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　２ 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：若手研究(B)
研究期間：2015～2019
課題番号：１５Ｋ１５９３５

研究課題名（和文）有限長のデータに対するデータ圧縮における圧縮率の理論限界の解析

研究課題名（英文）An Analysis of the Theoretical Limit of the Compression Rate in Data Compression for Finite-Length Data

研究代表者
松田　哲直（Matsuta, Tetsunao）

東京工業大学・工学院・助教

研究者番号：００６３８９８４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究では、有限長のデータに対する圧縮率の理論限界に対して精度の高い上界と下界を与えることを目的とし、次の主要成果を得た。(1)データを固定長の圧縮データに変換し、歪みを許容して元データに戻す場合のデータ圧縮に対して、精度の高い理論限界の上界と下界を与えた。特に上界については、これまで知られている中で最も精度が高くなることを数値計算で示した。(2)複数のデータを独立に分散して圧縮する場合のデータ圧縮をいくつか扱い、圧縮率の理論限界に対する新たな上界と下界を与えた。また、上界と下界の精度が高くなる場合があることを数値計算によって示した。

研究成果の学術的意義や社会的意義
本研究の成果によって、現実的な有限のデータ長におけるデータ圧縮法設計の有用な指標を与えることができた。さらに、分散してデータ圧縮する場合の圧縮率の理論限界についても多くの知見を与えることができた。以上から、近年のデータ量肥大化に対処するためのデータ圧縮における、理論的かつ現実的な基礎を築くことができたといえる。

研究成果の概要（英文）：The purpose of this study is to give accurate upper and lower bounds on the theoretical limit of the compression rate for finite-length data. The following main results are obtained. (1) Upper and lower bounds on the theoretical limit are given for a data compression that converts data into compressed data of fixed length and reproduces the original data allowing distortions. In particular, it is shown by a numerical calculation that the upper bound is the most accurate known so far. (2) Upper and lower bounds on the theoretical limit are given for several distributed data compressions that compress multiple data independently. It is also shown by a numerical calculation that the upper and lower bounds are quite accurate in some cases.

研究分野：情報理論

キーワード：情報源符号化　情報理論　データ圧縮　有限長

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

我々が普段パソコンやモバイル端末などで扱うデータの量は年々増加している。特に、画像や動画のデータ量の肥大化は著しい。このような状況から、データ圧縮の重要性が近年は特に高まっているといえる。

データ圧縮の理論研究を行う情報理論では、圧縮率の理論限界（すなわち理論的な下限）の解析を行っている。圧縮率の理論限界はデータ圧縮法設計の重要な指標であり、また、理論限界導出の過程には実用的なデータ圧縮法設計の指針となる多くの知見が含まれているため、理論限界の解析は工学的に非常に重要である。情報理論では特に、データ長が無限大に漸近する場合における、漸近的な圧縮率の理論限界を解析することが多い。しかしながら、現実的にはデータ長が無限大になることは無いため、データ長が有限の場合の解析を行うことが重要である。例えば、画像の圧縮でよく用いられる JPEG では、画像を 8×8 画素のブロックに分割してそのブロックごとに圧縮を行うが、１画素は 24 ビットあれば十分に表すことができるため、圧縮するデータのデータ長は高々1536 ビットである。

このような有限のデータ長に対する理論限界の解析は、有限長解析や非漸近的解析と呼ばれ、最近になって盛んに研究されるようになった。実際には、有限長解析は以前から僅かながら行われていたが、それらの成果の多くが精密さに欠けていたことと、情報理論では従来から漸近的な解析を主として扱ってきたことから、あまり注目されてこなかった。

有限長解析におけるデータ圧縮は、圧縮データから元データに戻す際に歪みを許容する（有歪み）かしない（無歪み）かと、固定長か可変長かとの組み合わせによって大きく 4 つに大別できる。但し、固定長とは元データを固定長の圧縮データに変換することを意味し、可変長とは元データをそれに応じた可変長の圧縮データに変換することを意味する。2012 年に Kostina と Verdu はこれらのうち、有歪み固定長データ圧縮を扱い、圧縮率の理論限界に対する上界と下界を与えた。彼女らの上界と下界は、科研費申請当時の 2014 年に知られている中で最も精度の高いものであった。図１に彼女らの上界と下界の例を示す。この図は、バイナリデータが定常無記憶情報源から生起することを仮定した場合に、歪みが許容値を超えてしまう確率を基準値以下にできる圧縮率の理論限界の上界と下界を示している。

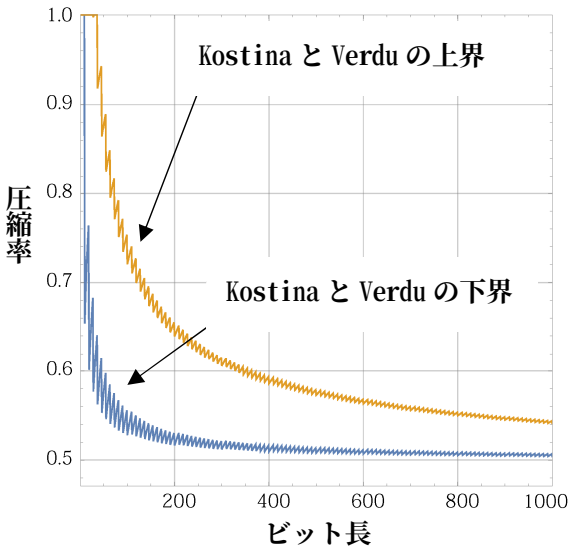

図１

但し、定常無記憶情報源とは、データの各ビットが統計的に独立に同一の分布に従って生起する情報源のことである。他方、無歪み及び有歪み可変長データ圧縮の圧縮率について、2014 年に Kostina らは理論限界の上界と下界を導出している。

有歪み固定長データ圧縮における Kostina と Verdu の成果では、データ圧縮の際にデータの生起確率が既知であることを仮定している。しかしながら、実際にはデータの生起確率を知ることは難しいため、生起確率が未知であっても動作するデータ圧縮が望まれる。これを達成するデータ圧縮はユニバーサルデータ圧縮と呼ばれているが、これに対する有限長解析はあまり詳細に行われていない。

また、上述したデータ圧縮では単一のデータを圧縮する場合を考えていたが、複数のデータを分散して圧縮する分散データ圧縮を考えることもできる。例えば、相関を有する複数の情報源から生起したデータを、それぞれ分散して独立に圧縮し、すべての圧縮データを同時に元のデータに戻すことを考えることもできる。分散データ圧縮の実用例としては、分散ストレージやセンサネットワークなどが挙げられる。分散データ圧縮の場合、分散したデータそれぞれに対して異なる圧縮率を与えることができるため、それらのトレードオフを評価するための圧縮率の領域を解析することになる。このとき、上界の代わりに内界を、下界の代わりに外界を解析する。

２．研究の目的

本研究の目的を簡単に述べると、データ圧縮の有限長解析における圧縮率の理論限界に対して、精度の高い上界と下界を与えること、及び有限長で優れた性能を有するユニバーサルデータ圧縮の開発を行うことである。具体的には、次の 4 つの課題に取り組む。

(1)　「固定長データ圧縮の理論限界」
有歪みや無歪みの固定長のデータ圧縮を扱い、精度の高い圧縮率の上界と下界を与えることを目指す。

(2)　「可変長データ圧縮の理論限界」
有歪みや無歪みの可変長のデータ圧縮を扱い、精度の高い圧縮率の上界と下界を与えることを目指す。

(3) 「ユニバーサルデータ圧縮の開発」
(1)と(2)のデータ圧縮に対して、圧縮率の理論限界を達成する、あるいはそれに近い値まで圧縮可能な、有限長で優れた性能を有するユニバーサルデータ圧縮を開発する。

(4) 「分散データ圧縮の有限長解析」
分散データ圧縮に対して有限長解析を行い、精度の高い圧縮率の内界と外界を与えることを目指す。また、有限長で優れた性能を有するユニバーサルデータ圧縮を開発する。

これらの課題を遂行して本研究の目的を達成することで、現実的なデータ長におけるデータ圧縮法設計の有用な指標を与えることができる。また、実用上重要なユニバーサルデータ圧縮の実現も期待できる。さらに、分散データ圧縮に対する有限長解析についても多くの知見を与えることができる。このように、本研究の目的を達成することには、近年のデータ量肥大化に対処するためのデータ圧縮における、理論的かつ現実的な基礎を築くことができるという意義がある。

３．研究の方法

研究目的で述べた 4 つの課題に対する研究の方法をそれぞれ述べる。

(1) 「固定長データ圧縮の理論限界」
有歪み固定長データ圧縮は無歪み固定長データ圧縮の一般化と考えることができるため、特に有歪み固定長符号化に注目する。理論限界の上界については、2012 年に Verdu によって与えられた一般的な上界の導出方法を応用して、さらに精度の高い上界を導出する。また、下界については、2012 年に Kostina と Verdu によって得られた下界が緩くなる原因を解明し、これを改良する方法を導出する。

(2) 「可変長データ圧縮の理論限界」
2014 年に Kostina らは、無歪み及び有歪み可変長データ圧縮の圧縮率の理論限界に対する上界と下界を与えている。また、彼女らは、定常無記憶情報源からデータが生起する場合における、上界と下界を閉じた式で与えている。定常無記憶情報源に対するこの上界と下界は精度が高く、これ以上の改善は容易ではないと考えられる。そこで、無記憶ではない情報源に対する上界と下界を閉じた式で与えることを検討する。

(3) 「ユニバーサルデータ圧縮の開発」
2010 年に研究代表者らは無歪み固定長データ圧縮に対して、有限長でも適用可能なユニバーサルデータ圧縮の設計手法を開発している。この手法を応用して、有限長で優れた性能を有するユニバーサルデータ圧縮を開発する。

(4) 「分散データ圧縮の有限長解析」
様々な分散データ圧縮に対して、上記(1)～(3)の解析方法を応用することで圧縮率の理論限界を導出する。特に、ユニバーサルデータ圧縮については、研究代表者らが開発したユニバーサルデータ圧縮の設計手法を応用する。

４．研究成果

以下では研究目的で述べた 4 つの課題に対する研究成果についてそれぞれ述べる。

(1) 「固定長データ圧縮の理論限界」
固定長データ圧縮の理論限界については次の成果を得ることができた。

① 有歪み固定長データ圧縮を扱い、圧縮率の理論限界に対する新たな上界と下界を与えた。また、定常無記憶情報源から生起したバイナリデータに対して、この上界を数値計算することで、Kostina と Verdu によって与えられていた最良の上界よりも真に精度が高くなることを示した。この数値計算例は図 2 に示した。図に示したとおり、研究代表者らの上界は Kostina と Verdu の上界よりも精度が高いことがわかる。

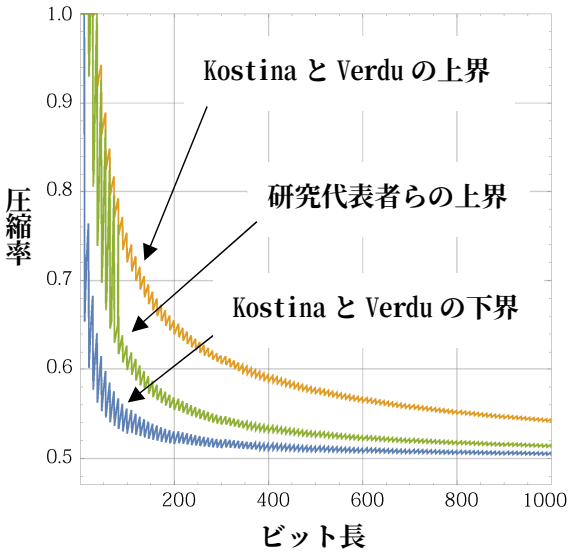

図 2

② 有歪みと無歪みデータ圧縮における圧縮率の理論限界が、Kolmogorov が導入した epsilon-entropy と呼ばれる量を用いて表せることを新たに示した。

(2) 「可変長データ圧縮の理論限界」
可変長データ圧縮の理論限界については、当初予定していたようなまとまった成果が得られていない。この理由としては、いくつかの詳細な研究が同時期になされていたことと、無記憶情報源以外の一般の情報源に対して、精度の高い閉じた式を与えることが困難であったことがあげられる。

(3) 「ユニバーサルデータ圧縮の開発」

研究代表者らが開発したユニバーサルデータ圧縮の設計手法を用いることで、有限長においてもユニバーサルデータ圧縮が可能であることが示された。特に、申請書らの手法は分散データ圧縮に対して有効であることが示されたが、この成果は以下の(4)-②において個別に述べる。

(4) 「分散データ圧縮の理論限界」

分散データ圧縮の理論限界については次の成果を得ることができた。

① 与えられたデータを複数の符号器それぞれにおいて異なる圧縮率で圧縮し、複数の復号器それぞれにおいて異なる歪みの許容値に基づいて圧縮データを元データに戻す場合のデータ圧縮を考える。このデータ圧縮に対して、データの定常性や無記憶性を仮定しなくても成り立つ、圧縮率の理論限界の領域に対する極めて一般的な内界と外界を新たに与えた。また、定常無記憶情報源については、データ長が無限大に漸近する場合における圧縮率の理論限界の領域が、これらの内界と外界から得られることを明らかにした。特に、定常無記憶情報源からバイナリデータが生起する場合に対して、理論限界に対する精度の高い上界と下界を閉じた式で与えた。

② 相関を有する複数の情報源から生起したデータをそれぞれ分散して独立に圧縮し、すべての圧縮データを同時に元のデータに戻すデータ圧縮を考える。このデータ圧縮に対して、データの生起確率が未知の場合に達成できる圧縮率の理論限界の領域に対する内界と外界を与えた。この内界は研究代表者らが開発したユニバーサルデータ圧縮の手法によって達成することができる。また、データ長が無限大に漸近する場合には、得られた内界と外界が一致することを明らかにした。

③ 上記②の場合とは異なり、すべての圧縮データを参照して一つのデータのみを元に戻す場合のデータ圧縮について検討した。このデータ圧縮に対して、圧縮率の理論限界の領域に対する新たな内界の導出方法を開発した。また、データ長が無限大に漸近する場合には、得られた内界と圧縮率の理論限界の領域とが一致することを明らかにした。

今後の展望としては、無記憶情報源以外の情報源に対して、精度の高い理論限界の上界と下界を閉じた式として導出することが挙げられる。これを達成する方法として、上述した等価な表現を利用することなどが考えられる。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Tetsunao Matsuta, Tomohiko Uyematsu	４．巻 E101.A
２．論文標題 Non-Asymptotic Bounds and a General Formula for the Rate-Distortion Region of the Successive Refinement Problem	５．発行年 2018年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 2110-2124
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E101.A.2110	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 Tetsunao Matsuta, Tomohiko Uyematsu	４．巻 E99.A
２．論文標題 New Non-Asymptotic Bounds on Numbers of Codewords for the Fixed-Length Lossy Compression	５．発行年 2016年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 2116-2129
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E99.A.2116	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 Tetsunao Matsuta, Tomohiko Uyematsu	４．巻 E102.A
２．論文標題 Achievable Rate Regions for Source Coding with Delayed Partial Side Information	５．発行年 2019年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 1631-1641
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E102.A.1631	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tetsunao Matsuta, Tomohiko Uyematsu	４．巻 -
２．論文標題 Coding Theorems for Asynchronous Slepian-Wolf Coding Systems	５．発行年 2020年
３．雑誌名 IEEE Transactions on Information Theory	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TIT.2020.2974736	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6件（うち招待講演　1件／うち国際学会　2件）

1．発表者名
Tetsunao Matsuta, Tomohiko Uyematsu

2．発表標題
Achievable Rate Regions for Source Coding with Delayed Partial Side Information

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

4．発表年
2018年

1．発表者名
Tetsunao Matsuta, Tomohiko Uyematsu

2．発表標題
Equivalent Representations for Two Source Coding Problems

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
Tetsunao Matsuta, Tomohiko Uyematsu

2．発表標題
A General Formula of the Achievable Rate Region for the Successive Refinement Problem

3．学会等名
電子情報通信学会ソサイエティ大会

4．発表年
2016年

1．発表者名
松田 哲直

2．発表標題
有歪み符号化における符号語数に対する有限長解析

3．学会等名
第39回情報理論とその応用シンポジウム（招待講演）

4．発表年
2016年

１．発表者名
Tetsunao Matsuta, Tomohiko Uyematsu

２．発表標題
Non-Asymptotic Bounds for Fixed-Length Lossy Compression

３．学会等名
2015 IEEE International Symposium on Information Theory（国際学会）

４．発表年
2015年

１．発表者名
Tetsunao Matsuta, Tomohiko Uyematsu

２．発表標題
Non-Asymptotic Bounds on Numbers of Codewords for the Successive Refinement Problem

３．学会等名
第38回情報理論とその応用シンポジウム

４．発表年
2015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