論文内容の要旨

Theory of Macroscopic Dynamics of Learning and Inference in Nonlinear Neural Networks
（非線形ニューラルネットワークにおける学習および推論の巨視的ダイナミクスの理論）

氏　名　　吉田　雄紀

2000年代後半からの深層学習技術の発展により，ニューラルネットワークは幅広い応用分野においてきわめて有用な機械学習手法となっている．具体的には，画像や音声の認識・生成・変換，翻訳や質問応答などの自然言語処理，囲碁などのゲーム，ロボティクス，その他，各学問分野で生じる予測タスク（例えばタンパク質の高次構造予測）など，あらゆる分野への応用が行われており，その多くで，従来手法よりも高い性能が深層学習技術によって実現されている．

しかしながら，ニューラルネットワークは理論的に未解明な点が多い．これには，性能の保証，推論の解釈性，最適な構造・ハイパーパラメータ，表現力，汎化性能の振舞いなど，様々な点でのブラックボックス性が挙げられるが，とりわけ本論文においては，ニューラルネットワークの学習がどのような条件の下で成功するかが明らかではない点に着目する．これは，主に二つの事由が念頭にある．一つは，1990年代にはニューラルネットワークの学習が今よりも困難と考えられていた点である．とりわけ，学習の中途において誤差の減少が長時間停滞する「プラトー現象」が問題となっていた．プラトー現象は理論的な研究が当時盛んに行われ，ニューラルネットの構造に内在する対称性から生じるパラメータ空間中の特異領域に起因することが指摘されていた．このように理論的にはニューラルネットワークの学習に不可避的に生じると考えられているプラトー現象であるが，近年のニューラルネットワークの実応用において見られることはほとんどない．すなわち，理論と実際に乖離が生じてきている．そしてもう一つは，現在においても，ニューラルネットワークの学習を最も成功させるためにはハイパーパラメータの設定を試行錯誤する必要がある点である．深層学習の周辺技術の発展に伴い，多様な手法が多数登場した一方で，ニューラルネットワークを学習させる上でのこれらの設定（ニューラルネットワークの構造，層や活性化関数の種類，パラメータの初期化方法，最適化アルゴリズムの種類やパラメータ，データの前処理方法，損失関数の設計など）の自由度は増加している．学習を成功させるためには，これらのハイパーパラメータを適切に設定することが以前にも増して重要となっているが，現状ではその多くを，ランダムサーチやグリッドサーチ等の試行錯誤をベースとした方法に頼っている．これらの二つの点を総括すると，深層学習の成功を支えている背後のメカニズムが十分に顧みられていない状況と言える．

上記の状況を動機として，本博士論文では，ニューラルネットワークの学習や推論中に生じるダイナミクスを統計力学的手法や平均場的手法により巨視的に解析し，ニューラルネットワークの学習が成功・失敗する条件を様々な観点で検討する．

本論文は，序論である第 1 章，本論である第 2–5 章，および結論である第 6 章からなる．

第 1 章では，導入的な内容を取り扱う．本論文はニューラルネットワークの理論研究に関するものであるが，それが重要である理由についてまず述べる．そして，本論文の第 2–4 章において重要な道具となる巨視的解析手法である統計力学的定式化に関して導入を行う．ニューラルネットワークは，典型的に極めて多数のパラメータを持ち，それらが学習中に勾配法にしたがって変化していくことで目的とする入出力関係を獲得することができる．この学習中のパラメータのダイナミクスを理論的に取り扱う必要があるが，非常に高次元の非線形力学系を直接扱うのは一般に困難である．一つの対処法としては，非線形ニューラルネットワークの代わりに線形ニューラルネットワークを考える方法がある．線形ニューラルネットワークの学習ダイナミクスは，一定の仮定の下で解析解を得ることができ[2]，理論的に取り扱いやすい．ただし，欠点として，非線形ニューラルネットワークならではの豊かな学習挙動（上述のプラトー現象も含まれる）が，線形ニューラルネットワークでは観察されず，解析不可能であるという点が挙げられる．一方，本論文で用いられる統計力学的手法は，1995年に Biehl et al. [3] および Saad et al. [4] により提案され，多数の微視的変数からなるダイナミクスを適切に縮約することで少数の巨視的変数のダイナミクスを得るという手法である．本手法を用いると，非線形ニューラルネットワークの学習ダイナミクスを，入力層の素子数が非常に多いという仮定の下で，少数次元の巨視的ダイナミクスに縮約することが可能である．本章では，彼らの議論をレビューすることにより，統計力学的手法の導入を行う．

第 2 章では，深層学習で用いられる正則化手法である Weight Normalization を用いた場合の非線形ニューラルネットワークの学習ダイナミクスを，統計力学的定式化を用いて解析する[5]．Weight Normalization と呼ばれる正則化手法が 2016 年に提案され[6]，学習が高速化することが経験的に知られていたが，そのメカニズムは不明であった．そこで本章では，本手法を用いた 2 層非線形ニューラルネットの学習ダイナミクスを統計力学的に解析・導出する．そして，本手法が学習係数への頑健性をもつことを明らかにし，またそのメカニズムについて論じる．

第 3 章では，非線形ニューラルネットワークの学習を困難にする「プラトー現象」に着目し，本現象がニューラルネットワークの出力素子数に依存して軽減されうることを示す[7]．ニューラルネットワークの学習中に，誤差が長時間にわたって減少せずに停滞する「プラトー現象」が知られるようになり，1990 年代後半から精力的に研究された．そして，プラトー現象の根本原因として，ニューラルネットワークのモデルに内在する対称性が，パラメータ空間の計量に特異性をもたらし，それが，非ゼロの測度の吸引領域をもつアトラクタ，いわゆるミルナーアトラクタを生み出すことが指摘されてきた[8]．しかしながら，その一方で，近年の深層学習の実応用においてプラトー現象が見られることは経験上ほとんどなく，理論と実際に乖離が生じていると言える．この乖離は，理論解析を行う上で前提としている仮定が現実的状況と一致していないことが原因のはずであるが，それらの不一致のうち何が本質的であるかは不明であり，解明されるべきと考える．本章では，出力層の素子数に着目した解析を行う．統計力学的手法を用いて非線形ニューラルネットワークの学習ダイナミクスを解析した既存研究では，いずれも出力素子数が 1 個と仮定されており，より現実的な設定である出力素子数が複数の場合が検討されていなかった．そこで本章の研究では，出力素子が複数次元ある 3 層非線形ニューラルネットワークの学習ダイナミクスを統計力学的に解析・導出することで，ネットワークの出力素子数がプラトー現象の重要な違いをもたらすことを指摘する．具体的には，複数の出力素子が存在する場合に，出力素子が 1 つの場合と比較して学習の失敗原因となるプラトー現象が軽減されることを明らかにする．

第 4 章では，非線形ニューラルネットワークの学習ダイナミクスが，学習データの統計性にいかに依存するかを解析する[9]. 第 2, 3 章で述べた研究を含め，統計力学的手法を用いて非線形ニューラルネットワークの学習ダイナミクスを解析した既存研究では，学習データ，とりわけ学習データ中の入力信号の統計性については考慮されていなかった．具体的には，統計力学的定式化による既存研究での解析では，入力信号は次元ごとに独立した標準ガウス分布から生成されるものと仮定されていた．しかしながら実際には，学習データの統計性に依存して学習挙動は変化しうると考えられ，また入力信号が高次元の等方的なガウス分布から生成されるという仮定は現実的ではない（多様体仮説によれば，現実のデータは高次元空間中に埋め込まれた低次元の多様体の近傍に集中していると考えられている[10]）．そこで本章では，学習データの統計性が学習ダイナミクスに及ぼす影響を調べるため，学習データに含まれる入力信号の統計性を一般化した場合に統計力学的枠組みを拡張する．そして，学習ダイナミクスおよびプラトー現象が，データの統計性へどのように依存するかを明らかにする．より具体的には，入力信号が従う分布の共分散行列が小さい固有値や分散した固有値をもつ場合に，プラトー現象が顕著に見られなくなることを明らかにする．

第 5 章では，第 2–4 章において非線形ニューラルネットワークの学習ダイナミクスを統計力学的手法により解析したのとは対照的に，深層ニューラルネットワークの内部を信号が伝播する際のダイナミクスを平均場の手法を用いて解析する．深層ニューラルネットワークの学習が成功するためには，多数の層にわたって信号や信号同士の相関が減衰・発散することなく伝播する必要がある．近年，このような信号伝播のダイナミクスを平均場的手法を用いて解析する試みがなされている．平均場的手法では，ランダムな重みを持つニューラルネットワークを考え，ネットワーク中の各層の幅（素子数）が無限大だと思い，各素子の活動を多変量ガウス分布で近似する．これらにより，伝播する信号の巨視的な統計性のみに関する決定論的な発展則を得ることができ，さらにそのダイナミクスの収束速度を議論することにより，信号伝播が成功する「深さスケール」と呼ばれる量を計算することができる[11]．伝播の巨視的ダイナミクスおよび深さスケールは，いくつかのシンプルなネットワークに対しては平均場的手法により解析的に求められ，さらに実際の訓練可能性によく一致するという報告がなされている．しかしながら，実用されるニューラルネットワークは典型的により複雑な構造をしており，解析的に深さスケールを求められるとは限らない．本章の研究では，任意の複雑なネットワークに対して深さスケールを数値的に評価する手法を確立し，実際に様々なタイプの構造を持つニューラルネットワークに対して深さスケールから訓練可能性を予測可能であることを示す.

最後に第 6 章では，本論文で述べた一連の研究が産業界および研究界に与えうる影響，および，今後の研究の課題や方向性について論じる．

参考文献

[1] Yuki Yoshida and Masato Okada. In Advances in NeurIPS, pages 1720–1728, 2019.

[2] Andrew M Saxe, James L McClelland, and Surya Ganguli. arXiv:1312.6120, 2013.

[3] Michael Biehl and Holm Schwarze. J. Phys. A: Math. and general, 28(3):643, 1995.

[4] David Saad and Sara A Solla. Phys. Rev. E, 52(4):4225, 1995.

[5] Yuki Yoshida, Ryo Karakida, Masato Okada, and Shun-ichi Amari. JPSJ, 86(4):044002, 2017.

[6] Tim Salimans and Diederik Kingma. In Advances in NeurIPS, 2016.

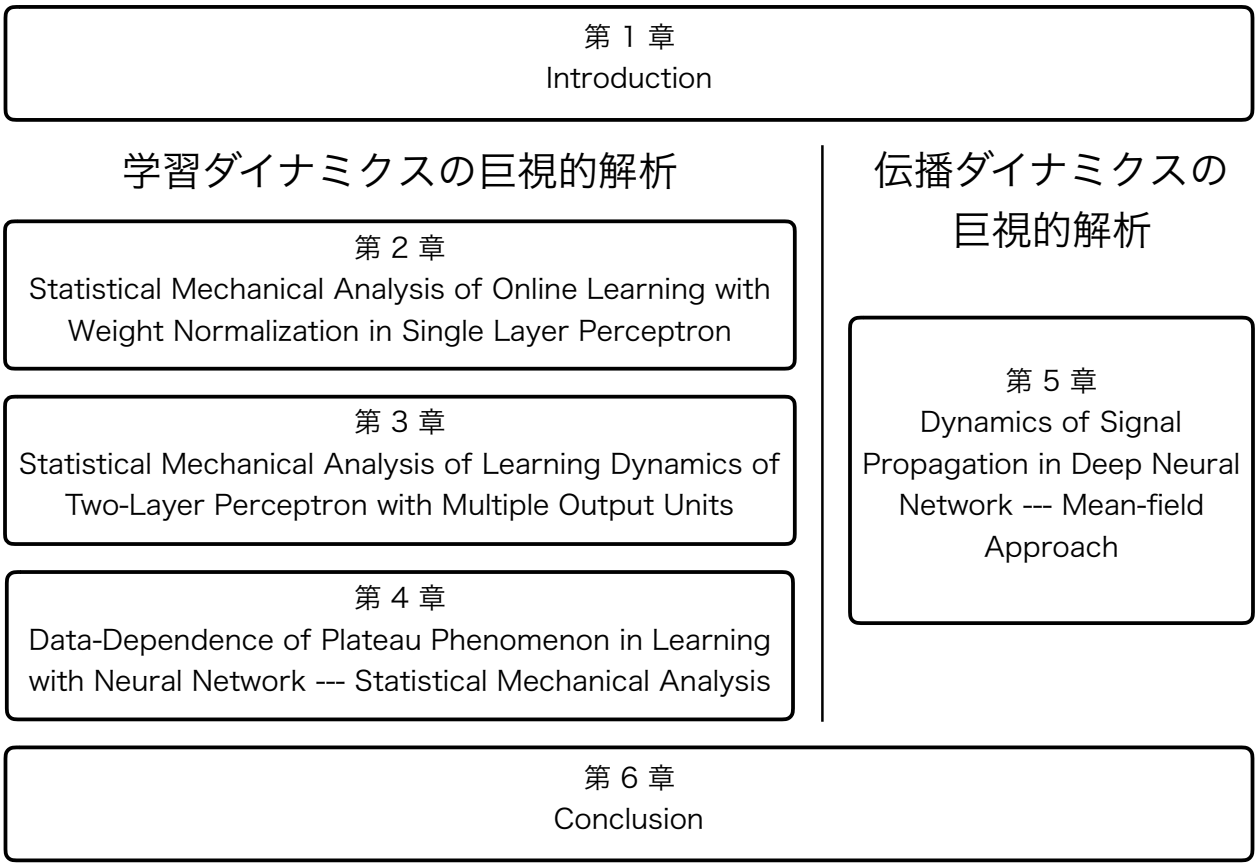
[7] Yuki Yoshida, Ryo Karakida, Masato Okada, and Shun-ichi Amari. J. Phys. A: Math. and Theoretical, 2019.

[8] Kenji Fukumizu and Shun-ichi Amari. Neural Networks, 13(3):317–327, 2000.

[9] Yuki Yoshida and Masato Okada. In Advances in NeurIPS, pages 1720–1728, 2019.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[11] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. arXiv:1611.01232, 2016.



図：本論文の構成.