公開

研 究 終 了 報 告 書

「ロバスト性と汎化性能を両立する機械学習法の確立」

研究期間: 2021 年 10 月～2024 年 3 月
研 究 者: 藤澤 将広

1. 研究のねらい

インターネットの普及やスマートデバイス等の著しい発展に伴い，膨大な情報の自動的収集・蓄積が可能となった現代の「ビッグデータ」時代では，世界中の産業・企業が，ビッグデータ解析を通した新たな技術革新を産み続けている．その革新を支えている情報工学技術の一つが機械学習であり，学習精度・適応可能タスクの急速な発展に伴って，ますますその重要性が高まっている．しかしながら，ビッグデータ収集の自動化は，予期せぬシステムエラーやユーザーによる敵対的な情報入力などによって，データ内に自然発生的・攻撃的な汚染，すなわち敵対的摂動や外れ値が混在するリスクを高める結果を産んだ．このような「汚染データ」は学習精度に大きく影響を及ぼし，医療診断や自動運転等の誤りの許されない場面で深刻な予測エラーを引き起こすため，汚染の影響を取り除く機械学習の開発が必要とされ，今日も様々な研究が報告されている．また，自動化により大量な情報の迅速な取得が可能になったことで，機械学習は，次々に得られる膨大な未知のデータに対しても高い精度を発揮する「汎化性能」が求められるようになった．優れた汎化を達成するための機械学習は，最も注目されている研究分野の一つとなり，著しい速度で研究・開発がなされている．

しかし，汎化性能とロバスト性との間には多くの文脈でトレードオフの関係，すなわち，ロバスト性を高めることで汎化性能が失われる，あるいはその逆の現象が起こり得ることが報告されている．機械学習の応用の社会浸透が加速していることを鑑みれば，この「ロバスト性」と「汎化性能」は，安心・安全な社会応用を促すためにどちらも両立すべき性能と言える．それにも拘らず，この問題に対処する研究は，限定的な汚染・タスク・モデル下でのトレードオフ緩和あるいはトレードオフ現象の存在証明など，対処療法的・現象発見的な研究に留まっている．

本研究のねらいは，汎化性能に関する理論の観点から，タスクやモデルの設定をできるだけ限定しない条件下で，ロバスト性と汎化性能とのトレードオフ関係を解析し，そのメカニズムや，解消可能性，理論限界についての示唆を与える，あるいは両性能を両立可能にするアルゴリズムを提供することにある．そして，それらを通して，**誤予測リスクと隣合わせにある実社会応用に高い信頼性・安全性を提供することが究極の目標である．**

2. 研究成果

（1）概要

前半の研究期間（研究開始～1 年半）では，PAC-Bayes 汎化誤差上界と呼ばれる，汎化性能指標の一つである汎化誤差の上界であり，上界最小化という形で学習アルゴリズムが構成可能な理論的枠組みに焦点を当てた．特に，この上界と変分推論の目的関数との関係と，最小記述長に基づく汎化の議論に注目し，汚染された事前情報に対するロバスト性と汎化性能を両立する変分推論アルゴリズムの構成を試みた．簡単な概要を以下に整理する．

公開

・初期研究：事前分布誤設定へのロバスト性と汎化性能向上を実現する経験変分ベイズ法

経験変分ベイズ法は，予測の意味で不適切な事前分布を適応的に修正することで，最小記述長原理の観点から汎化性能の獲得を目指す変分推論アルゴリズムである．この方法は計算負荷も小さく実装も容易な一方で，深層ベイズ学習においては，学習途中の低精度な事後分布に影響されることで予測性能を低下させると言われている．当初の研究では，この方法に「事後分布の情報を信用しすぎない」工夫を施すことで，近年の高度な事前分布修正法と遜色ない性能を達成することを示した．さらに，経験変分ベイズ法が目的関数の値を増大させてしまう現象に着目し，これを回避する新しい経験変分ベイズ法を提案した．

・初期研究の失敗と考察，研究方針の転換

現状では，この初期研究は失敗に終わっている．理由としては，特に深層学習では，汎化性能の基準としてよく用いられる PAC-Bayes 上界における最小記述長，つまりモデルの複雑度指標と損失関数の良いトレードオフを達成することによる上界値の小ささと汎化性能との相関に基づく議論では必ずしも良い汎化性能を保証し得ないことが報告されており，これまで**一般的とされてきた汎化学習理論を土台にした汎化性能とロバスト性の理論的探求や開発したアルゴリズムの正当性の担保が困難**であったことが挙げられる．

この事実に直面したことで，汎化性能とロバスト性の関係に現実的な解釈や議論を提供するためには，現在一般的となっている汎化理論に則るのではなく，**新しい汎化理論の構成が重要**なのではないかという仮説を立てるに至った．また，学習アルゴリズムに対する詳細な解析を提供するためには，**学習アルゴリズム特有の性質や学習の途中経過の詳細に着目した理論解析が重要なのではないか**と思うようになった．そこで，後半の 1 年では，この仮説を主眼に据えて研究を遂行した．結果的には，**「研究1：確率的勾配 Langevin 動力学に対する時間非依存な情報理論的汎化誤差上界」**および**「研究2：Stein 変分勾配降下法の弱収束性の理論保証」**の2つの研究成果を得ることができた．次節にて，これらの詳細を提供する．

（2）詳細

・【研究1】：確率的勾配 Langevin 動力学に対する時間非依存な情報理論的汎化誤差上界

確率的勾配 Langevin 動力学（Stochastic gradient Langevin dynamics; SGLD）に基づく確率的最適化は，大規模モデルを実用的にする上で中心的な役割を担っており，それがもたらす汎化性能の特徴を理論的に理解することは，より良いアルゴリズムを模索する上で重要である．しかし，現状の汎化誤差上界や超過リスク上界は，非現実的な学習率の調整を行わなければ，学習アルゴリズムの進行，および学習率の減少に応じて発散してしまい，正確な汎化性能の理解を妨げている問題があった．

そこで，当該研究では，訓練データ分布と学習したパラメータ分布間の相互情報量に基づく新たな情報理論的汎化誤差上界を，訓練データの変化に対する**出力の安定性の時間発展を追跡する**ことで導出した．この汎化誤差上界は，**汎化性能とよく相関することが知られている訓練データの変化に対する目的関数の勾配の安定性**で表現され，汎化の実態をよく反映したものとなっている（**図1参照**）．また，この上界は，既存上界の課題であった，アルゴリズムの繰り返し数に対して線形なオーダーでの学習率減衰という非現実的な仮定や，その逆数に

公開

$$|\mathrm{gen}(\mu, P_{W_T|S}; L)| \leq \sqrt{\frac{2c_1\sigma_g^2}{n}\left(1 \wedge \frac{\eta T}{4\beta c_{\mathrm{LS}}}\right)(V_\nabla + c_2)} = \mathcal{O}(\sqrt{(\eta T \wedge 1)/n})$$

汎化誤差

訓練データの変化に対する勾配の安定性

図 1：本研究で導出された SGLD の新たな汎化誤差上界.

表 1：既存上界と提案上界との比較. 最右列が各上界のオーダーを示す. n, t, ηはそれぞれ標本数，アルゴリズムの繰り返し数，学習率を表す. 提案上界（最下行）は，学習率を$\mathcal{O}\left(\frac{1}{t}\right)$に設定せずとも，$t \to \infty$で発散せず，$\eta$の減少に伴う発散も生じない.

Study	Assumptions for a loss function	Expected generalization error bound
(S) Raginsky et al. [29] (Thm. 2.1.)	Dissipative, Smoothness	$\mathcal{O}(\eta t + e^{-\eta t/c} + 1/n)$
(S) Mou et al. [23] (Thm. 1.)	Bounded, Lipschitz	$\mathcal{O}(\sqrt{\eta t}/n)$
(S) Mou et al. [23] (Thm. 2.)	Lipschitz, Sub-Gaussian, (Weight decay) [6]	$\mathcal{O}(\sqrt{\eta \log(t+1)/n})$
(I) Pensia et al. [28] (Cor. 1.)	Lipschitz, Sub-Gaussian	$\mathcal{O}(\sqrt{\eta t/n})$
(I) Negrea et al. [25] (Thm. 3.1.)	Sub-Gaussian	$\mathcal{O}(\sqrt{\eta t/n})$
(S) Farghly and Rebeschini [10] (Thm. 3.1.)	Lipschitz, Smoothness, Weight decay	$\mathcal{O}((\eta t \wedge 1)(1/n + \sqrt{\eta}))$
(S) Farghly and Rebeschini [10] (Thm. 4.1.)	Dissipative, Smoothness	$\mathcal{O}((\eta t \wedge 1)(\sqrt{\eta^{-1}}/n + \sqrt{\eta}))$
(I) Wang et al. [35] (Thm. 1.)	Sub-Gaussian	$\mathcal{O}(\sqrt{\eta t/n})$
(I) Ours (Thm. 4 and Cor. 1)	Dissipative, Smoothness, Sub-Gaussian*	$\mathcal{O}(\sqrt{(\eta t \wedge 1)/n})$

Diverges as $t \to \infty$ unless $\eta = \mathcal{O}(1/t)$ or $\eta = \mathcal{O}(1/\log(t+1))$

Diverge as $\eta \to 0$ without weight decay

Do not diverge!

依存するが故の発散を迂回しつつ，SGLD の最適化繰り返し数に伴う発散を回避する（**表1参照**）. この研究では**ロバスト性との関わりについての議論は行われていない**が，**具体的な学習アルゴリズムに注目し，その出力の安定性を通した汎化性能解析を行う**ことが，より現実的な汎化性能理解に繋がり得るという知見が得られた.

このアプローチは，**今後ロバスト学習と汎化性能解析との関係性を理論的に解析する際の手段として有用な候補の一つとなることが期待**される. また，本研究成果は，大阪大学講師の二見太先生が遂行されている「JST さきがけ研究：信頼される AI」の採択課題：「情報理論を用いた不確実性に関する学習理論の展開」との連携により創発された.

・【研究2】（投稿中より一部詳細非公開）：Stein 変分勾配降下法の弱収束性の理論保証

複雑な確率モデルにおけるベイズ推論では，事後分布を解析的に得ることは困難となる. 変文推論は，事後分布を近似的に得るための方法論の一つであり，中でも Stein 変分勾配降下法（Stein variational gradient descent; SVGD）は，一般の変分推論のように正規分布などの取り扱いが容易な分布の制約を課すことなく，柔軟に事後分布を近似可能にする方法である. SVGD は実用上では優れた近似精度を達成する一方で，その収束性については，カーネル Stein 距離と呼ばれる，**カーネル関数に基づく弱い距離尺度の下での保証が主流となっていた.** これにより，SVGD による近似においては，事後分布への弱収束性が必ずしも保証されないという問題を残していた. より現実的な距離尺度，例えば**カルバック・ライブラー（Kullback–Leibler; KL）距離尺度**上での事後分布への収束性は，近似推論の有用性を担保する上では重要な条件である.

SVGD によって得られる近似分布の事後分布への弱収束性を示すためには，**近似分布と目標分布とのカルバック・ライブラー（Kullback–Leibler; KL）距離基準**で，アルゴリズムの繰り

公開

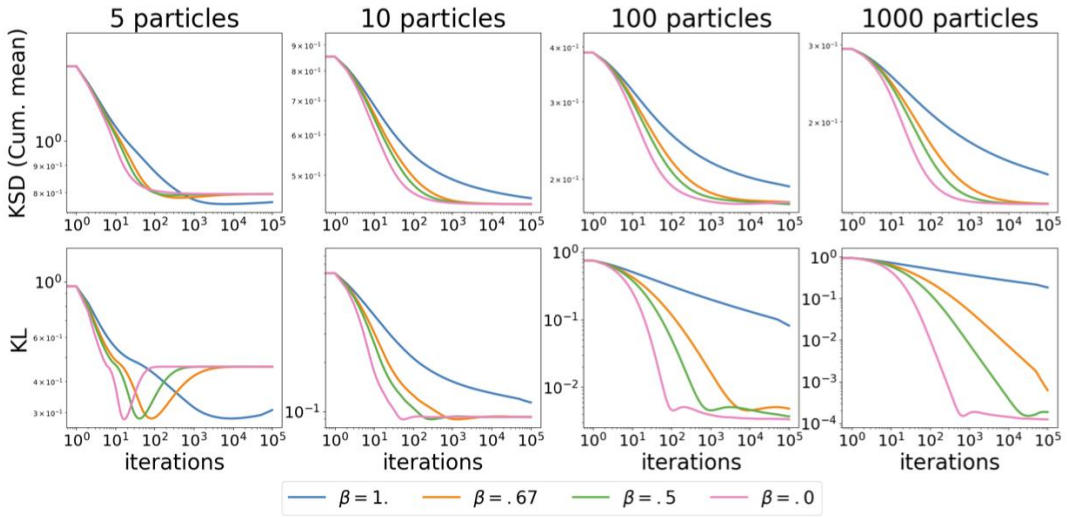

図 2：SVGD による正規分布の近似推論実験．βは学習率の減衰率を司るパラメータ．粒子数（particles）が大きくなればなるほど，適切な学習率減衰のもとでは，アルゴリズムの繰り返し数に対して準線形に KL 距離（第２行）の値が減少する傾向が見て取れる．

返し数に対して収束することを示せば良い．常套手段としては，**対数ソボレフ不等式と呼ばれる，KL 距離とフィッシャー情報量との関係性を利用**した証明方法が挙げられるが，**SVGD では，この関係性の成立が非常に困難**である．よって，この方法に頼らない代替手段を考える必要が生じる．

そこで当該研究では，「(ε,δ)−近似勾配流」という，勾配流の近似精度を測る新しい尺度を導入し，これに基づく収束解析を試みた．この結果，無限粒子および有限ステップ幅の下での SVGD が，**近似分布と目標分布との KL 距離基準で，アルゴリズムの繰り返し数に対して準線形収束**することを示すことに成功した．実際の SVGD は有限粒子による近似推論を用いるため，無限粒子設定は実用上とかけ離れているが，数値実験上は，粒子数の増大に連れて，準線形収束のような挙動を示すことが確認された．

【研究１】と同様に，本研究成果は，大阪大学講師の二見太先生が遂行されている「JST さきがけ研究：信頼される AI」の採択課題：「情報理論を用いた不確実性に関する学習理論の展開」との連携により創発された．

3．　今後の展開

2 年半の ACT−X 研究，特に【研究１】を通じて，現実的な汎化性能に関する理論理解を提供するためのアプローチとして，**各アルゴリズムの特徴に着目し，出力の安定性を綿密に解析する**ことが，有効となり得るという示唆が得られた．今後も本アプローチの適用を通して，幅広い学習アルゴリズムの汎化性能解析を進めることが重要な研究展開の１つとなることが予想される．例えば，初期研究にて失敗に終わっていた経験変分ベイズの汎化解析と提案アルゴリズムの理論保証などに本アプローチを適用し，**汎化性能と事前分布の誤設計に対するロバスト性を両立する新たな学習アルゴリズムを理論保証とともに提供する**ことは，当該 ACT−X 研究課題の目標である，「ロバスト性と汎化性能を両立する機械学習の確立」を達成する上で喫緊の課題である．

公開

また，敵対的学習や外れ値に頑健な学習などといったロバスト学習が汎化に及ぼす影響を上記のアプローチで解析することも，ロバスト性の獲得とそれによる汎化性能への影響に関する詳細を明らかにする上で検討すべき方向性である．さらに，既存の理論体系で両性能のより良いトレードオフを達成できる手法が開発できそうであれば，上記研究と並行して積極的に研究を展開することも検討している．

4. 自己評価

研究目的の達成状況：研究計画の進行に関しては，初期研究が予想以上に困難を極めるなど，前途多難であり，当初の計画通り，ロバスト性と汎化性能との関わりに関する研究成果を揃えることはできなかったと言わざるを得ない．一方で，初期研究での失敗を契機に，より現実的な汎化性能解析へと方針を一旦切り替えたことが功を奏し，特に**【研究1】**は論文成果となり，当該目標の達成に向けて，進度は遅くとも着実に前進することができたと言える．**【研究2】**は現状，汎化解析に直接的には結びついていないが，現実的な距離尺度上での事後分布への収束性保証は，近似推論としては学習アルゴリズムの有用性に直結する条件であるため，それを示すことを優先した．ACT-X 期間中に，この後続研究として，汎化性能解析に繋げることができなかったのが悔やまれる．以上を鑑みると，本研究課題の研究目的のうち，汎化性能解析の側面では今後の研究方針に繋がる論文成果を挙げることができたが，ロバスト性との議論まで展開することができなかったというのが研究達成状況の総括である．

研究の進め方：ACT-X の構想通り，研究代表者主導による研究遂行ができた．また，数年を費やした初期研究が失敗したにも関わらず，さきがけ研究者との連携により，短期間で2つの研究成果を挙げられたことは，効率的かつ有効な研究実施体制を築けたといって差し支えないであろう．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って失われていた行事，とりわけ国際会議のオフライン開催が再び執り行われるようになったことで，招待公演や採択論文の現地発表のために必要となる出張費用を中心に執行を行なった．新型コロナウィルス感染症蔓延時に計画的に予算計画を少額に設定し，最終年度に多くの予算を残していたことが，多くの会議参加を可能にし，これがひいては，国内外の共同研究ネットワークの構築に活きたと考えている．

研究成果の科学技術及び社会・経済への波及効果: 情報の迅速な自動取得・蓄積が日常化された昨今では，実運用している機械学習技術が，次々に得られる膨大な未知のデータに対して良く汎化することが，信頼性の高い機械学習運用における重要な要素の一つとなっている．本研究成果は，汎化性能は，学習アルゴリズムから得られるどの情報と相関しているかを説明づけるものであり，この知見は，深層学習などといった複雑なモデルの汎化性能の理論解析の際の基礎としても役立つことが期待される．汎化性能の実態が理論的な形で明らかとなることで，多くの機械学習技術の信頼性評価が可能となり，人工知能技術の安心安全な実社会応用の促進と，それに伴う経済発展を後押しできる可能性を秘めている．また，この人工知能技術の積極的な応用を促すことで，長期的に見れば，これから起こり得る科学技術イノベーションを誘発する一要素となり得る．頑健性の議論への展開し，この流れをさらに一押しできる成果への昇華を図りたい．

公開

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： 2 件 （うち1件は投稿中により非公開）

1. Futami, F.* and **Fujisawa, M.***† Time-Independent Information-Theoretic Generalization Bounds for SGLD. In Advances in Neural Information Processing Systems (NeurIPS), 2023 (to appear). (*: Equal contribution, †: Corresponding author)
当該論文では，確率的勾配 Langevin 動力学（Stochastic gradient Langevin dynamics; SGLD）に対し，訓練データ分布と学習したパラメータ分布間の相互情報量に基づく新たな情報理論的汎化誤差上界を，訓練データの変化に対する**出力の安定性の時間発展を追跡する**ことで導出する．この汎化誤差上界は，既存上界の課題であったステップサイズに対する厳しい仮定を迂回しつつ，SGLD の最適化繰り返し数に伴う発散を回避する．また，この上界は，**汎化性能と関わりの深い，訓練データの変化に対する目的関数の勾配の感度**で表現され，より汎化の実情を反映したものとなっている．
2. **Fujisawa, M.*** and Futami, F.* Convergence of SVGD in KL divergence via approximate gradient flow. Under review. (*: Equal contribution)
（投稿中より情報非公開）
3.

（2）特許出願

研究期間全出願件数： 0件

（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）

主要な学会発表：

・二見太，藤澤将広（共同筆頭著者）．SGLD のための時間非依存な情報理論的汎化誤差上界．第 26 回情報論的学習理論ワークショップ．北九州国際会議場，10 月 30 日，2023.

・藤澤将広，二見太（共同筆頭著者）．「近似勾配流」が実現する SVGD の KL 距離下での準線形収束保証．第 26 回情報論的学習理論ワークショップ．北九州国際会議場，10 月 31 日，2023.

・Futoshi Futami, Masahiro Fujisawa (Equal Contribution) Time-Independent Information-Theoretic Generalization Bounds for SGLD. The 37th Conference on Neural Information Processing Systems (NeurIPS2023), Dec. 15, 2023.