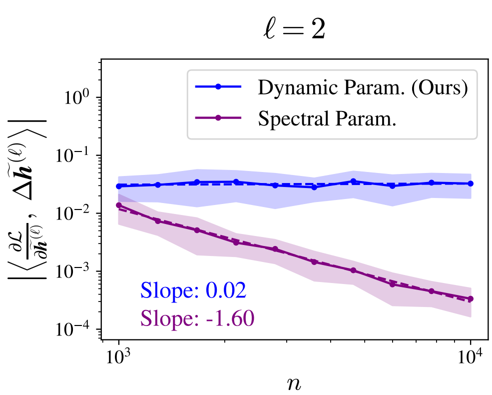

図 1: Dynamic Parametrization (DP) では各層で式 5 を実現することができる．[図 3.3] の $\ell = 2$ を抜粋．

上記の結果は，性能にも影響を与えることが期待される．実際，$n_1 = n_3 = n_5 = 10^4$，$n_2 = n_4 = 500$ のようにニューロン数の差異が非常に大きい状況で両者のパラメータ設定を比較した結果，DP の方が Spectral Parametrization に比べて訓練損失が効果的に減少することが確認できた．また，このときのテスト損失を図 2 に示す．結果より，テスト損失に関しても DP の方が低い傾向にあることがわかる．この結果は，ボトルネック構造を持つ深層ニューラルネットワークを DP で設定することで，訓練性や汎化性能を向上できる可能性を示唆している．

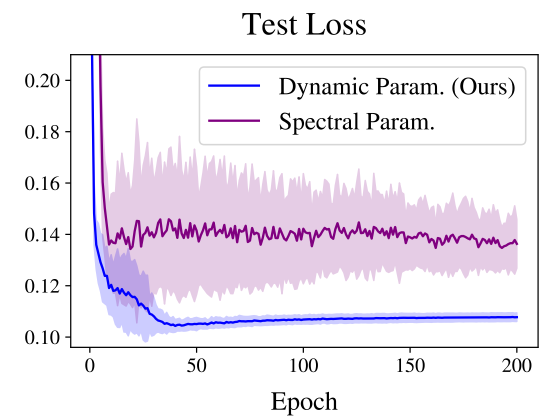

図 2: Dynamic Parametrization (DP) ではより効果的な学習が行われる．[図 3.4] の Test Loss を抜粋．

参考文献

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

[2] Greg Yang and Edward J. Hu. (2021). Tensor programs IV: Feature learning in infinite-width neural networks, *International Conference on Machine Learning*, *PMLR* **139**, 11727-11737.

[3] Greg Yang, James B. Simon, and Jeremy Bernstein. (2023). A spectral condition for feature learning, *arXiv preprint arXiv:2310.17813*.

[4] Dhruva Karkada. (2019). The lazy (NTK) and rich (μP) regimes: a gentle tutorial, *arXiv preprint arXiv:2404.19719*.

[5] Alex Krizhevsky. (2009). Learning multiple layers of features from tiny images, Technical Report, University of Toronto.