早稲田大学大学院 基幹理工学研究科

博 士 論 文 概 要

論 文 題 目

階層型の確率モデル族に基づく
ユニバーサル情報源符号化に関する研究

A Study on Universal Source Coding Based on Hierarchical Probabilistic Model Class

申 請 者

宮	希望
Nozomi	MIYA

数学応用数理専攻 情報理論研究

2016 年 5 月

情報理論の一分野である情報源符号化問題はデータ圧縮や機械学習等の基礎理論として近年盛んに研究が行われている．情報源符号化の代表的な問題設定として，ある確率構造を仮定した情報源から出力される情報源系列を符号語と呼ばれる別の系列に符号化し，符号語を伝送や蓄積した後，一切誤り無く元の情報源系列に復号する事が考えられる．これは無歪み情報源符号化と呼ばれる．無歪み情報源符号化において可変長，すなわち符号語長が情報源系列毎に異なる事を許容した場合，代表的な評価基準として平均符号語長が挙げられる．平均符号語長とは各情報源系列に対する符号語長の情報源の確率分布に関する期待値である．特に，情報源に定常エルゴードな確率分布を仮定した場合，任意の無歪み符号に対し，情報源系列１シンボルあたりの平均符号語長の下限が漸近的にエントロピーレートで与えられる事および１シンボルあたりの平均符号語長が漸近的にエントロピーレートに等しくなる，ある無歪み符号が存在する事がシャノンの情報源符号化定理として知られている．エントロピーレートとは情報源の確率分布から一意に定まる量である．さらに，情報源の確率構造を利用する事により，情報源系列１シンボルあたりの平均符号語長がエントロピーレートに漸近する実用的な符号化アルゴリズムまで提案されている．その代表的な例としてハフマン符号や算術符号等が挙げられる．

一方，ユニバーサル符号と呼ばれる情報源の確率構造の一部が未知である情報源符号化の問題は実用的に重要な研究となっている．特に，情報源が定常エルゴードな確率分布であれば，情報源系列１シンボルあたりの平均符号語長がエントロピーレートに漸近する符号を構成する事が重要なテーマとなってくる．ユニバーサル情報源符号化における代表的な問題設定として,情報源の確率構造に対し,例えば，情報源系列の出現確率が１次マルコフ過程と呼ばれる直前の値のみに依存する確率モデルに従う事は既知であるが遷移確率の確率パラメータは未知であるという様な仮定を置く事が挙げられる．さらに，出現確率が直前の長さｋの系列に依存するｋ次マルコフ過程である事，すなわち確率モデルはマルコフ過程であると仮定されているが次数ｋが未知，つまり何次のマルコフ過程かは未知であるという問題設定等も考えられている．前者をモデル既知の確率モデル，後者をモデル未知の確率モデルとそれぞれ呼ぶ事にする．

確率モデルの仮定に基づくユニバーサル符号においては，ある評価関数を導入した下でその評価関数の最適化に基づいた符号化に関する研究が一部で行われている．その際，評価関数として対数損失関数が導入されるのが一般的である．対数損失関数の情報源の確率分布に関する期待値は冗長度と呼ばれ，これは平均符号語長とエントロピーレートの差分を表す．特に，各時点において観測される情報源シンボルを逐次的に符号化する問題は，逐次的に観測されるデータを基に母集団分布を予測する統計的予測問題と等価であるとみなせ，さらに，冗長度は統計的予測問題におけるカルバックライブラー情報量と等価であるとみなせる.統

計的予測問題においては度々，仮定した確率モデルが母集団の確率分布を表現できるか否かという事が問題となる一方，ユニバーサル符号化問題では確率モデルに基づく符号化においてさえもそのような事はあまり議論されていない．

ユニバーサル符号は様々な問題設定や評価基準から研究が行われているが，確率モデルのパラメータに事前分布を導入し，符号語長をベイズ基準の下で最小化するベイズ符号は，理論的にも重要な符号である．情報源符号化においては，例えば圧縮対象のテキストの種類に応じて,ある単語の出現頻度がどの程度であり，さらにその頻度がそれまでの文脈にどの程度依存して変わってくるか等が事前に分かっている場合が多いため，パラメータに事前分布を仮定する事は比較的自然である．特にベイズ符号の場合，その性能を理論的に容易かつ精密に解析する事が可能であり，実際，平均符号語長がエントロピーレートに漸近する事が明らかにされているが，その漸近式は $o(1)$ まで明らかにされており，収束オーダーも明確である．さらに，情報源がマルコフ過程のように高次のモデルが低次のモデルを含む入れ子構造を持つ階層型のモデル未知の確率モデルに対し，ベイズ符号の平均符号語長が精密に解析されており，ベイズ符号の有効性が理論的に証明されている．

本研究ではまず，モデル既知の確率モデルに基づくベイズ符号において情報源の確率分布がその確率モデルでは表現されない分布だった場合の平均符号語長を精密に漸近評価している．さらに，その結果を応用し，階層型のモデル未知の確率モデルに基づくベイズ符号において，情報源の確率分布がその確率モデルでは表現されない分布だった場合の平均符号語長を精密に漸近評価し，そのように仮定が崩れた場合でもベイズ符号が有効である事を論じている．

本研究ではさらに，区分定常情報源に関するベイズ符号の応用について考察を行っている．区分定常情報源とは，モデル未知の確率モデルで表される情報源の一種であり，ある定常分布に従って情報源系列を発生させている情報源の確率パラメータがある時点で突然に変化し，その後またある時点まではその変化したパラメータを持つ定常分布に従って系列を発生させ，またパラメータが変化するといった情報源である．インターネット上でやり取りされるファイルのデータ圧縮等で利用されている bzip2 は，ブロックソート法と呼ばれるユニバーサル符号を実装したものであるが，これまでのところ，その符号化性能は理論的に明確に解析されていない．本研究では，モデル未知の確率モデルの一種である，次数 k が未知のマルコフ過程を仮定した情報源に対し，ブロックソート法の特徴的前処理であるブロックソート後の系列が区分定常情報源からの出力系列とみなせる事を利用し，区分定常情報源のベイズ符号を応用する事により，実用性をほとんど失う事無く理論的にも明確な符号化アルゴリズムを提案し，その有効性を論じている．

本論文の構成は以下の通りである．

第 1 章では本研究の背景および目的について述べる．

第 2 章では平均符号語長を評価基準とした無歪み可変長情報源符号化の研究背景を数理的に述べた上で確率モデルに基づくユニバーサル情報源符号化の問題設定について述べ，その中でも特に理論的に重要なベイズ符号を定義し，その性能解析に関する従来研究について述べる．

第 3 章では本研究成果の 1 つである，情報源の確率分布を表現できない確率モデルに基づくベイズ符号の性能解析について述べる．

第 4 章ではもう一方の成果である，ブロックソート法に区分定常情報源のベイズ符号を応用した符号化アルゴリズムを提案する．

最後に第 5 章では本研究成果をまとめ，結論及び今後の展望について述べる．

早稲田大学　博士（工学）　学位申請　研究業績書

氏 名　　宮 希望　　印

（2016 年 4 月　現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
1.論文	Evaluation of the Bayes Code from Viewpoints of the Distribution of Its Codeword Lengths IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E98-A, no.12, pp.2407-2414, Dec. 2015 Shota SAITO, Nozomi MIYA, Toshiyasu MATSUSHIMA
2.論文	Fundamental limit and pointwise asymptotics of the Bayes code for Markov sources Proceedings of 2015 IEEE International Symposium on Information Theory, pp.1986-1990, Hong Kong, June 2015 Shota Saito, Nozomi Miya, Toshiyasu Matsushima
3.論文○	Asymptotics of Bayesian Inference for a Class of Probabilistic Models under Misspecification IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E97-A, no.12, pp.2352-2360, Dec. 2014 Nozomi MIYA, Tota SUKO, Goki YASUDA, Toshiyasu MATSUSHIMA
4.論文	Evaluation of the minimum overflow threshold of Bayes codes for a Markov source Proceedings of 2014 International Symposium on Information Theory and Its Applications, pp.211-215, Melbourne, Australia, Oct. 2014 Shota Saito, Nozomi Miya, Toshiyasu Matsushima
5.論文	Asymptotics of MLE-based Prediction for Semi-supervised Learning Proceedings of 2014 International Symposium on Information Theory and Its Applications, p.343, Melbourne, Australia, Oct. 2014 Goki Yasuda, Nozomi Miya, Tota Suko, Toshiyasu Matsushima
6.論文○	Asymptotics of Bayesian estimation for nested models under misspecification Proceedings of 2012 International Symposium on Information Theory and its Applications, pp.86-90, Honolulu, USA, Oct. 2012 Nozomi Miya, Tota Suko, Goki Yasuda, Toshiyasu Matsushima
7.講演	プライバシー保護機能を持つ分散型正則化ロジスティック回帰に関する一考察 電子情報通信学会パターン認識・メディア理解研究会（PRMU），大阪府，2016 年 1 月 増井秀之，宮希望，松嶋敏泰
8.講演	一般情報源に対する Slepian-Wolf 符号化問題の 2 次の達成可能レート領域の別表現 電子情報通信学会情報理論研究会（IT），福岡県，2015 年 3 月 齋藤翔太，宮希望，松嶋敏泰
9.講演	非定常情報源に対する Bayes 符号のオーバーフロー確率における最小しきい値の評価 第 37 回情報理論とその応用シンポジウム（SITA2014），富山県，2014 年 12 月 守屋貴司，齋藤翔太，宮希望，松嶋敏泰

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
10.講演	定常エルゴードマルコフ情報源に対する Bayes 符号の符号語長の漸近正規性と重複対数の法則 第 37 回情報理論とその応用シンポジウム（SITA2014），富山県，2014 年 12 月 齋藤翔太，宮希望，松嶋敏泰
11.講演	パターン認識における AdaBoost の予測誤り率改善に関する一考察 電子情報通信学会情報論的学習理論と機械学習研究会（IBISML），茨城県，2014 年 9 月 増井秀之，都築遼馬，宮希望，松嶋敏泰
12.講演	推薦対象ユーザのクラスが未知の推薦問題におけるマルコフ決定過程を用いた推薦システムに関する一考察 電子情報通信学会情報理論研究会（IT），兵庫県，2014 年 7 月 岩井秀輔，宮希望，前田康成，松嶋敏泰
13.講演	真の分布を含むとは限らない階層モデル族に対するベイズ推定の漸近評価 第 36 回情報理論とその応用シンポジウム（SITA2013），静岡県，2013 年 11 月 宮希望，須子統太，安田豪毅，松嶋敏泰
14.講演	半教師付き学習における一致推定量に基づく予測の漸近評価 第 36 回情報理論とその応用シンポジウム（SITA2013），静岡県，2013 年 11 月 安田豪毅，宮希望，須子統太，松嶋敏泰
15.講演	ベイズ符号のオーバーフロー確率における最小しきい値の評価 第 36 回情報理論とその応用シンポジウム（SITA2013），静岡県，2013 年 11 月 齋藤翔太，宮希望，野村亮，松嶋敏泰
16.講演	真のモデルを含まないパラメトリックモデル族に対するベイズ予測の漸近評価 電子情報通信学会情報理論研究会（IT），岡山県，2011 年 7 月 宮希望，須子統太，安田豪毅，松嶋敏泰