研 究 終 了 報 告 書

「大自由度ニューラルネットワークの学習に潜む幾何学的構造の解析と信頼性評価への展開」

研究期間: 2019 年 10 月～2022 年 3 月
研 究 者: 唐木田 亮

1. 研究のねらい

多層のニューラルネットワークにおける機械学習は深層学習（Deep Learning）と呼ばれる。深層学習は画像認識、音声認識や言語翻訳といった典型的な情報科学のタスクにおいて、既存の機械学習手法を上回る性能を発揮している。しかし、応用研究主導での開発が進んでおり、モデルの設計やアルゴリズムの選択には多くのヒューリスティクスや恣意性があり、性能を発揮する条件やメカニズムの多くが未解明である。また、数理的な最適性や性能保証、結果の信頼性評価や解釈性が限定的である点も問題となっている。

こうした問題の解決に向けて、数理的な研究の需要が近年高まっているが、深層学習特有の難しさがある。まず、深層学習は超高次元のパラメータおよび多段階の非線形変換に基づいているため、数理的な解析が難しい。また、様々な構造のモデルや学習手法が乱立しており、個別のモデルでの個別の数理研究が進んでいる。こうした状況に対し、多様な深層学習を系統的・普遍的に解析できる数理的基盤を構築することは、性能比較の見通しをよくするために重要である。 そこで、本提案研究は、 様々なモデルや学習アルゴリズムに普遍的に適用できる数理を統計力学的解析によって構築することを目的とする。具体的には、パラメータが無限次元の極限（大自由度極限）をとることで様々なモデルで普遍的に成立する平均場理論、ランダム行列理論、さらに学習ダイナミクスの理論（Neural Tangent Kernel 法）を利用する。また、幾何学的な特徴づけを行うことで、異なるモデルや学習アルゴリズムの挙動に統一的な説明を与える。これらの数理に基づき、深層学習の課題であるヒューリスティクスや恣意性の排除、および予測の信頼性保証を目指す。

2. 研究成果

（1）概要

本研究課題では、【課題 1】Neural Tangent Kernel（NTK）法に潜む情報幾何学的構造の同定、【課題 2】学習ダイナミクスの解明、【課題 3】信頼性評価への展開，を計画し遂行した。【課題 1】は NTK の基本的な性質を調べることを目的とし、具体的な成果としては深層学習において基本的な設定となっている正規化手法（Batch normalization や Layer normalization）が与える効果を 5-(1)における主要な原著論文 1、回帰問題(平均二乗誤差ロス関数)と識別問題(クロスエントロピーロス関数)の共通点を原著論文 3 で報告した。いずれも Fisher 情報行列の固有値評価に基づいており、ロス関数あるいはパラメータ空間の曲がり方を反映している。特に、原著論文 1 は学習のハイパーパラメータである学習率に定量的な示唆を与えており、恣意性の排除にも貢献している。また原著論文 3 は【課題 3】の入力の摂動に対する頑健性を、入力次元における Fisher 情報行列の対応物を考えることである程度明らかにしている。

【課題 2】については、まず原著論文 2 で自然勾配法の収束ダイナミクスを解明した。自然勾配法はロス関数の幾何学的な歪みを補正することで高速な収束を実現するが、深層学習

で実用的に用いられているヒューリスティックな近似自然勾配も、厳密な自然勾配法と同じ速さで収束できることを NTK regime において明らかにした。この成果は国際会議 NeurIPS において採択率およそ 1%のオーラル発表に選ばれ、分野にとって重要な貢献を与える成果となったと考えらえる。また合わせて国際セミナーでの発表も行った(5-(3)参照)。このほかにも、本課題では確率的な最適化の検証をテーマにあげていたが、これに関連した成果としては、共同研究の論文成果(5-(3)参照)により、1 サンプルの条件つき Fisher 情報行列をとおしてオンライン学習初期の学習率設定に定量的な知見を与えた。

【課題 3】では少数データへの応用として NTK を使った Gaussian Process (GP) の適用をあげていたが、これについては方針転換を行った。すなわち、実際の深層学習で少数データにおいて効果を発揮する学習手法を探索することに注力し、Data Augmentation の共同研究を行い論文成果を得た。

そのほか【課題 1,2】の成果をまとめた総説記事の出版、関連する研究者と交流を深めるためのワークショップ開催、全 10 件の口頭発表を行った。

(2)詳細

【課題 1】Neural Tangent Kernel 法に潜む情報幾何学的構造の同定

成果としては 5-(1)における主要な原著論文 1 と 3 などを得た。まず、原著論文 1 は NTK に基づく学習ダイナミクスの解明の第一歩として、深層学習で広く使われている正規化手法の解析を実施した。具体的には、幅が十分に大きい深層モデルにおいて、Batch normalization (BN) が NTK および Fisher 情報行列の最大固有値を抑える仕組みを数理的に明らかにした。 初期値周辺においては、BN がロス関数の歪みを抑え、勾配法の学習率を大きくとれることが示唆され、これは BN の経験的な知見とも一致する。とくに、BN は最大固有値のネットワーク幅への依存性を軽減しており、どのような大きさ(幅)のモデルでも大きな学習率で高速な訓練の収束が期待できる。さらに、本研究では BN の幅への依存性を軽減するためなら、最終層に BN をかけるだけで十分であることを明らかにした。図 1 は実際に深層モデルを再急降下法で訓練した結果であり、左図が BN なし、右図が最終層にのみ BN の簡易版をかけた結果である。赤線は理論線である。図から分かるように、最終層での BN が幅に依存しない学習率でロス関数を大きく下げることに成功している。 なお、この効果は BN に特有であり、類似手法である Layer normalization では実現できないことも数理的に明らかにした。

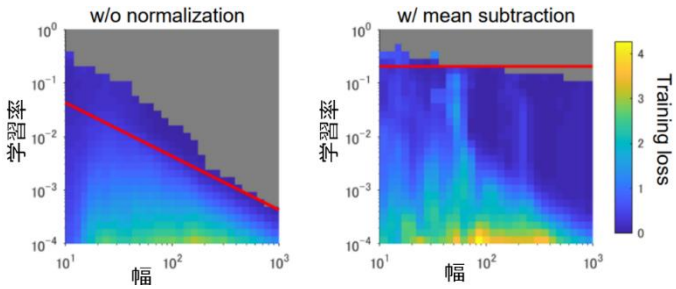

図 1: 訓練が進みやすい学習率の設定

次に原著論文 3 では、これまで研究代表者が行ってきた NTK および Fisher 情報行列の固有値の知見をレビューしつつ、識別問題として定式化したモデルでは Softmax 関数が固有値のばらつきを変調することを明らかにした。BN の効果など基本的な性質は回帰問題と共通であるが、固有値のばらつきは増大し、これは既存の深層学習の実験ともコンシステントな結果

であることを議論した。

なお、当初の計画では CNN や RNN に代表される典型的なアーキテクチャでも NTK の導出も予定していたが、この問題を解決した複数の関連論文が現れたため、上記の原著論文 1 と 3 の内容に集中する形をとった。結果的には課題 1 に対して十分な成果を上げられたといえる。また、情報幾何に関連しては、確率モデルの幾何学的な構造を反映した自然勾配法のダイナミクス解析を原著論文 2 で実施しており、これも広い意味では NTK に潜む情報幾何学的な構造の研究といえるだろう。これに関しては課題(2-2)で報告する。

【課題 2】NTK 法に基づいた学習ダイナミクスの解明

(2-1) 確率的最適化の効果

関連する成果としては 5-(3)における主要な共同研究による原著論文[Hayase & Karakida, AISTATS 2021]を得た。この研究では Dynamical isometry 下における Fisher 情報行列の固有値解析を行った。 非常に層数が多い深層モデルにおいて、逆誤差伝播の発散・消失を防ぎながら訓練するためには、アーキテクチャとパラメータ初期値に一定の条件(dynamical isometry)が必要であることが知られている。これは直交行列初期値で実現でき、逆誤差伝播の信号が定める行列の固有値分布が 1 点に集中することを利用している。 Dynamical isometry が成立するモデルで、特殊な Fisher 情報行列(サンプルごとの条件付き Fisher)をランダム行列理論/自由確率論で解析したところ、 この Fisher 情報行列でも固有値の集中が見られることが明らかとなった。 一定の条件下では固有値分布を陽に計算することができて、数理的にも興味深い。集中した固有値の値から、学習初期において確率的最適化を促進する適切な学習率(η)の上限を説明することができ、数値実験との一致もみられた(図 2; 赤線が理論線)。この条件付き行列はたとえば online 学習(すなわち、mini-batch size 1 の SGD)の学習初期のダイナミクスに現れ、ダイナミクスの収束に影響している。我々の実験結果は、online 学習の収束を決める学習率が理論によってえられた条件付き行列の最大固有値によって説明できることを示している。

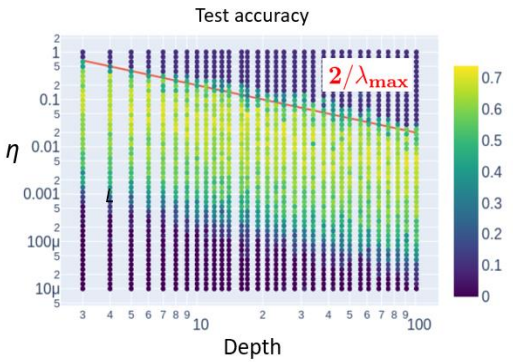

図 2: 適切な学習率における超深層モデルの訓練

(2-2) 収束を高速化するパラメータ座標系の開発

成果としては 5-(1)における主要な原著論文 2 などを得た。研究当初は自然勾配に weight normalization のような特殊な座標系の開発も想定していた。これに対し最終的には、自然勾配法における実用上もっとも重要なヒューリスティクス、Fisher 情報行列の近似、を対象として集中的に研究を遂行した。自然勾配法は Fisher 情報行列の逆行列を使ってパラメータ空間の曲がり方を考慮した勾配を使うことで、ステップ数に対して高速に訓練を進めることができる。しかしながら、深層学習ではパラメータ数が非常に多いため、逆行列計算に計算コストがか

かり、 厳密な自然勾配を利用することは難しい。そこで、Fisher 情報行列を逆行列計算が容易になるように近似したヒューリスティックな近似自然勾配法の開発が進められてきたが、数理的には近似の妥当性が未解決となっていた。こうした近似手法は自然勾配法が本来もっていた一般座標変換に対する不変性も失っており、数理的な見通しも悪い。我々は、この近似自然勾配法に対して、幅が十分に大きい深層モデルの NTK regime において解析し、深層学習で経験的に使われてきた近似アプローチが厳密な自然勾配と同一の訓練ダイナミクスを実現できることを解明した。一定の条件のもと、層ごとの対角ブロックを用いる layer-wise 近似、unit-wise 近似、さらには対角ブロックを行列積で置き換える K-FAC 近似のすべてで、NTK regime における訓練ダイナミクスは厳密な自然勾配と同じ速さで収束可能である。これらの近似自然勾配法では一般化された NTK 行列が対角行列になる”等方性条件”（図 3 上部)が成立しており、厳密な自然勾配と同様、関数空間の勾配が等方的になっている。通常の最急勾配(Gradient Descent; GD)の NTK ダイナミクスは NTK 行列の条件数に依存して収束レートが遅くなる一方、自然勾配では等方性条件の定数 α のみに依存しており、この α に対して学習率を適切にスケールすることで、高速な訓練の収束が実現できる（図 3)。

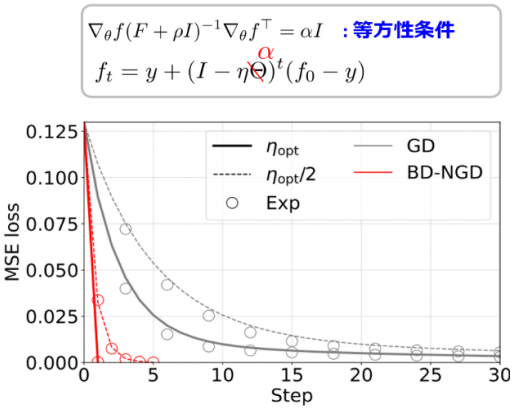

図 3: 近似自然勾配法（Layer-wise 対角ブロック近似; BD-NGD)の拘束な収束

なお、NTK regime の訓練済みモデルはカーネルリッジレス回帰と等価であり、各近似自然勾配法で異なるカーネルを用いた回帰を行っていると解釈できる。数値実験では、いずれの勾配法も同程度の汎化性能を達成しており、最も計算量が低い K-FAC 近似で非常に高速かつ汎化性能の高い解にたどりつけることが示唆される。 さらに、unit-wise 近似よりも粗い Fisher 情報行列の近似では等方性条件が必ずしも成立しないことを数値実験的に確認し、今後の自然勾配法の開発においても議論の基盤になる知見を与えた。

【課題 3】信頼性評価への展開

(3-1) 少数データへの応用

当初の予定では、少数データへの応用として NTK を使った Gaussian Process(GP)の適用をあげていたが、研究期間中に同様の方針で成果をあげた複数の関連論文が現れたため、方針転換をはかった。この課題には、通常の深層学習が少数データで高い性能を発揮するのが難しいという前提がある。この前提を確認するため、Data Augmentation によってデータ数を水増しした深層学習の性能評価を実験で行った。結果として、データ数が多い場合に比べて性能の劣化は避けられないものの、Data Augmentation は少数データの深層学習でも予測性能の向上に大きく貢献することが明らかになった(5-(3)の[Takase, Karakida & Asoh, Neurocomputing, 2021]参照)。Data Augmentation の効果を NTK regime において検証することは今後の課題であるが、その土台となる実験の知見を与えることができた。

(3-2) 入力ノイズに対する予測の耐性

成果としては、5-(1)における主要な原著論文 3 を得た。当初は adversarial example と関連づけた応用寄りの出口を視野にいれていたが、結果的にはそれよりも基礎的な数理的な成果を得た。具体的には、入力次元で Fisher 情報行列に対応した計量行列を導入し、ランダム結合のモデルでは、この行列が外れ値的に大きな固有値をクラス数個だけ持つことを示した。これは少数の次元において入力ノイズに対する耐性が低いことを示しており、adversarial attack を避けるのがランダム初期値近傍ですでに困難な可能性を示唆している。

3. 今後の展開

ACT-の研究期間中にも、人工知能技術および機械学習手法として革新的なニューラルネットワークモデルの提案および学習手法の開発は日進月歩で進んでおり、本研究で扱えなかった対象も多い。こうした革新的な手法はこれまでにない着想や方針から提案されている場合が多く、非常に高い性能を発揮する一方でメカニズムが未解明あるいは設定の恣意性が多く残っている傾向にある。まずは、今後の深層学習にとって重要となる可能性を秘めた新しい手法を本研究と同様のアプローチで解析し、理論的な知見を蓄積していくことが直近の課題として必要な展開と考えられる。

本研究では大自由度ニューラルネットワークにおける解析可能な数理モデル、別の言い方をすれば”可解モデル”をつかって体系的に学習メカニズムや性能を評価することを目的とし成果を挙げた。今後、より多くの対象を可解モデルで検証することは、対象の理解を進めるだけでなく、その可解モデルの有効範囲を明らかにするうえで重要なプロセスといえる。現実的な設定でみられる現象を各可解モデルがどのような条件下であれば再現できるのか分類を明らかにしていくことが数年から 5 年単位で必要な展開と考えらえる。

本研究成果ではヒューリスティクスのメカニズムを理解し恣意性を排除するなかで、既存の近似アルゴリズム等の設定に一定の定量的な洞察を与えたが、必ずしも手法の最適性や最良性をすべて明らかになったわけではない。たとえば自然勾配法の解析においては、近似手法が満たすのがのぞましい基準を同定したが、この基準を満たすもっとも計算コストが低い近似手法が何であるかは未解決である。また訓練が進みやすい学習率の設定も学習初期に限定されており、学習終盤までを通してのぞましいスケジューリング設定は未解明の点が多い。このように、より踏み込んだアルゴリズム開発、アーキテクチャやハイパーパラメータの選択の追究は数年から 5 年単位で将来的に必要な展開と考えらえる。

以上の展開をとおして理論と実験の行き来、ギャップの補完を進めるなかで、ニューラルネットワークの学習を理解するための数理的な基盤が整うと考えられる。これに立脚した教育や考え方を共通言語にして、より体系だって効率のよい学習手法の開発・実問題への応用を促進させていくのが 10 年単位での展開と考えられる。

4. 自己評価

研究目的の達成状況: 研究全体の総括としては、おおむね順調に進行できたといえる。特に, 課題 1-1 と課題 2-1, 2-2 は論文成果を得られたため十分な達成といえる. 課題 1-2 については学会発表で成果報告を行った点、課題 3 には直接的な解決ではないが関連手法の開発で間接的に問題の解決を与える論文成果が得られた点で、研究目標をある程度達成できたといえるだろう。

研究の進め方: 代表者主導での研究が集中して進められた点で ACT-X の枠組みにあった

研究が遂行できた。共同研究としては、外部機関の研究者や所属機関の研究者とも研究成果を挙げられている。 さらに、ACT-X の参加研究者と共同研究を行い、論文出版につなげることができた点でも、有効な研究実施体制を築けたといえるだろう。研究費執行については、新型コロナウィルス感染症の蔓延により、当初予定していた海外出張や滞在を断念せざるをえない状況となったが、代わりに計算資源を要求する課題に取り組むことで柔軟に研究費の執行に努めることができたと考えられる。

研究成果の科学技術及び社会・経済への波及効果: 大規模なニューラルネットワークは現在の人工知能・機械学習技術の中核を担っている要素のひとつであり、今後の科学技術において欠かせない技術と考えられる。本研究成果は、ヒューリスティクスが乱立し恣意性が多いこの技術の現状に対し、数理的な視点から体系だった理解と手法開発の指針を与えるものであり、今後の開発の基礎として役立つことが期待される。 将来的には数理的に自然かつ効率的な学習手法の開発へ進むことで、計算コストや資源の消費を抑える効果が期待できるとともに、社会への人工知能技術の浸透とそれに伴う経済発展を後押しできるだろう。また数理に基づきながらも常に実際の応用で観測される現象をテーマとする研究視点は ACT-X が推進する独自性に沿った研究でもあり、科学技術イノベーションにつながることも期待される。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:5件

1. Ryo Karakida, Shotaro Akaho, Shun-ichi Amari “The normalization method for alleviating pathological sharpness in wide neural networks”, In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 6406—6416, 2019.
深層学習で広く使われている normalization 手法の解析を行った. 具体的には, 幅が十分に大きい深層モデルにおいて, Batch normalization が Fisher 情報行列の最大固有値を抑え, 初期値近傍のロスのゆがみを抑えることを明らかにした. これは勾配法の学習率を大きくとることを可能にし, 安定した学習を実現できることを示唆しており, 経験的な知見とも一致する. この効果は類似手法である Layer normalization では必ずしも実現できないことも数理的に示しながら, これまでヒューリスティックに開発が進んできた normalization 手法に, 普遍的に成立するメカニズムの系統的な理解・定量化を与えた
2. Ryo Karakida & Kazuki Osawa, ″Understanding Approximate Fisher Information for Fast Convergence of Natural Gradient Descent in Wide Neural Networks″, In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 10891--10901, 2020.
深層学習では学習の高速化のために自然勾配法の開発が進んできたが, 厳密な自然勾配は計算コストが大きいため, ヒューリスティックな近似の開発が進められてきた. この近似自然勾配法に対して, 幅が十分に大きい深層モデルの NTK regime において訓練ダイナミクスの解析を行った. その結果, 深層学習で経験的に使われてきた複数の近似自然勾配が, ステップ数に対して, 厳密な自然勾配と同じ速さで収束できることが明らかとなった. いずれの

近似も関数空間での勾配が等方的になる条件を満たしており，これが NTK regime において最も早い収束を実現するために重要である.
3. Ryo Karakida, Shotaro Akaho, Shun-ichi Amari, “Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks”, Neural Computation, vol. 33(8), pp. 2274–2307, 2021.
機械学習や統計学において重要な役割を果たす Fisher 情報行列の固有値をランダム結合における深層モデルで解析した．特に本研究では識別問題に着目し，Softmax 関数に基づくクロスエントロピーロス関数が固有値に与える影響を明らかにした．典型的な活性化関数やパラメータスケールのモデルでは幅が十分に大きくなると，クラス数個だけ外れ値として大きな固有値が発生する．回帰問題と同様，特定の正規化手法で固有値のスケールをそろえることができ，対応する NTK の固有値でも同様の傾向がみられることを検証した.

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

主要な共同研究による原著論文:

- Tomohiro Hayase & Ryo Karakida, “The Spectrum of Fisher Information of Deep Networks Achieving Dynamical Isometry”, In Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), 2021.
- Tomoumi Takase, Ryo Karakida & Hideki Asoh, ″Self-paced data augmentation for training neural networks″, Neurocomputing, vol. 442, pp. 296–306, 2021.

主要な学会発表:

- Ryo Karakida, “Fisher Information of Deep Neural Networks With Random Weights”, The 11th International Chinese Statistical Association (ICSA) International Conference, 中国, 2019/12.
- Ryo Karakida & Kazuki Osawa, “Understanding Approximate Fisher Information for Fast Convergence of Natural Gradient Descent in Wide Neural Networks”, NeurIPS, オンライン, 2020/12 (オーラル発表).
- Ryo Karakida, タイトル同上, Math Machine Learning Seminar MPI MIS + UCLA, オンライン, 2021/03 (国際セミナー，招待講演).

著作物:

- 唐木田亮, “深層神経回路網の幾何 ～統計神経力学とのつながり～”（総説), 数理科学, 2020/10.