公開

研 究 終 了 報 告 書

「最適化アルゴリズムの平均感度解析」

研究期間： 2019 年 10 月～2023 年 3 月
研 究 者： 吉田 悠一

1. 研究のねらい

機械学習、オペレーションズ・リサーチ、データマイニングなどの応用分野における多くの問題は、最適化問題として定式化することができる。それらを解くアルゴリズムは、意思決定や知識発見の道具として、社会や科学研究において幅広く利用されている。

最適化問題の殆どは NP 困難であり、コンピュータで効率的に最適解を計算するアルゴリズムを持たないと信じられているため、最適化問題の理論研究の多くは近似アルゴリズムに関するものである。しかし最適化の応用が広がるにつれ、近似性能だけでアルゴリズムの性能を測ることが必ずしも適当でない状況が増えてきている。具体的には

- 観測や実験の精度の限界により真のデータの一部しか手に入らない
- 入力データが時間とともに変化する
- 入力データが他のデータを変換することで作られており、その変換に恣意性がある

など、入力データが現実を正確に反映しているとは必ずしも言えない状況である。アルゴリズムを利用した意思決定や知識発見を適切に行うためには、このような状況下でのアルゴリズムの出力の品質を保証する必要がある。

この状況を踏まえ、研究代表者らは 2019 年に「アルゴリズムの平均感度」という概念を提唱した。直感的には、アルゴリズムAの平均感度が小さいとは、入力の一部がランダムに欠落する時に、元の入力に対してAが出力する解と、欠落後の入力に対してAが出力する解があまり変化しないことを言う。変化の度合いの測り方、つまり解の間の距離をどう定義するかは問題による。

平均感度が小さいアルゴリズムが出力した解は、入力データが不正確であっても、真のデータが与えられた時と同様である。よって得られた解は、その後の意思決定や知識発見に安心して使うことができると言える。しかし残念ながら、現在使用されているアルゴリズムの多くの平均感度は小さくない。

平均感度は研究代表者により提唱されたばかりの概念である。本研究の大きな目的は「これまで研究されてきたあらゆる最適化問題を、平均安定なアルゴリズムが存在するかという観点から問い直すこと」である。本研究はこれまでのアルゴリズム研究に「平均感度」という新しい軸を増やす試みであり、アルゴリズム研究の領域が大幅に広がると期待している。

2. 研究成果

(1)概要

平均感度をグラフの問題に対して定義を行い、最小全域木問題、最小カット問題、最大マッチング問題などに対して平均感度の低いアルゴリズムを構築した。次に多くの問題が動的計画法により解けることに着目し、動的計画法の計算過程を表現できる最大重み鎖問題に対して平均感度の低いアルゴリズムを構築した。最大重み鎖問題に対して帰着することで、最長増加部分列問題、最長共通部分文字列問題、ナップサック問題、RNA 畳み込み問題に対

する平均感度の低いアルゴリズムを構築した。次にデータマイニングで用いられる様々な問題に対して平均感度の観点から考察を行なった。まず、スペクトラルクラスタリングと呼ばれるグラフのクラスタリング手法が、クラスタ構造がある場合には安定であることを示した。次にk-means と呼ばれるベクトルデータに対するクラスタリングの最適化問題に対して、k-means++と呼ばれる既存手法の平均感度が低いこと、任意の近似アルゴリズムに対して、近似度を殆ど損なわずに平均感度を下げられることを示した。また決定木学習や階層クラスタリングに対して精度と平均感度のトレードオフを利用者が自由に選べるような手法を開発した。平均感度はデータの一部を削除する離散的な変更を考えているが、重みつきの問題に対して、重みの変化が連続的に変化する場合の解の変化を定量化したリプシッツ定数を定義した。リプシッツ連続、すなわちリプシッツ定数が有限なアルゴリズムを最小全域木問題、最短路問題、最大マッチング問題に対して構築した。これらの研究成果を通じて、平均感度という概念は多様な問題を扱える汎用性の高い概念であることが確認できた。

(2)詳細

<u>平均感度の提案とグラフアルゴリズムの平均感度</u>

まず平均感度をグラフの問題に対して以下のように定義する。グラフ$G = (V, E)$に対する決定性アルゴリズムAの出力を$A(G)$と書いたとき(典型的には$A(G)$は頂点集合や枝集合である)、アルゴリズムAのグラフGに対する平均感度とは

$$\frac{1}{|E|}\sum_{e\in E} |A(G)\,\Delta\, A(G-e)|$$

である。ここで$S\,\Delta\,$Tは集合S, T間の対称差を表す。

次に、様々なグラフの問題に対して平均感度の解析を行った。以下nを入力グラフの頂点数とする。

- 最小全域木問題:最小全域木問題に対する有名なアルゴリズムであるプリム法とクラスカル法を解析し、プリム法の平均感度は$\Omega(n)$になりうるのに対し、クラスカル法の平均感度は$O(1)$で常に抑えられることを示した。計算量などの観点からはプリム法とクラスカル法には大きな差は無かったが、平均感度の観点では大きく違うことが明らかになった。
- 最小カット問題:最小カット問題に対して$(2+\epsilon)$近似を行う平均感度$n^{O\left(\frac{1}{\epsilon \mathrm{OPT}}\right)}$のアルゴリズムを示した。ここでOPTは最適値である。また任意の近似アルゴリズムの平均感度は近似比によらず$\Omega(n^{\frac{1}{\mathrm{OPT}}})$であることを示し、上記のアルゴリズムの平均感度がほぼタイトであることを示した。
- 最大マッチング問題:乱択の貪欲アルゴリズムが近似度$1/2$、平均感度$O(1)$を達成することを示した。また最大マッチングに対する$(1-\varepsilon)$近似アルゴリズムで平均感度が(nに依存しない)ϵの関数のものがあることを示した。提案したアルゴリズムは既存のストリーミングアルゴリズムに基づいたアルゴリズムである。また、ラムゼー理論を用いることにより定数近似の決定性アルゴリズムの平均感度は$\Omega(\log^* n)$であることを示した($\log^* n$はiterated logarithm)。これにより乱択アルゴリズムと決定性アルゴリズムは、平均感度に関して真に能力が異なることが明らかになった。

これらの成果は離散アルゴリズムのトップ会議である The 32nd ACM-SIAM Symposium on Discrete Algorithms (SODA) (代表的な論文 1、Boston University の Nithin Varma 氏との共著)と The 12th Innovations in Theoretical Computer Science (ITCS)（CMU の Samson Zhou 氏との共著）に採択された。

動的計画法の平均感度

動的計画法を用いて解くことができる様々な問題に対して、平均感度の低いアルゴリズムの設計を行った。最初に、多くの動的計画法の計算プロセスを表現できる最大重み鎖問題を考える。これは有向無閉路グラフ$G = (V, E)$と頂点に対する重み$w: V \to \mathbb{R}_+$が与えられ、G上のパス$P = (v_1, \dots, v_k)$で重み$\sum_{v \in P} w(v)$ が最大になるパスPを求める、という問題である。この問題に対して、近似度が$1 - \epsilon$で平均感度が$O(\epsilon^{-1} \log^3 n)$のアルゴリズムを構築した。
次に最長増加部分列問題、最長共通部分文字列問題、ナップサック問題が最大重み鎖問題に自然に帰着できることに着目し、これらの問題に対する近似度が$1 - \epsilon$で平均感度が$O(\epsilon^{-1} \log^3 n)$のアルゴリズムを構築した。後にナップサック問題に関しては平均感度を$O(\epsilon^{-1})$まで落とせることを示した。
最後に RNA 畳み込み問題と呼ばれるバイオインフォマティクスで盛んに研究されている問題を扱った。この問題の最大重み鎖問題への帰着は非自明であり、$n^{\mathrm{polylog}(n)}$ 個の頂点を持つ有向無閉路グラフを構築する必要があった。結果として近似度が$1 - \epsilon$で平均感度が$O(\epsilon^{-1} \log^7 n)$のアルゴリズムを構築することに成功した。
これらの研究成果をまとめたものは離散アルゴリズムのトップ会議である The 33rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (代表的な論文 2, RA の隈部壮氏との共著)と The 30th Annual European Symposium on Algorithms (ESA)（RA の隈部壮氏との共著）に採択された。

データマイニング手法の平均感度

代表的なデータマイニング手法に、与えられたデータを適切に分割するクラスタリングがある。データを「理解」するためにクラスタリングを用いる場合は、その結果が頻繁に変わると「理解」も変わってしまうことになり都合が悪い。そこで代表的なクラスタリング手法である、スペクトラルクラスタリングとk-means と呼ばれる問題に対して、平均感度の解析を行なった。

- スペクトラルクラスタリング：この手法は与えられたグラフからラプラシアンと呼ばれる行列を生成し、その固有ベクトルにおける各頂点の値を用いることでクラスタリングを行う。ラプラシアンの固有値は全て非負の実数であり、下からi番目の固有値をν_iと書くことにする。このときスペクトラルクラスタリングの平均感度はν_2 / ν_3^2に比例することを示した。これは、上手くクラスタリングできるグラフではスペクトラルクラスタリングが安定していることを示唆している。逆に上手くクラスタリングできないようなグラフでは、そもそもクラスタリングを行うべきではない。つまりスペクトラルクラスタリングを使うべき状況では安心して使ってよいということが明らかになった。またこの理論的なバウンドが実験的にも成り立つことを確認した。本研究成果はデータマイニングのトップ会議である The 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)（中国科学技

術大学の Pang Peng 氏との共著）に採択された。

- k-means: この手法では、与えられたd次元ベクトルデータを用いて、k個の中心点$v_1, \dots, v_k \in \mathbb{R}^d$を選び（与えられたデータ中にない点でも良い）、各点を一番近い中心点に割り当てることでk個のクラスタを得る。まずk-means++と呼ばれる$O(\log n)$近似アルゴリズムの解析を行い、その平均感度が$O(k)$であることを示した。次に、任意の近似アルゴリズムに対して、近似度を少し悪化させるだけで平均感度を$O(\mathrm{poly}(d, k, \log n))$に下げられることを示した。直感的にはまずデータをその性質を保ったまま圧縮し、圧縮したデータに対して与えられたアルゴリズムを適用することで、平均感度を下げる。また平均感度の小さいアルゴリズムを利用することにより、将来の更新が分からないオンライン環境下において、解の更新時間や解の変更度合いをうまく抑えたアルゴリズムを作れることを示した。本研究成果は機械学習のトップ会議である The 36th Conference on Neural Information Processing Systems (NeurIPS)に採択された（代表的な論文 3、NEC の伊藤伸志氏との共著）。

またその他にも代表的なデータマイニング手法である決定木学習や階層クラスタリングについての平均感度の解析も行い、精度と平均感度の間のトレードオフを利用者が調節できる手法を構築することに成功した。前者の結果は機械学習のトップ会議である The 11th International Conference on Learning Representations (ICLR)に採択された（大阪大学の原聡氏との共著）であり、後者の結果は現在投稿中である（大阪大学の原聡氏、京都大学の竹内孝氏との共著）。

<u>その他の成果</u>

- 分解可能劣モジュラ関数の疎化：（単純な）劣モジュラ関数の和で表現される劣モジュラ関数の圧縮について研究を行った。劣モジュラ関数は性質の良い離散関数であり、様々な理論的・実用的な応用がある。特に単調な劣モジュラ関数の最大化は、サイズ制約やマトロイド制約などの様々な制約下で定数近似ができることが知られており、データの要約や影響最大化など様々な用途に用いられている。このような応用で現れる多くの劣モジュラ関数は、比較的単純な形をした劣モジュラ関数の和で表現できることが知られているので、和から一部の劣モジュラ関数がランダムに欠損するときの最大化アルゴリズムの平均感度を低く抑えることはできるかについて研究を行った。結果として平均感度の低い最大化アルゴリズムを構築することはできなかったが、副産物として劣モジュラ関数の和で表現される劣モジュラ関数は非常に小さく圧縮できることが分かった。具体的には、劣モジュラ関数$f: 2^E \to \mathbb{R}$が$f = \sum_{i=1}^{n} f_i$ と表現され、各f_iが劣モジュラ関数のとき、$O\left(\frac{|E|^2}{\epsilon^2}\right)$個の非ゼロ成分を持つ$w \in \mathbb{R}^N$が存在して、$f = (1 \pm \epsilon)\sum_{i=1}^{n} w_i f_i$を満たす。つまりどんなに沢山の関数の和で表現されていても、少ない個数の関数でよく近似できることを意味しており、以降は省メモリで高速に処理を行うことができる。これらの研究成果をまとめたものは人工知能のトップ会議である The 36th AAAI Conference on Artificial Intelligence (AAAI)（Simon Fraser University の Akbar Rafiey 氏との共著）に採択され

た。

- リプシッツアルゴリズム: 平均感度は入力の一部がランダムに欠損した場合の出力の変化を測るものであったが、入力が重み付けされており、その重みが微小に変化することを考える。この時に出力の変化が重みの変化に比例した量で抑えられる、つまりアルゴリズムをリプシッツ連続にできるかについて研究を行なった。平均感度と違い変化が起こった部位に関して平均を取ることができないため、リプシッツ連続性を示す方が平均感度を抑えることよりも難しい。結果として、最小全域木問題、最短路問題、最大マッチング問題などに対してアルゴリズムをリプシッツ連続にできることが分かった。本研究成果は現在投稿中である(RA の隈部壮氏との共著)。

3. 今後の展開

平均感度の研究を数年間続けて、平均感度が一過性の話題ではなく、非常に多くの問題に適用できる汎用性の高い概念であることを確認することができた。これまで理論と応用の両面から平均感度に関する研究を行なってきたが、解くことができた問題よりも、それによって生じた未解決問題の方が多い。以下に現在課題として残っているもののうち特に重要なものを説明する。

- 個別の問題の平均感度の改善:例えば現在最大マッチングに対する$(1-\epsilon)$近似アルゴリズムの平均感度で最良のものは$exp\left(exp\left(exp(\epsilon^{-1})\right)\right)$である。これは明らかにタイトではないので改善したい。また最大カット問題もよく研究されている基本的な問題であるが、非自明な近似を得るためには半正定値計画問題を解く必要があり、その解をベクトルとして取り出す際の感度を抑える方法が分かっていない。
- 他の計算モデルとの関連:平均感度を抑えたアルゴリズムを利用してオンラインアルゴリズムや動的アルゴリズムを構築できることが分かっている(代表的な論文 3)。しかし両者は本質的に異なるのか、または逆向きの変換も可能なのかについては分かっておらず、今後も研究が必要である。
- リプシッツ性:アルゴリズムのリプシッツ性は本研究課題の終わり頃に提案した概念であり、どのような問題であればリプシッツ連続なアルゴリズムを得られるのか全体像がまだ分かっていないため、今後も研究を進めていく。
- 安全性・効率性・再現性の担保:アルゴリズムの感度を抑えることで、現在のアルゴリズムを利用したシステムの問題点である安全性・効率性・再現性が担保できる可能性がある。例えば、微小な変化によって学習したモデルの出力が大きく変わる「敵対的攻撃」はモデルの感度を下げることで防ぐことができる。次に、深層学習において、一部を変更すると全体の挙動が大きく変わるため品質保証が難しいという問題(changing anything changes everything, CACE)が知られているが、これも各部分の感度を抑えることで緩和できる可能性がある。また、不正なデータを混入させることで学習アルゴリズムの出力結果を操作する「データ汚染」と呼ばれる攻撃が知られているが、これも学習アルゴリズムの感度を抑えることで緩和できる可能性がある。

4. 自己評価

- 研究目的の達成状況：研究目的として(1)個別の最適化問題に対する平均安定なアルゴリズムの構築、(2)平均感度の解析に対する方法論の確立、(3)応用分野における平均安定アルゴリズムの有用性の確認、(4)平均感度の亜種の研究の4つを挙げていた。(1)に関しては、最小全域木問題、最小カット問題、最大マッチング問題、最大重み鎖問題、最長増加部分列問題、最長共通部分文字列問題、ナップサック問題、RNA 畳み込み問題などさまざまな問題に対する解析を行うことができた。(2)に関しては、(1)の結果を得る上で統計物理学的な手法、局所計算アルゴリズムの利用、正則化など、様々な問題に使える一般的な手法を示すことができた。その一方で下限を得るための方法論の確立までには至らなかった。(3)に関しては、機械学習やその応用を専門としている研究者との共同研究により、スペクトラルクラスタリング、k-means、決定木学習、階層クラスタリングについての成果を得ることができた。(4)については、平均感度をさらに精緻化したリプシッツ性についての結果を得ることができた。これらの成果を踏まえると、当初予定していた研究目的の多くは達成できたといえる。
- 研究の進め方：計算機サーバを購入することで計算機実験を効果的に進めることができた。基本的には個人研究であったが、リサーチアシスタントとして隈部壮氏を迎えることで理論研究を効果的に進めることができた（代表的な論文 2 など）。他にも研究補助員が雇用できる予定であったが、残念ながら先方の病気のため実質的な研究を進めることはできなかった。

研究成果の科学技術及び社会・経済への波及効果：科学技術への波及効果という意味では、平均感度やリプシッツ性といったアルゴリズムの新しい評価軸を導入したことで、これまでに考えられていた多くの問題を問い直すことができ、その影響は大きい。平均感度は様々な計算モデルと関連があることが分かっており、それらとの間を行き来することで新たな結果が生み出されると期待することができる（例えば局所計算モデルにおける下限を平均感度の下限から導出することに成功している）。またワークショップや学会での発表を通じて、アルゴリズムの安定化は企業でのニーズがあることが確認できたので、企業研究者などと連携をとりながら平均感度やリプシッツ性の概念を波及させていきたい。

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. Nithin Varma and Yuichi Yoshida. Average Sensitivity of Graph Algorithms. Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA). 2021, 684—703.
平均感度をグラフの問題に対して定義し、最小全域木問題、最小カット問題、最大マッチング問題など様々な問題に対して平均感度の解析を行った。また平均感度の小さいアルゴリズムの設計に局所計算アルゴリズムが用いることができることを示し、前者の二彩色問題に対する下限から後者の下限を示すことに成功した。
2. Soh Kumabe and Yuichi Yoshida. Average Sensitivity of Dynamic Programming. Proceedings of the 33rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). 2022, 1925—1961.

動的計画法を用いて解くことができる様々な問題に対して、平均感度の低いアルゴリズムの設計を行った。最初に、多くの動的計画法の計算プロセスを表現できる最大重み鎖問題を考え、そこに帰着することで最長増加部分列問題、最長共通部分文字列問題、ナップサック問題、RNA 畳み込み問題に対する平均感度の低いアルゴリズムを構築した。
3. Yuichi Yoshida and Shinji Ito. Average Sensitivity of Euclidean k-Clustering. 36th Conference on Neural Information Processing Systems (NeurIPS). 2022.
与えられた点集合をk個のクラスタに分割する問題の平均感度を調べた。まずk-means++と呼ばれる$O(\log n)$近似アルゴリズムの解析を行い、その平均感度が$O(k)$であることを示した。次に、任意の近似アルゴリズムに対して、近似度を少し下げるだけで平均感度を$O(\mathrm{poly}(d, k, \log n))$に下げられることを示した。これらの平均感度が動的アルゴリズムやオンラインアルゴリズムの設計に応用できることを示した。

(2)特許出願

特になし

(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)

- 原聡, 吉田悠一. 決定木学習の安定化. 第 25 回情報論的学習理論ワークショップ 最優秀プレゼンテーション賞. 2022.
- Soh Kumabe and Yuichi Yoshida. Lipschitz Continuous Algorithms for Graph Problems. Submitted.
- Satoshi Hara, Koh Takeuchi, and Yuichi Yoshida. Average Sensitivity of Hierarchical Clustering. Submitted.