研 究 終 了 報 告 書

「柔軟な価値観を持つ機械学習のアルゴリズム開発と性能保証」

研究期間： 2020 年 11 月～2022 年 3 月
研 究 者： ホーランド マシュー ジェームズ

1． 研究のねらい

学習能力のある機械は信頼できるか？ 医療の現場や企業の採用活動の最前線に「AI」(人工知能)と呼ばれる機械学習技術が急速に浸透している今、この問いは重要性が増す一方である。ここ数年の間、機械学習の研究者は、AI 手法の信頼性向上を目指して「解釈しやすさ」や「説明可能性」を高めるための技法を盛んに開発している。しかし、これらの技法は AI が出した結果のアドホックな解釈を手助けするツールにとどまっている。また、機械学習の厳密な工学的性能保証とは完全に切り離されていることから、利用者の期待と実際の挙動の間には依然として重大な乖離があり、このツールだけでは真の信頼にはつながらない。

上記の背景を踏まえて機械学習の「工学的性能」と「利用者の認識」を分離しては、AI が確固たる信頼を得ることはできない。本研究では、学習済みの AI が出した結果ではなく、その学習過程にこそ信頼性のカギがあると考え、その過程を左右する核心的な要素を利用者の制御対象とする新たな学習アルゴリズムの開発と解析を探求している。

切り口として、右図で示すように機械学習アルゴリズムを駆動させる「評価基準」に着目する。具体的には、理論と実践の両面において、ほとんどすべての機械学習の手法は「平均的な性能の最適化」を主な学習原理としている。たとえば、理論的な性能保証はほぼ例外なく期待損失（損失値の確率分布の期待値）を基準とし、実践では学習後のテストデータを使って、性能指標の平均値に基づいて良し悪しを判断するなど、この「平均重視」はさまざまな側面から窺える。計算がしやすく、また確率論的な数理解析も容易にできるなど、平均を重視すること自体は至って自然ではあるが、これによって明確な価値判断が暗黙の裡に下されていることは広く認識されていない。平均重視の学習アルゴリズムの特徴として、顕著に成績が落ちてしまう「極端な事例」の影響を強く受ける一方、頻発する事例を疎かにしてしまう傾向がある。このような振る舞いが望まれる場合もあれば、極端なケースを無視して大半の事例を優先すべき場合も当然あるが、現在の機械学習では、このような選択はできない。

本研究の最大の狙いは、「機械学習の評価基準を選ぶこと」、そのプロセス自体を学習アルゴリズム設計の方法論に盛り込むことで、性能保証を捨てることなく、「価値判断の表層化」を合理的な形で実現することである。

2.　研究成果

研究成果について、「研究期間の総括」と「研究項目の達成」に分けて説明する。前者はACT-Xの研究期間を総合して、個人的な収穫について説明する。後者は研究計画書で掲示して項目と照らし合わせながら、より具体的な研究成果の説明に当たる。

研究期間の総括

このACT-X研究の成果を広く共有するにあたって、具体的な研究成果の説明に飛び込む前、本研究の全体を俯瞰したい。まずは背景として、私がACT-Xに応募した当初（2020年春）では、自分の研究実績のほとんどは機械学習アルゴリズムの理論的な性能保証や効率的な実装方法など、いわば「工学的な技術としての信頼性」を追求するものであった。技術を使う「ユーザー」のことは全く考慮していない。一方、社会人向けに機械学習の基礎を教える幾度の機会に恵まれ、多様なAIの利用者と交流を深めることによって、より広い「AIの信頼性」の重要性を鮮烈に認識するとともに、自分が傾倒していた理論ベースのアプローチの限界を痛感した。その結果として理論にとどまらずAIの「真の信頼性」を実現したい、という強い気持ちを抱くようになった。

この新たな研究の方向を追求する上で、より多様なバックグラウンドを持つ研究者との交流が重要なポイントとなると考えた。2020年度からJSTさきがけ「信頼されるAI」という領域が誕生することは知っていたが、自分のビジョンはまだ磨き切れておらず、さきがけではなくACT-Xに応募することとした。それでも応募時から、ACT-Xに採択された場合、最初の1年はできる限り多くの方と交流し、自分のビジョンを共有し、フィードバックをいただき、2021年度にさきがけに挑戦できるところまで自分の研究構想を増強させたい、と決心していた。

ACT-X採択後、自身が検討していたアイディアを様々な研究者と共有した。機械学習を専門とする方、それを技術として使う方、これから使ってみたいという方など、実に多様な出会いに恵まれた。特に、アドバイザーの内田先生や内田先生の「グループ」の皆様（ACT-I出身者も含む）との交流がオンラインながらも濃密で、貴重な意見、助言、エールなどを頂いた。そのなか、中途半端だった私の「評価基準」の概念を入念に練り、損失分布を制御対象とする際に注目すべき「汎化指標」の概念へと開花させた。その結果として、学習アルゴリズムの理論保証との結びつきがより明確に見えてきただけでなく、AIユーザーの日頃のワークフローとの結びつきも自ずと掴めるようになった。より堅固な土台ができたおかげで、私は自信を持って確実に自身の構想を築くことができて、スムーズにさきがけ採択までこぎ着けた。

博士課程を終えて3年となるこの時期にこの経験ができたのは、大変貴重である。この経験はACT-Xの研究課題に限らず、私がこれから取り組むすべての研究に生かせると確信しており、この1.5年の最大の成果とも言えるのではないかと考える。

研究項目の達成

当初から大きく3つの達成目標を掲げている。研究計画書で使った名称や表現をそのまま引用して簡潔に述べると、以下の3点である。

1. 解釈性と推定しやすさを両立させ、期待損失を含むAI評価基準クラスを構築する。
2. 提案基準に対して効率的な学習則を設計し、誤差の信頼区間を導出する。

3. 基準の変化が提案手法の挙動に及ぼす影響を実証的に示す。

この 3 つの達成目標の達成度合いについては、早期卒業もあって 1.5 年という短い期間であるため決して 100%ではないが、それでもどの目標も、納得できる水準まで達成できたと考える。以下では、目標別の研究成果を説明し、その関連業績も合わせて紹介する。

目標 1（評価基準クラスの構築）

この目標はこの研究課題の基軸となる重要なものである。研究開始の当初は機械学習アルゴリズムの「評価基準」という表現を使って、捉えたい性質が少し漠然としていたが、次第に「テスト損失の分布に持たせたい性質とは何か？」という問いを軸に考えるようになった。その結果、数学的に表現しやすい性質、すなわち分布の「位置」、「ばらつき」、「裾の影響」という 3 つの性質をきっちり捉えられるようなリスク関数族の設計に取り組んできた。CVaR や spectral risk、exponential smoothing など、既存の著名な「非期待値の指標」を網羅的に調査した結果、明らかな盲点を見つけることができた。その盲点を体系的に解消する指標族という位置づけを目指し、特に「ばらつきを測る関数の対称性」や「ばらつきを測る起点」に着目して、既存手法では表現しえない分布の性質を円滑に捉えられるクラスを着実に築き上げることができた（図 1 参照）。

数学的に捉えたい損失分布の性質がはっきりした結果、「評価基準」という表現をやめて、学習則の「汎化指標」と呼ぶようになった。まさに「汎化能力をどう測るか」ということが新しい機械学習の方法論を展開していく上で重要であると思い、テクニカルな進歩を経て、私の構想の基本的な概念が成熟してきた。

（関連業績）

- Learning with risks based on M-location. Matthew J. Holland. arXiv:2012.02424. Machine Learning に投稿中（査読待ち）。
- 若手研究発表賞（日本神経回路学会 2022 年度全国大会）
- Learning with risk-averse feedback under potentially heavy tails Matthew J. Holland and El Mehdi Haress. AISTATS 2021. Proceedings of Machine Learning Research 130:892-900, 2021.

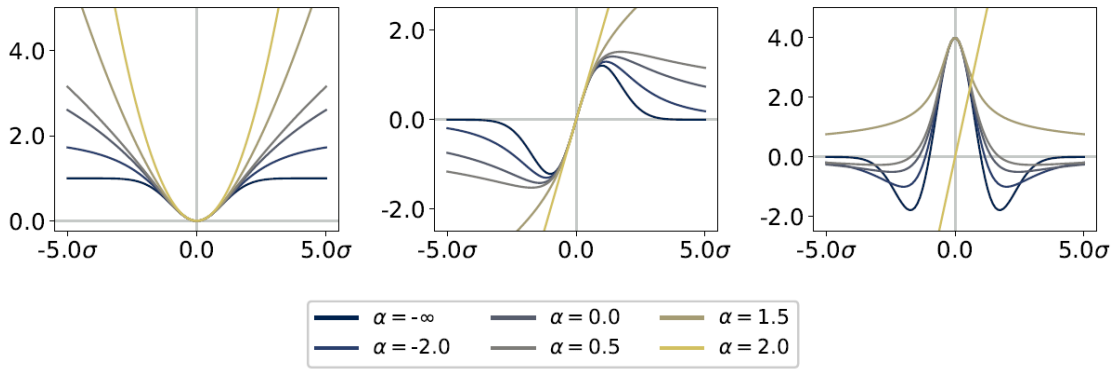

図 1

目標 2（学習則の設計と解析）

新たな汎化指標族の下、確率的最適化の「常識」がそのまま通じるケースもあれば、条件的に厳しくなることもある。今回開発している汎化指標族は「対称性のある関数でばらつきを測る」というアプローチを取っており、その結果として、元の損失関数が凸関数であっても、全体の最小化問題が凸性を有しない。加えて、微分できない点を含む汎化指標をスペシャルケースとして指標族に入れているため、non-convex かつ non-smooth な確率的最適化問題に直面する。研究を始めた当初、勾配法を使ってどこまで証明できるかは不明であったが、入念な調査を経て、機械学習で使われる多くの損失関数が持つ滑らかさと私の指標クラスの特徴を合わせると、ある種の弱い凸性を担保することができる。弱い凸性の下での勾配法の挙動は 2019 年から少しずつ明らかになってきており、分布な heavy-tailed のときに高い確率の保証をつけることはまだ難しいが、真の誤差の滑らかな近似の停留点への接近は平均的に約束できる。このような保証が可能であることは当初想定していなかったので、期待以上の結果である。また、新たに設計した汎化指標族のほかに、勾配情報が入手困難な場合の効率的な手法の開発と解析、オンラインで学習する場合の学習則の頑健性を保持させる方法など、様々な方法へ展開している。

（関連業績）

- Anytime Guarantees Under Heavy-Tailed Data. Matthew J. Holland. AAAI 2022 (to appear).
- Robustness and scalability under heavy tails, without strong convexity. Matthew J. Holland. AISTATS 2021. Proceedings of Machine Learning Research 130: 865-873, 2021.
- Spectral risk-based learning using unbounded losses. Matthew J. Holland and El Mehdi Haress. arXiv:2105.04816. AISTATS 2022 に投稿中。

目標 3（評価基準と手法挙動の実験検証）

豊かな表現力を持つ汎化指標を作り、効率的なアルゴリズムを設計したとしても、機械学習のユーザーが「どの指標をいつ選ぶべきか」ということを意識し、日頃のワークフローを推進しない限り、本研究の成果は道半ばで終わってしまう。従って、汎化指標が損失分布の変化によって、どのように変容していくか、また、汎化指標を近似したフィードバックをダイナミックな学習過程に取り入れたとき、その挙動や最終的な結果がどのような影響を受けるかなど、端的にいえばその「使い勝手」を究明していく必要がある。この目標はそれを数値実験によって定量化・可視化することを主旨とする。

実験の進捗は順調であり、丹念に設計した数値シミュレーションから実データを使った実験まで、幅広く実行し、明確な知見を蓄積しつつある。大型の深層学習モデルや超大規模なデータセットへの応用は今後の課題となるが、それでも今回提案している新しい汎化指標を導入することによって、はっきりとした影響が確認できている（図2参照）。これは当初の想定を大きく超える結果であるといえる。その理由は、汎化指標は学習の前に決めているのに、長く複雑な学習過程を経てもなお、汎化指標の特性が学習結果に鮮明に現れているからである。これは作り込んだシミュレーションではなく、標準的な実データを対象とした確率的勾配法を単純に本提案のフィードバック形成方法と組み合わせた結果であるため、現実にあるAIの現場でも、学習前に選んだ

汎化指標が非常にノイジーな学習過程を経ても影響を及ぼせることを示唆した。この実験的検証は始めたばかりで、疑問がまだたくさん残っているが、初期成果として、「AIのパフォーマンスに対する価値観」を体系的に学習過程に反映させるアプローチが見えてきている。

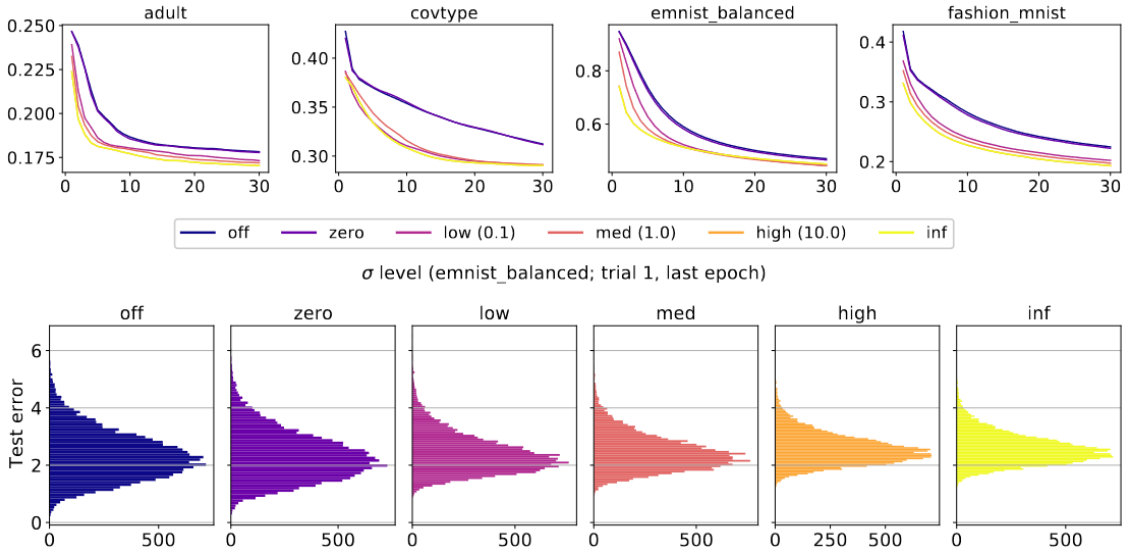

図 2

（関連業績）

- Learning with risks based on M-location. Matthew J. Holland. arXiv:2012.02424. Machine Learning に投稿中（査読待ち）。

3. 今後の展開

学問分野の開拓

汎化指標は学習の核心部分であるため、それを制御の対象とするこの試みは、基礎的な学習理論、効率的なアルゴリズムの設計、応用に当たっての工夫と実装テクニック、さらには広義の AI の透明性（解釈可能性、説明性）など、AI のあらゆる側面に関係する。本提案の課題(A)と(B)の成果を中心として、一つの「原型」を明示するとともに、未来の AI を見据えた技術的な課題や社会的に議論すべき点をあぶり出したい。一人の研究者として、これらの成果を通じて「より多様な汎化指標を考慮した機械学習」（Machine learning under diverse risks）ともいえる新しい研究分野が拓けることを期待する。

AI ワークフローの変革

本構想は、科学技術分野をはじめ、機械学習を活用するすべてのユーザーに対して、その「使い方」を大きく変えることを期待する。課題(A)と(D)によって、汎化指標という概念を定着させ、その合理的で解釈しやすい原型を提示し、ユーザーに「AIのパフォーマンス」を客観視させる。課題(B)と(C)によって、ユーザーのパフォーマンスに対する価値観をバックエンドで実現し、コーディ

ングの作業に際して可視化も含めてサポートできる。特に(C)の成果によって、「ユーザーが望むパフォーマンス」と「損失分布の性質」を、ユーザーとシステムのインタラクションによって結びつける本提案の仕組みが革新的であると考える。科学技術の現場におけるAIの試行錯誤を減らし、価値表現を明確化させることは、まさしくAIを使った研究開発プロセスそのものの透明化に貢献することを期待している。

4. 自己評価

本研究課題では「AI の性能をどう測るか」という機械学習分野への疑問を呈し、それに一つの答えを示すとともに、その答えが既存の機械学習の性能を超えることも、それを説明することも、可能であることを示唆している。研究問題の視点は高い独自性があり、また問題の解消に向けた取り組みは理論保証と実用性の両立を常に念頭に入れていることから、今後の AI 技術のさらなる発展に寄与できると考える。その上で、ユーザーがやみくもに試行錯誤をするのではなく、AIの不安定な挙動を解消するアプローチの体系化は長いスパンで考えると、全く新しい方法論に大きく貢献する可能性があると考えている。以上のことから、学術的な新規性、社会実装の可能性と AI 技術の社会的信頼性、知的財産権への発展可能性など、総合すると、この期間の研究成果を高く評価する。

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:4 件

1. Matthew J. Holland. Scaling-Up Robust Gradient Descent Techniques. Proceedings of the AAAI Conference on Artificial Intelligence 35(9) 7694-7701, 2021.
   [概要] 損失やその勾配の分布が不都合の場合の自動的な対処法の開発と解析。凸性が利用できる状況に限定されるが、データを分割して複数の安価な候補からロバストな最終候補を選ぶ手法の頑健性と学習効率を示した。
2. Matthew J. Holland. Robustness and scalability under heavy tails, without strong convexity. Proceedings of Machine Learning Research 130:865-873, 2021.
   [概要] 凸性が使えない状況下での勾配法の頑健化法の提案。従来の confidence boosting をロバストな推定量の設計によって格上げし、損失も勾配も heavy-tailed で、なおかつ目的関数の凸性がないという状況でも使える汎用的な手法の性能解析。
3. Matthew J. Holland and El Mehdi Haress. Learning with risk-averse feedback under potentially heavy tails. Proceedings of Machine Learning Research 130:892-900, 2021.
   [概要] 期待損失に変わって損失分布の CVaR を最小化する際、弱いモーメント条件の下でも成り立つ保証を持つ推定法およびそれを活かした学習則を合わせて開発・解析。
4. Matthew J. Holland. Anytime Guarantees Under Heavy-Tailed Data. Proceedings of the AAAI Conference on Artificial Intelligence (to appear, 2022/02).

［概要］ 勾配情報を計算する候補の系列と、最終的に実際に使う候補の系列が一致しないことが多いが、本提案では、それらを一致させる anytime 化の効果が高い確率で頑健に保たれるような新しい手法を設計し、理論解析を経てそのオンライン学習への応用展開の可能性を示唆。

（2）特許出願

該当なし。

（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）

1. 若手研究発表賞（日本神経回路学会 2022 年度全国大会）
2. 国際会議 ICML 2021 ワークショップ（Workshop on Distribution-Free Uncertainty Quantification）での研究発表（Making learning more transparent using conformalized performance prediction）
3. 招待講演：日本ソフトウェア科学会第 38 回大会
4. その他の国内会議発表（IBIS ワークショップ）