

2023 年度年次報告書

次世代 AI を築く数理・情報科学の革新

2023 年度採択研究代表者

熊野 創一郎

東京大学 大学院情報理工学系研究科

大学院生

拡張平均場理論を用いた敵対的訓練の理論的解析

研究成果の概要

敵対的頑健性に関する二つの研究を行った。一つは平均場理論を利用した敵対的訓練の研究である。平均場理論はニューラルネットワークの訓練可能性や表現力を解析する理論的枠組みであるが、既存の理論ではネットワークの局所的な情報遷移にしか注目できず、敵対的訓練の解析はできなかった。本研究では、既存の平均場理論を拡張した新たな平均場理論を提案し、それを利用した敵対的訓練の解析を行った。主要な結果としては、敵対的損失の経験的にタイトな上界や敵対的訓練における訓練可能性、ネットワークパラメータの時間発展などが挙げられる。また敵対的訓練ではネットワークの表現力が時間に応じて単調減少し、幅がそれを防ぐことを明らかにした。我々の結果は総じて敵対的訓練における幅の重要性を示唆しており、これは一般に層の深さが重要視されるネットワークの設計理論に新たな視点を与える。もう一つの研究では敵対的摂動からの学習という分野における未解決問題の理論的説明を試みた。敵対的摂動は人間の目にはただのノイズのように見えるが、ニューラルネットワークはそれらから学習し、汎化が可能である。この非直感的な汎化は敵対的摂動がクラスに固有な特徴を持っていることを示唆しており、敵対的摂動及び敵対的画像の起源を解明するにあたって重要な意味を持つ。しかしながら現在は実験的な証拠しか与えられていない。本研究では、敵対的摂動からの学習を一隠れ層ニューラルネットワークと直交データに対して行い、そのダイナミクスを証明した。我々の結果は入力次元とデータセットのサンプル数に依らず、敵対的摂動がクラスの特徴を持ち、ネットワークがそれらから汎化可能であることを示している。また我々の理論結果はこれまで主に L_2 距離によって制約されていた敵対的摂動とそれらからの学習が、 L_0 距離によって制約されたスパースな敵対的摂動に対しても成り立つことを証明した。

【代表的な原著論文情報】

- 1) Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Adversarial Training from Mean Field Perspective. In Conference on Neural Information Processing Systems (NeurIPS), volume 36, 2023.
- 2) Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Theoretical Explanation for Generalization from Adversarial Perturbations. In Conference on Neural Information Processing Systems (NeurIPS) Workshop on Mathematics of Modern Machine Learning (M3L), 2023.
- 3) Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Theoretical Understanding of Learning from Adversarial Perturbations. In International Conference on Learning Representations (ICLR), 2024.