様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ４ 月 ２７ 日現在

機関番号：１４７０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ０４１４１

研究課題名（和文）ビッグデータ解析のためのデータ圧縮法の開発

研究課題名（英文）Investigation of data compression algorithms for big data analysis

研究代表者
葛岡　成晃（Kuzuoka, Shigeaki）

和歌山大学・システム工学部・教授

研究者番号：６０４５２５３８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：関数計算のためのデータ圧縮，および，多端子仮説検定問題に関する研究に取り組んだ．関数計算のためのデータ圧縮に関する研究では，主に関数の引数が2の場合を扱っていた既存結果を引数が3の場合に拡張する成果を得た．さらに，計算を行うノードに観測データを送信する際の通信路雑音を考慮した設定で研究を進め，通信路雑音を考慮しない従来研究の結果を拡張することに成功した．多端子仮説検定問題に関する研究では，複数の達成可能最適誤り指数の間に成り立つ一般的な関係式を導出した．
さらに副次的な成果として，情報源符号化と推測問題との関係性に関する成果や，VF情報源符号化に関する成果も得られた．

研究成果の学術的意義や社会的意義
情報機器の発達により蓄積されるデータが飛躍的に増大しつつある現在，従来のデータ処理技術では扱うことが困難な膨大なデータ（ビッグデータ）を効率よく扱うための手法が必要とされている．具体的には，観測データをネットワークを介して通信する際に，データの宛先で実行するデータ処理に適した形で通信するためのデータ圧縮技術が必要になる．本研究の成果は，そのような通信のためのデータ圧縮技術開発に貢献するものである．

研究成果の概要（英文）：Problems of data compression for function computation and multiple terminal hypothesis testing are investigated.
In the study of data compression for function computation, we extended the existing result, which mainly dealt with the case where the argument of the function is two, to the case where the argument is three. In addition, the study was conducted in a setting that takes into account the channel noise between data-observers and the node performing the computation, and extended the existing result that did not take into account the channel noise. In the study of the multiple terminal hypothesis testing, we derived general formulas which show relationship among multiple achievable error exponents. As secondary results, we also obtained results on the relationship between source coding and guessing problems, as well as results on VF source coding.

研究分野：情報理論

キーワード：情報理論　多端子仮説検定　分散関数計算

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

情報機器の発達により蓄積されるデータが飛躍的に増大しつつある現在，従来のデータ処理技術では扱うことが困難なほど膨大なデータ，いわゆるビッグデータを効率よく扱うための手法が必要とされている．このような背景から，データ圧縮という基本的な技術についても，従来とは異なる視点から研究・開発することが求められる．具体的には，下記のような視点が重要である．(i) 本当に必要とされているのはデータそのものではなく，そこから引き出される知見であること．(ii) 例えば複数のセンサが観測したデータをサーバに送るといった，ネットワークを介したデータのやり取りが前提とされること．

このような視点に立つと，従前の「データをハードディスクに保存して，後日そのデータを読み出す」ことだけを目的としたデータ圧縮技術では，次世代の情報処理のための基礎技術としては不十分であり，情報の収集・取捨選択・共有といったプロセスまで考慮した，新しいデータ圧縮技術が必要とされる．このような問題は，情報理論分野では，「多端子仮説検定問題」および「分散関数計算問題」として定式化され研究されてきた．多端子仮説検定および分散関数計算では，「送信データ量」と「判断・計算の正確さ」との最適なトレードオフ，および，それを達成する最適な符号化方法が研究されている（より正確には，データに基づく検定を目的とするのが「多端子仮説検定問題」であり，データを処理した計算結果を得ることを目的とするのが「分散関数計算問題」である）．ただし，これらの問題は一般的には未解決であり，例えば何らかの制約や仮定の下での特別な場合についてのみ最適なトレードオフが求まっている状況であった．

２．研究の目的

前項で述べた背景に対して，本研究では，下記(1)(2)を目標とする．

(1) 多端子仮説検定および分散関数計算における未解決問題に取り組む．具体的には，従来研究で課されていた制約条件や仮定についてより詳細に検討し，より弱い条件・仮定で同様の結果が得られることを証明する．

(2) 上記(1)の研究で得られた成果・知見に基づいて，具体的なデータ圧縮アルゴリズムを開発する．

３．研究の方法

上述の目的を達成するため，以下の方針・手法で研究を進める．

(1) まず分散計算問題に取り組む．特に，符号器が２つだけの２対１のネットワーク通信における分散計算問題を取り上げ，「送信データ量」と「判断・計算の正確さ」との最適なトレードオフの限界について研究する．

(2) ２対１のネットワーク通信における分散計算問題のためのデータ圧縮法の開発に取り組む．単なる試行錯誤ではなく，理論研究において得られた知見に基づいて，具体的なアルゴリズムを考案する．

(3) 研究の進捗に応じて，多端子仮説検定問題についても，同様に理論研究およびそれに基づくデータ圧縮法の開発に取り組む．

４．研究成果

（１）関数計算のためのデータ圧縮法に関する成果

関数計算のためのデータ圧縮法に対する理論的な最適レート領域を求めることは一般的には未解決問題であるが，関数の二分法に着目することで最適レート領域を決定できることが知られていた．従来研究では関数の引数が 2 の場合を中心に研究されていたが，本研究では関数の引数が 3 の場合に拡張する結果を得た（文献 1）．

さらに，計算を行うノードに観測データを送信する際の通信路雑音を考慮した設定で研究を進めた．前述の従来研究では，計算を行うノードに観測データを送信する際の通信路雑音は無視できると仮定されていた．これは暗に，ノード間の通信で通信路符号化を利用していることを仮定している．しかしながら，一般的なネットワーク通信の場合，情報源符号化と通信路符号化を同時に最適化したほうがよりよい結果になることが知られている．そこで本研究では「関数計算のための情報源-通信路結合符号化問題」について研究し，通信路雑音を考慮しない従来研究の成果を拡張することに成功した．この結果は国際会議 ITW で発表した（文献 2）．

（２）多端子仮説検定問題の誤り指数に関する成果

第２種の誤り確率の達成可能な最適値は，第１種の誤りに対する制約条件の違いにより，複数の定義があり得る．本研究では，複数の達成可能最適誤り指数の間に成り立つ一般的な関係式を導出した．さらに，一対多の有歪み情報源符号化に関する研究も行った．この場合，誤り確率（歪みが許容値を超える確率）に対する制約の違いにより，達成可能な最適レートが複数定義される．本研究では，他端子仮説検定に関する成果と同様の証明手法を用いて，複数の達成可能最適レー

トの間に成り立つ一般的な関係式を導出した（文献 3）.

さらに，ネットワーク通信を念頭に，二つの離れた地点間で双方向通信が許される場合の多端子仮説検定問題について検討した．既存研究ではデータの観測地点からセンターへの一方向通信を行う場合のみ考えられており，双方向通信が許される場合についての研究は十分ではなかった．それに対して本研究は，二つの地点間でやり取りされる情報量の総ビット数がある条件を満たす場合には，双方向通信は誤り指数を改善しないことを証明できた.

（３）情報源符号化と推測問題との関係性に関する成果

データ圧縮法に関する研究の副産物として，情報源符号化と推測問題（観測データに関する推測を行う推測を行う問題）との関係性に関する成果が得られた．ビッグデータに対して推測を行う場合，全ての結果を列挙して正解に到達するまで推測を行うことは現実的ではなく，途中で推測を打ち切ることが必要になる．本研究では，このように途中で推測を打ち切ることを許容する推測問題と，復号誤りを許す情報源符号化との関係を考察した．具体的には，条件付スムース Renyi エントロピーの新たな定義を提唱し，情報源符号化問題および推測問題のいずれの問題に対しても，提唱した条件付スムース Renyi エントロピーを用いることで符号化定理を記述できることを証明した（文献 4）.

この研究成果は，さらに発展させる事が出来た．具体的には，Salamatian らによって 2019 年に提案された推測問題の一種である非同期推測問題について取り上げ，推測結果が真の値と完全に一致していない場合も許容する，いわゆる有歪み推測の場合に Salamatian らの結果を拡張した．この結果は，情報理論のトップカンファレンスである ISIT に採録された（文献 5）.

（４）VF 情報源符号化に関する成果

データ圧縮法に関する研究の副産物として，VF（可変長-固定長；Variable-Length to Fixed-Length）情報源符号化に関する成果が得られた．本研究では，主に FV（固定長-可変長；Fixed-Length to Variable-Length）情報源符号化について考察してきた．しかしながら，FV 符号化と比較すると，符号語長が一定になる VF 符号化はデータ圧縮後のファイルの取り扱いが容易になるという長所がある．これは，実用化の観点からすると重要な点である．本研究では，定常性を満たすとは限らない一般的な情報源に対する VF 情報源符号化に関して取り組んだ．とくに，オーバーフロー確率，すなわち，符号化率が所与の閾値を超える確率の性質を考察した．その結果，情報源のエントロピー・スペクトル上限によって，最適な達成可能閾値が特徴づけられることを証明した（文献 6）.

＜引用文献＞

1. 上木成樹，葛岡成晃，“3 入力関数を計算するための分散符号化法の達成可能領域に関する一検討，” 信学技報，IT2020-6，pp. 31-36，2020 年 5 月.
2. N. Joki and S. Kuzuoka, “A study on the joint source-channel coding for computing functions: An approach from a dichotomy of functions,” in Proc. of 2021 IEEE Information Theory Workshop (ITW2021), Kanazawa, Japan, Oct. 17-21, 2021.
3. S. Kuzuoka, “A unified approach to error exponents for multiterminal source coding systems,” IEICE Trans. Fundamentals, vol. E101-A, no. 12, pp. 2082-2090, Dec. 2018.
4. S. Kuzuoka, “On the conditional smooth Renyi entropy and its applications in guessing and source coding,” IEEE Trans. Inform. Theory, vol. 66, no. 3, pp. 1674-1690, Mar. 2020.
5. S. Kuzuoka, “Asynchronous guessing subject to distortion,” in Proc. of 2021 IEEE International Symposium on Information Theory (ISIT2021), pp.2008-2012, Melbourne, Australia, July 12-20, 2021.
6. S. Kuzuoka, “A study on the overflow probability of variable-to-fixed length codes,” in Proc. Of 2020 International Symposium on Information Theory and its Applications (ISITA2020), pp.26-30, October 24-27, 2020.

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Shigeaki Kuzuoka	４．巻 66
２．論文標題 On the conditional smooth Renyi entropy and its applications in guessing and source coding	５．発行年 2020年
３．雑誌名 IEEE Transactions on Information Theory	６．最初と最後の頁 1674-1690
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TIT.2019.2937318	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shigeaki Kuzuoka	４．巻 E101-A
２．論文標題 A Unified Approach to Error Exponents for Multiterminal Source Coding Systems	５．発行年 2018年
３．雑誌名 IEICE TRANS. FUNDAMENTALS	６．最初と最後の頁 2082--2090
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E101.A.2082	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計13件（うち招待講演　0件／うち国際学会　7件）

１．発表者名 N. Joki and S. Kuzuoka
２．発表標題 A study on the joint source-channel coding for computing functions: An approach from a dichotomy of functions
３．学会等名 2021 IEEE Information Theory Workshop（国際学会）
４．発表年 2021年

１．発表者名 S. Kuzuoka
２．発表標題 Asynchronous guessing subject to distortion
３．学会等名 2021 IEEE International Symposium on Information Theory（国際学会）
４．発表年 2021年

１．発表者名
N. Joki and S. Kuzuoka

２．発表標題
A study of computability in distributed computing with joint sourcechannel coding over multiple-access channel

３．学会等名
第44回情報理論とその応用シンポジウム

４．発表年
2021年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
A study on the overflow probability of variable-to-fixed length codes

３．学会等名
2020 International Symposium on Information Theory and its Applications（国際学会）

４．発表年
2020年

１．発表者名
上木成樹，葛岡成晃

２．発表標題
関数計算のための情報源・通信路結合符号化に関する研究～関数の二分法によるアプローチ～

３．学会等名
電子情報通信学会情報理論研究会

４．発表年
2021年

１．発表者名
上木成樹，葛岡成晃

２．発表標題
3入力関数を計算するための分散符号化法の達成可能領域に関する一検討

３．学会等名
電子情報通信学会情報理論研究会

４．発表年
2020年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
On the conditional smooth Renyi entropy and its application in guessing

３．学会等名
2019 IEEE International Symposium on Information Theory（国際学会）

４．発表年
2019年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
A study on the overflow probability of variable-to-fixed length codes

３．学会等名
第42回情報理論とその応用シンポジウム

４．発表年
2019年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
A study on variable-to-fixed length coding of general sources

３．学会等名
電子情報通信学会情報理論研究会

４．発表年
2019年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
On overflow probability of variable-to-fixed length codes for non-stationary sources

３．学会等名
the 11th Asia-Europe Workshop on Concepts in Information Theory（国際学会）

４．発表年
2019年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
On the smooth-Renyi entropy and guessing allowing error

３．学会等名
第41回情報理論とその応用シンポジウム

４．発表年
2018年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
Properties and applications of the smooth Renyi entropy

３．学会等名
AMS Sectional Meeting（国際学会）

４．発表年
2019年

１．発表者名
Shigeaki Kuzuoka

２．発表標題
A study on the problem of channel resolvability for channels with countable input alphabet

３．学会等名
2018 International Symposium on Information Theory and its Applications（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関