公開

研 究 終 了 報 告 書

「求解軌道のマクロ表現によるアルゴリズム制御理論の創出」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 坂田 綾香

1. 研究のねらい

計測技術の向上や記憶媒体の大容量化がもたらした高次元データは、その利活用を通して世の中の流れを大きく変えつつある。統計的機械学習は、高次元データに潜む特徴を抽出し、予測や推定に役立てるための技術基盤である。統計的機械学習における多くの問題は、高次元変数の推定問題として定式化される。しかし解析解が構成できるのは一部の推定問題に限られる。解析解が構成できない場合は、何らかのアルゴリズムにより解が求められる。多層ニューラルネットワークなどの複雑なモデルが導入される昨今、推定問題の複雑さも高まっており、効率的なアルゴリズムの開発は急務である。

そこで、複雑な推定問題に対しても適用可能な、アルゴリズムによる求解の理論的限界を評価する方法と、その限界を達成するようにアルゴリズムを制御する方法を開発することが本研究の目標である。これにより、実用化可能な統計的機械学習手法の多様性が高まり、情報からの知の抽出を通した社会発展に貢献する。

その目標に向け、本研究では以下(A)～(C)の目的達成を目指す。

A) 求解軌道のマクロ表現に基づく、軌道の確率的揺らぎの評価

高次元空間上で、アルゴリズムを通して解に至る軌道(求解軌道)をそのまま把握することは、極めて難しい問題である。そこで、統計量などを用いることで、統計的機械学習における特徴抽出の考え方をアルゴリズムに適用し、求解軌道の低次元表現を得ることを考える。この手続きを、これが成功すれば、求解軌道の振る舞いを把握することができると考える。

求解軌道は、入力データに内在するランダム性などの影響を受けて揺らぐ。よって使用するデータがたまたま非典型的なものであった場合、求解軌道が予想よりも大きく揺らぐことがある。この揺らぎを把握することでアルゴリズムの信頼性を評価し、ランダム性が求解に与える影響を明らかにする。特に、軌道を特徴づける統計量と、揺らぎの関係性についての定量化を目指す。

B) 能動的データサンプリング法に対する理論の構築

求解軌道は、入力データについての関数であるとみなすことができる。したがって、データを逐次的に取得する場合など、入力データに対する解軌道の振る舞いを知ることができれば、データ取得にタイミング、適切な回数、などの指針が得られると期待される。

ここでは、能動的なデータサンプリング法に基づき、逐次的にデータを取得するアルゴリズムについて、その性能やアルゴリズム軌道の安定化についての方法を開発する。従来法よりも計算量の少ない方法が得られれば、能動的学習に対する強力な手法となると期待される。

C)　アルゴリズムのパラメータ制御方法の開発

取得されたデータによらずに安定な動作を示すアルゴリズムを開発するには、求解軌道の揺らぎをいかに制御するかが鍵である。アルゴリズムが持つパラメータをデータの性質に応じて適切に制御することで、アルゴリズムの動作を安定化する方法を開発する。

具体策として、アルゴリズムの各ステップで信頼性が高くなる方向へとアルゴリズムのパラメータを変化させる方法が考えられる。

以上の研究を通し、幅広いデータおよび推定問題に適用可能な汎用的アルゴリズムを確立することでコンピューティングの革新に貢献することを目標とする。

2. 研究成果

(1)概要

【グループテストにおける能動的データサンプリング法の開発】

能動的なデータサンプリングを行うことでアルゴリズムの性能を安定化させる方法について研究した。具体的問題として、グループテストと呼ばれる問題を扱った。我々は、観測にノイズが含まれるグループテストの問題において、ノイズを除去しながら真の状態を復元するアルゴリズムの研究を行った。さらに、能動的なデータサンプリングを、確率伝搬法とベイズ予測誤差の評価により実現する方法を提案した。

【Query by Committee 法によるベイズ予測誤差評価に対するアルゴリズム開発】

上記のグループテストの問題は、扱う変数が離散値であることが問題の性質を簡単にしている。そこで、連続変数に対して近似的にベイズ予測誤差を評価する query by committee 法を導入し、その理論解析とアルゴリズム開発をおこなった。query by committee 法には、これまでマルコフ連鎖モンテカルロ法などによる数値計算法が提案されていた。しかし計算量の観点から実用的ではないため、近似アルゴリズムを query by committee 法に適用することで計算時間を大幅に短縮できること、またその理屈について説明した。

【非凸スパース正則化に対する非凸性制御法の提案】

一般に、昨今のスパース推定などで用いられている正則化関数は、凸関数が多い。一方で、非凸関数を用いた方が高性能となることが示唆されている。しかしながら、非凸関数の扱いは容易ではなく、一般に局所解がアルゴリズムの収束を妨げることが知られる。そこで、非凸正則化最小化問題に対してアルゴリズムを構成し、安定に収束するための「非凸制御法」を提案した。

【近似確率伝搬法を用いた交差検証誤差の評価方法】

一般化線形モデルと罰則付き最尤法における予測誤差について、一般化近似確率伝搬法を用いて統計モデルの予測精度を評価する方法を提案した。予測誤差は、ベイズ予測分布により定義され、これは能動的データサンプリングにおける、データの不確実性にも対応する。解析の結果、予測誤差は軌道の揺らぎにより記述できること、その揺らぎが確率伝搬法による軌道を特徴づけるマクロ変数により与えられることがわかった。したがって本研究は、マクロ表現の元、軌道の揺らぎを記述できる理論の基盤となると考えている。

(2)詳細

「1. 研究目標」に記載した内容と、研究成果の関連は以下の通りである。

(A) 求解軌道のマクロ表現に基づく、軌道の確率的揺らぎの評価

軌道の確率的な揺らぎを評価する際の発想は、「軌道はデータの確率変数列と見做せる」という点に基づく。この考えを導入すると、軌道の固定点に対して展開されてきた統計学における揺らぎの評価方法が、軌道に対しても適用できるとの着想に至った。軌道の固定点は、一般的には何らかの推定問題の解に一致する。この解が、入力データによりどのように変化するのかと言う点については、古くから統計学で議論されてきた。統計学的な考え方では、揺らぎの少ない推定量が良い推定量であり、また未知データに対しても頑健性を持つモデルが良いモデルであると考える。このような考え方のもと、未知データに対する揺らぎの評価が長年行われてきている。情報量規準や、交差検証誤差も、このような文脈で理解することができる。

情報量規準や交差検証誤差は、予測誤差の推定量である。予測誤差とは、未知データに対するモデルの当てはまりを評価する指標である。まず、我々は一般化近似メッセージパッシング(GAMP)アルゴリズムを用いて、予測誤差と推定値の揺らぎの関係性を明らかにした。具体的には、一般化線形モデルの予測誤差が、推定値の揺らぎから評価できるという、物理学における「揺らぎと応答」に対応する関係性を発見した。すなわち、データに対してどの程度推定結果が変化するのかを知りたければ、現在得られている推定結果の分散を見れば良い、ということを意味する。この事実は、事後分布がラプラス近似できる系であれば常に成り立つと期待されるため、普遍的な関係性の抽出に成功したといえる。

この関係性を軌道に応用すれば、軌道上の各点で、外的要因による軌道の揺らぎやすさを評価できることになる。軌道の理論への応用については現在研究中であるが、固定点に関する理論の結果については、以下の投稿中論文にまとめた。

A. Sakata

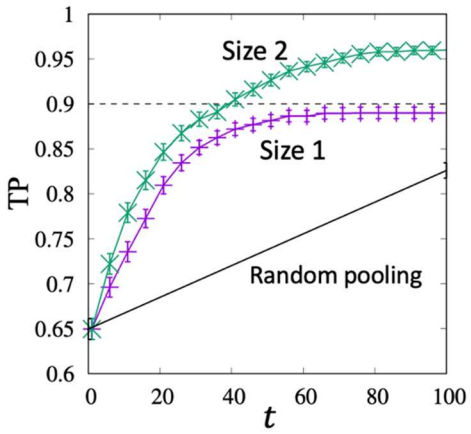
"Prediction Errors for Penalized Regressions based on Generalized Approximate Message Passing"

arXiv:2206.12832 (2022)

(B) 能動的データサンプリング法に対する理論の構築

2019~2020 年度は、能動的なデータサンプリングを行うことでアルゴリズムの性能を安定化させる方法について研究した。具体的問題として、グループテストと呼ばれる問題を扱った。この問題は 0,1 で指定された各要素の状態を、少ない数の観測数から復元することを目標とする。一般には、1 をとる要素数が十分小さいと仮定する。集団の中に含まれる、状態 1 の要素が異質なものであると考え、この 1 をとる要素を検出することが目標である。例えば、0 を非感染者、1 を感染者とすれば集団の中から感染者を見つける問題となり、また 0 を正しい製品、1 を不良品とすれば、製品集団の中から不良品を見つける問題と読み替えることができる。

本研究では、観測にノイズが含まれる設定を考え、そのノイズを除去しながら真の状態を復元するアルゴリズムの研究を行った。さらに、データ取得法を工夫することでアルゴリズムの性能を上げる方法を検討した。データ取得法の基準は、ベイズ予測分布である。ベイズ予測分布とは、既存データから構成したモデルが実際のデータ生成過程とどの程度合っているかを表す指標である。既存データに可能な新しいデータを追加した際に、どの程度推定精度が改善されるかを評価し、推定精度を大きく改善させるデータを逐次的に取り込んでいくという方法により計算効率を上げる、というのがここで用いたアイデアである。



実際に、グループテストと呼ばれる問題において、ベイズ予測分布に基づくデータ取得法を実行した。グループテストとは離散変数の推定問題であり、ベイズ予測分布が一つの統計量で表されるという利点がある。一般の問題ではベイズ予測分布の評価において大きな計算量が必要となるが、グループテストでは一つの統計量を評価すればよいため、データ取得にかかる計算量が少なく済む。

グループテストにおいてベイズ予測分布に基づくデータ取得を行なった結果、ランダムなデータ取得方法と比べて、学習に必要なデータ数が約 7 割で済むことがわかった。右図は、ランダムなデータ取得(random pooling)と、ベイズ予測分布に基づく二種類のデータ取得法を比較したものである。縦軸は推定精度、横軸はデータ数を表している、能動的データンプリング法の方が、少ないデータ数で高い推定性能を示していることがわかる。データ数が増えるほど計算時間が増えるため、このデータ数の削減には大きな意味がある。

一方で、先行研究では、データ間に相関がある場合に一部のアルゴリズムについて収束にかかる時間が大きく延びるということが示されていた。何らかの基準に基づくデータ取得法は、データ間に相関を生じさせることが知られる。よって私自身も、データ取得法を効率化すれば、そのトレードオフとしてアルゴリズムの収束時間が大きくなってしまうことを懸念していた。しかしながら本研究ではそのような傾向は見られなかった。なぜデータ間に相関があるのにアルゴリズムの挙動がほとんど変わらないのか、それはグループテスト以外の問題についても一般的であるのかといった、さきがけ研究の発展としても理解すべき点が新たに示された。

上記のグループテストの問題は、扱う変数が離散値であることが問題の性質を簡単にしている。具体的には、能動的データ取得の根拠となるベイズ予測分布が求められるという利点がある。一方で、連続変数に対してベイズ予測誤差を評価するには、指数関数オーダーの計算量が必要になるという問題点がある。この問題点を解決するため、query by committee とよぶ学習法を導入した。この方法では、取得済みのデータから構成した事後分布を用いて、committee member と呼ぶ確率変数を生成する。その committee member と整合性が取れないデータを積極的に取得する。「整合性がとれない」というのは、現状の事後分布ではうまく

記述できていないという意味であり、そのデータを適切に記述することができるよう、事後分布を修正する必要がある。そのデータの「整合性のなさ」は committee member の経験分布から特徴付けられるが、この経験分布は committee member 数無限大の極限でベイズ予測分布に一致する。

この query by committee 法に対しては、これまでマルコフ連鎖モンテカルロ法などによる数値計算法が提案されていた。しかし計算量の観点から実用的ではないため、まずは現代的なベイズ近似推論アルゴリズムを構築するところから研究を開始した。本研究では、確率伝搬法とよぶ近似アルゴリズムを query by committee 法に適用し、計算時間を大幅に短縮した。

以上の研究は以下の論文として出版された。

A. Sakata
"Bayesian inference of infected patients in group testing with prevalence estimation"
J. Phys. Soc. Jpn. 89, 084001 (2020). Editors' choice

A. Sakata
"Active pooling design in group testing based on Bayesian posterior prediction"
Phys. Rev. E 103, 022110 (2021).

また、以下の投稿中論文も本件に関連するものである。

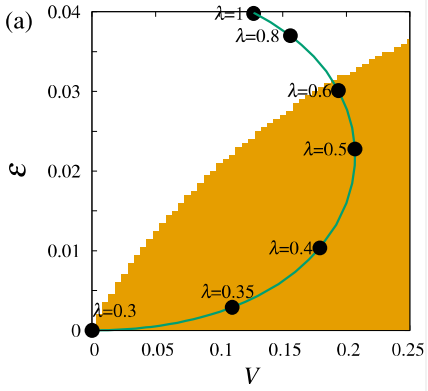
A. Sakata and Y. Kabashima
"Decision Theoretic Cutoff and ROC Analysis for Bayesian Optimal Group Testing"
arXiv:2110.10877 (2021).

(C) アルゴリズムのパラメータ制御方法の開発

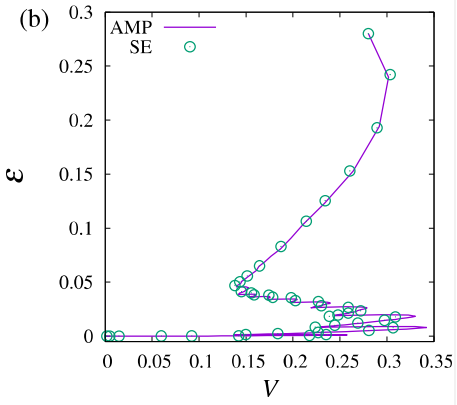
アルゴリズムに含まれるパラメータを制御する例として、非凸制約最小化問題を研究した。凸制約最小化問題ではアルゴリズムの挙動が非常に安定であり、特に L1 制約最小化問題などでは固定点への引き込み領域が発散するという利点がある。すなわち、初期条件を fine tuning しなくても、L1 制約最小化問題では(固定点が存在するときは)固定点に収束する。一方で、このような性質は非凸制約最小化問題においては存在せず、初期条件の選び方により性能が大きく変わるという問題点があった。初期条件を fine tuning せずに、アルゴリズム軌道を自動的に引き込み領域に導くような技術が開発できないかと考え、圧縮センシングにおける確率伝搬法に対するアルゴリズム制御法を検討した。

確率伝搬法とはベイズ推定に対する近似アルゴリズムであるが、圧縮センシングのようにデータ観測手法がランダム行列で与えられている場合は、アルゴリズム軌道が二つのマクロ変数(ここでは V と E とする)の軌道により近似可能ということが知られている。また、推定の成功は V=E=0 への到達として理解できる。したがって、アルゴリズムの制御の結果としてどのような軌道を通って V=E=0 に到達するのか、２次元平面で理解できるという利点がある。

ここで考えるアルゴリズム制御法とは以下の通りである。アルゴリズムの時間発展は、ある非凸関数によって与えられている。その非凸関数の「非凸度合い」をパラメータλで特徴づける。このλが小さいほど性能が良いが、V=E=0 への引き込み領域が小さくなるという問題点をまず発見した。右図(a)の色付きの領域は、λ=0.3 での V=E=0 への引き込み領域である。V=E=0 に到達するためには、E〜0.01 としなければならないが、E は物理的には真の解と推定値の二乗誤差を意味しており、すなわち真の解と十分近い初期条件を選ばなければ推定成功しないことを意味する。真の解を発見すべく推定を行いたいにもかかわらず、そのためには真の解に十分近い点を初期条件としなければならない、ということでは実用上ほとんど意味がない。右図(a)の実線は、引き込み領域の縮小に伴う困難をアルゴリズム制御法により解決する方法の概略である。研究の結果、パラメータλが大きいとき、V=E=0 に収束することはできないが、しかし大きな引き込み領域を持つため、アルゴリズムは安定的に収束することがわかった。このアルゴリズムの安定性を利用し、λを段階的に下げてくることで、λ=0.3 への引き込み領域に自動的に入ることができる、というのがここで考えたアルゴリズム制御法である。



右図(b)は本研究で実装したアルゴリズム制御法の結果である。実線が実際のアルゴリズムの軌道で、丸は理論予想である。λを大きい値から段階的に下げることで、V=E=0 への引き込み領域に到達し、V=E=0 に収束することができる。



この方法は、更なる改善が見込まれること、また他の問題に対しても適用可能であると考えられることから、さきがけ研究終了後も引き続き研究を行なっていく。

以上の研究は以下の論文に出版されている。

Ayaka Sakata and Tomoyuki Obuchi

"Perfect reconstruction of sparse signals with piecewise continuous nonconvex penalties and nonconvexity control"

Journal of Statistical Mechanics: Theory and Experiment, vol. 2021(9), 093401 (2021).

3. 今後の展開

本研究の成果が将来的な社会実装に繋がるためには、次のような展開が必要であると考える。

まず、非凸性制御などのアルゴリズムのパラメータ制御法についての新しい知見をもとに、複

雑な推定問題に対する推定法のパッケージを開発することができれば、将来的にはデータサイエンス、製薬、医療統計といった、統計的機械学習と関連する分野のうち社会生活に近い分野において、機械学習の方法論が有効活用されると期待される。さらに、坂田は機械学習のみならず、進化生物学における学習理論の展開についても研究しており、アルゴリズムの安定性に対する理論は、「生物がなぜ安定な表現型発現ダイナミクスを示すのか」といった問いに対する基礎理論として、進化生物学の分野でも新しい展開が期待されると考えている。

研究発展のタイムスパンとしては、パッケージ化については1～2年程度で実現すると考えている。本研究で扱った非凸制約最小化問題のみならず、他の非凸問題等に対する適用を調査することが追加で必要である。有効性の確かめられた問題群について、共通部分と個別部分をそれぞれ洗い出し、それらを踏まえたパッケージ化を実現することが望ましいと考えている。また、進化生物学との関連性については、すでに研究を進めており、いくつかの新規な結果が得られている。2023 年度中の論文投稿が可能であると考えている。現状では、簡単なタンパク質の模型におけるダイナミクスの安定性について議論しているが、安定性を示すダイナミクスを持つ生物システムは多数にわたる。これらのシステムが持つ普遍的性質について、ダイナミクスをアルゴリズム軌道と捉えることで、本研究を応用させた学祭研究を展開していきたい。この展開は、おそらく数十年レベルでの研究が必要になると考えているが、機械学習と進化生物学の協働という新分野の創生につながると考えている。

4. 自己評価

本研究は、統計的機械学習の問題を対象として、アルゴリズムの挙動の現状を明らかにし、その問題点を解決する方法を提案することを目的としてきた。研究開始当初から、発見的アルゴリズムの性能、根拠については知られていない点が多く、本研究の結果を通して、新しい知見をもたらすことを目指してきた。

実際の研究では、ベイズ推定の文脈で提案されたグラフィカルモデルによる表現とその上で定義される確率伝搬法と呼ばれるアルゴリズム、また統計物理学の方法を用いて、アルゴリズム軌道に関する性質についてのデータ収集と、解析的な表現の導出を試みた。その結果、能動的データサンプリング法や非凸問題といった複雑な推定問題に対するアルゴリズムの現状を明らかにすることができ、性能改善方法を提案することができた。また、アルゴリズム軌道の揺らぎについての新しい知見を得ることができ、国内外での発表や論文執筆を通じて、研究成果を発信することができた。特に Journal of Physics 誌に、本件に関する Review Paper の依頼をいただけたことは、とても大きい成果であると考えている。

研究を進める上で、グラフィカルモデルのうち、有効グラフについての知識が浅かったため、ベイジアンネットワークにおけるアルゴリズムについては、本さきがけ研究内で新規な結果をもたらすことができなかった。しかし、さきがけ研究をきっかけに、東工大、東大などで集中講義の機会をいただき、その中で再度、現代的な知見を勉強し直すことができた。これをもとにさきがけ研究を発展させていきたい。また、研究費執行状況についても、難しい点があった。技術支援員を 3 名雇用する予定であったが、2 名の雇用にとどまり、そのうち1名も体調の問題で想定の研究活動を遂行できず、研究費の執行が予定通りに進まなかった。情報系の人材不足、またコロナ禍という難しさが背景に

あったと考えている。前者については、研究のみならず、研究をきっかけとした教育活動にも貢献することで、状況を改善していきたいと考えている。

幸い本研究の遂行期間中に、イタリアの G. Parisi 博士が本研究でも用いた解析手法であるレプリカ法に関連する研究でノーベル賞を受賞し、我々の研究分野への注目度が増した。日本物理学会などでは、関連シンポジウムが開かれ、私も招待いただいた。このような機会にさきがけ研究を遂行していたおかげで、研究成果の社会への波及効果を高めることができたことに感謝している。今後も、私が関わる統計科学、統計物理、情報科学の分野において、新しい知見を発見しつつ、その社会的発信にも努めていきたい。

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:8件

1. A. Sakata, "Bayesian inference of infected patients in group testing with prevalence estimation" Journal of Physical Society of Japan 89, 084001 (2020) **Editors' choice**
グループテストにおいて、検査が有限の確率で偽の結果を返す場合について，ベイズ推論とそれに対応する確率伝搬法を導入し，プールに対して行われた検査結果から感染者を同定する問題を扱った．EM 法によるパラメータ推定と BP アルゴリズムを組み合わせることで，検査における有病率と誤差確率を推定することができること、また階層ベイズモデルを導入し，感染者の特定と有病率の推定を同時に行うことができることを示した。
2. A. Sakata, "Active pooling design in group testing based on Bayesian posterior prediction", Phys. Rev. E 103, 022110 (2021).
グループテストにおける能動的検査法についての研究成果である。本論文では，ベイズ推論の枠組みの中で，予測分布に基づく適応的なプールの設計法を提案した．提案手法を確率伝搬法により実行した結果，事前に決定したランダムなプールを用いた集団検査と比較して，より正確に感染者を同定することができた．
3. Ayaka Sakata and Tomoyuki Obuchi, "Perfect reconstruction of sparse signals with piecewise continuous nonconvex penalties and nonconvexity control" Journal of Statistical Mechanics: Theory and Experiment, vol. 2021(9), 093401 (2021).
次元より少ない測定値からのスパース信号の再構成問題を、非凸スパース正則化最小化も大として定式化した。正則化関数の非凸性は非凸性パラメータによって制御され、L1 ペナルティはこれらのパラメータに関する極限値として含まれる。解析的に得られた再構成限界は L1 限界を克服し、非凸パラメータが適切な値であればベイズ最適設定のアルゴリズム限界も克服すると期待される。しかし、比較的高密度な信号の再構成が理論的に期待される非凸性パラメータが小さい場合、解析と密接な関係にある AMP と呼ばれるアルゴリズムでは完全な再構成ができず、解析解とのギャップが生じた。このギャップは、完全再構成解の引き込み領域

が縮小していることと、またある領域で AMP が発散する挙動により生じることが、state evolution と呼ぶ方法により明らかになった。また、非凸関数の形状を制御し、AMP の軌道を引き込み領域に誘導することで、理論とアルゴリズムのギャップの一部を解決した。

(2) 特許出願

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)

[主な学会発表(招待講演)]

- Ayaka Sakata, Active pooling design in group testing based on Bayesian posterior prediction ISM-Bristol Joint Seminar 2020 年 9 月 10 日
- 坂田綾香,「非凸スパース正則化による信号復元とアルゴリズム軌道の制御」統計学会春季大会 2021 年 3 月 13 日
- 坂田綾香,「グループテストにおける確率伝搬法と最適カットオフ」離散数学とその応用研究集会 2022 年 8 月 19 日
- 坂田綾香,「学習と進化におけるレプリカ対称性の破れ」日本物理学会 2022 秋季大会 シンポジウム「Parisi のスピングラス理論と複雑系研究の発展」2022 年 9 月 10 日
- 坂田綾香,「統計的機械学習における近似計算アルゴリズムとその理論」日本学術会議 第12回計算科学シンポジウム 2022 年 12 月 5 日

[そのほか]

- Journal of Physical Society Japan, Editors' choice (2020 年 8 月)