早稲田大学大学院 基幹理工学研究科

博士論文審査報告書

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 7 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のみならず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では，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ことを考える．可変長，すなわち符号語長が情報源系列毎に異なることを許容した場合，符号の代表的な評価基準としては符号語長の情報源確率分布に関する期待値である平均符号語長が挙げられる．例えば，情報源に定常エルゴード確率過程を仮定した場合，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こと，さらにその限界を達成する具体的符号化法として，情報源系列 x^n の確率 $P(x^n)$ を用いたハフマン符号や算術符号等も示されている．

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問題は実用的にも重要な研究となっている． ユニバーサル符号の問題設定はいくつかあるがその代表的なものとして，情報源の確率構造を，パラメタライズされた確率分布を用い，確率質量関数または確率密度関数 $p(x^n \mid \theta_m, m)$ で表すことが多い．ここで，$m \in \mathcal{M}$ は確率モデルを示すインデックス，θ_m は $\Theta_m \subseteq \mathbf{R}^{k_m}$ 上の k_m 次元実数値パラメータである．例えば，m は m 次のマルコフ過程であることを示し，パラメータ θ_m は m 次マルコフ過程の遷移確率を表している．今後，混乱をまねかない場合，m で $p(x^n \mid \theta_m, m)$ で表されるモデルそのものを示すことも，そのモデルのクラスを $\mathcal{M}$ で示すこともあるとする．

このように表現された情報源の確率構造のどの部分を未知とするかによって，いくつかの問題設定が考えられる．ほとんどの問題設定でパラメータ空間 Θ_m は既知でありパラメータ θ_m については未知であると仮定しているが，モデル $m \in \mathcal{M}$ については一つのモデル m を仮定する場合（モデル既知と呼ぶ）と，モデルのクラス $\mathcal{M}$ のみを仮定し，モデル m は未知である場合（モデル未知と呼ぶ）の 2 つが考えられる．例えば，k 次のマルコフ過程であることを仮定する場合はモデル既知であり，1 次から k 次のマルコフ過程のどれかであると仮定する場合がモデル未知に対応する．

これらの確率構造の一部が未知の仮定の下で，符号化の為の確率質量関数または確率密度関数 $Ap(x^n)$ を決定することが，ユニバーサル符号の主要問題となり，リスク関数としては $Ap(x^n)$ を用いて符号化された平均符号語長とエントロピーレートとの差である冗長度で評価されることが多い．さらに情報源の確率構造の一部が事前にある程度仮定できる情報源符号化問題では，確率パラメータ θ_m やモデル m に事前分布を仮定することも多く，リスク関数を事前分布で期待値をとったベイズリスク関数を評価基準とした場合に最適な符号化であるベイズ符号が提案されている．モデル未知の場合のベイズ符号の符号化確率は以下となり，

$$Ap^*(x^n) = \sum_{\mathcal{M}} P(m) \int_{\Theta_m} p(x^n \mid \theta_m, m) w(\theta_m \mid m)\, d\theta_m, \tag{1}$$

モデルクラス $\mathcal{M}$ の全てのモデルの事前確率 $P(m)$ と，そのモデル m のもとでの確率パラメータ θ_m の事前確率 $w(\theta_m \mid m)$ で 2 重に重み付けすることが最適となることが示されている．

このユニバーサル符号の一括符号化確率 $Ap(x^n)$ の決定と同等なもう一つの表現である逐次符号化確率 $Ap(x_t|x^{t-1})$ の決定問題は，統計学の真の分布と予測分布間のカルバックライブラー情報量をリスク関数とした予測分布の決定問題と同等であることも示されている．両者の違いは，情報源符号化の場合は，符号化確率を決定することに主眼がおかれ，モデル m や確率パラメータ θ_m を特定せず，重み付けにより符号化確率を求めているが，統計学の場合は，未知であるモデルを AIC 等のモデル選択規準を用いて一つのモデル $\hat{m}$ に特定したもとで，パラメータも最尤推定量 $\hat{\theta}_{\hat{m}}$ 等で一つに固定して予測分布を求める手法が多く用いられている．

本研究は，ユニバーサル情報源符号化問題のベイズ符号について 2 つのテーマで研究を行なっている．

第 1 章では，序論，第 2 章では問題設定を述べている．

第 3 章では，一つ目のテーマについて論じており，従来研究でおかれていた仮定が崩れた場合のベイズ符号の冗長度の厳密な漸近評価を行なうことにより，仮定が崩れた場合でもベイズ符号が有効であることを明らかにしている．従来の研究では，一つのモデル m を仮定するモデル既知の問題設定でも，モデルのクラス $\mathcal{M}$ のみを仮定してモデル m は未知とするモデル未知の問題設定でも，情報源の真の分布 $p^*(x^n)$ が仮定されたモデルやモデルクラスで表現可能である，つまり $p^*(x^n) \in \{p(x^n \mid \theta_m, m) \mid \theta_m \in \Theta_m, m \in \mathcal{M}\}$ とする問題設定で研究が行なわれていた．特に，冗長度の Clarke らや以下の Gotoh らによる厳密な漸近評価がベイズ符号の有用性を示す重要な結果となっていた．

$$E\left[\log \frac{p^*(X^n)}{Ap^*(X^n)}\right] = \frac{k_{m^*}}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta^*_{m^*} \mid m^*)}}{w(\theta^*_{m^*} \mid m^*)} - \log P(m^*) + o(1). \quad (2)$$

ただし，E は $p^*(x^n)$ による期待値を表し，$I(\theta_m \mid m)$ はフィッシャー情報行列として以下の式で定義され，その極限値は存在すると仮定する．

$$I(\theta_m \mid m) = \lim_{n\to\infty}\frac{1}{n}E\left[\frac{\partial \log p(X^n \mid \theta_m, m)}{\partial \theta_m}\frac{\partial \log p(X^n \mid \theta_m, m)}{(\partial \theta_m)^T}\right]. \quad (3)$$

また，$p^*(x^n) = p(x^n \mid \theta^*_m, m)$ なる全ての θ^*_m の内，次数 k_m が最小となるものを $\theta^*_{m^*} \in \Theta_{m^*}$ と定義する．これは，m^* を次のように定義することと同等である．

$$m^* = \arg\min_{m\in\mathcal{M}}\left\{k_m \,\middle|\, {}^{\exists}\theta_m, {}^{\forall}x^n, p(x^n \mid \theta_m, m) = p^*(x^n)\right\}. \quad (4)$$

この結果から得られる重要な知見として，例えば，情報源の真の分布を表現可能な十分大きい次数 k のマルコフ過程のモデル一つを仮定したベイズ符号に対して，1 次から k 次までのマルコフ過程のクラスを仮定し，どの次数かは未知としたベイズ符号の方がエントロピーレートへの収束レートが速く，有用であることが示される．これは，前者は仮定された次数の k 次モデルのみを用いて圧縮を行なうのに対して，後者が高々 k 以下の次数となる真の分布を表現可能な最小次数のモデルを用いたものと等価な符号化を行なっていることに起因する．

本研究では，どちらの仮定においても，真の分布が表現不可能な場合について平均符号語長等を漸近的に評価し，両者の有用性も比較し考察している．統計学においても，真の分布が表現不可能な問題設定は，モデル選択問題における竹内の TIC や，最尤推定量に関する Nishii の研究等が行なわれており，本研究もそれらの重要な成果を取り入れて評価を行なっている．主要な結果の一つとしては，モデルクラスのみを仮定したベイズ符号において，そのクラスでは真の情報源分布が表現不可能な場合の冗長度の厳密な漸近評価を以下のように行なっている．

$$E\left[\log \frac{p^*(X^n)}{Ap^*(X^n)}\right] = \frac{k_{m^0}}{2}\log\frac{n}{2\pi} + D\left(p^n \| p^n_{\theta^0_{m^0}, m^0}\right)$$
$$-\frac{1}{2}\mathrm{tr}\left(I(\theta^0_{m^0} \mid m^0)J(\theta^0_{m^0} \mid m^0)^{-1}\right) + \log\frac{\sqrt{\det J(\theta^0_{m^0} \mid m^0)}}{w(\theta^0_{m^0} \mid m^0)} - \log P(m^0) + o(1). \quad (5)$$

ただし，$D\left(p^n \| p^n_{\theta^0_{m^0}, m^0}\right)$ は真の分布にカルバックライブラー情報量の意味で最も近い確率分布の内，最小次数のものと真の分布とのカルバックライブラー情報量であり，以下のように定義される．

$$D\left(p^n \| p^n_{\theta^0_{m^0}, m^0}\right) = E\left[\log\frac{p(X^n)}{p(X^n \mid \theta^0_{m^0}, m^0)}\right]. \quad (6)$$

また，$J(\theta_m \mid m)$ は以下の式で定義され，その極限値は存在すると仮定する．

$$J(\theta_m \mid m) = \lim_{n\to\infty} \frac{1}{n} E\left[-\frac{\partial^2 \log p(X^n \mid \theta_m, m)}{\partial\theta_m (\partial\theta_m)^T}\right]. \tag{7}$$

この結果から得られる興味深い知見は，例えば，真の分布が表現不可能な場合でも，一つの k 次のマルコフ過程モデルを仮定するのに対し，一見無駄な，より次数の小さい 1 次から k 次のモデル全部を含むクラスを仮定するベイズ符号の方が，エントロピーレートへの収束レートが同等かそれより速くなる場合もあり得ることが示された．

第 4 章では，２つ目のテーマを論じており，既に実装されている代表的データ圧縮アルゴリズムへのベイズ符号の適用の有効性を理論面から明らかにしている．ベイズ符号は前章の例に用いたマルコフ過程の他にも様々な情報源についても構成可能であり，区分定常情報源に関しても研究がされている．区分定常情報源とは，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ことを繰り返す情報源である．情報源のパラメタライズされた確率表現 $p(x^n \mid \theta_m, m)$ において，m は変化時点を表すインデックス，θ_m は変化点間の各区分の定常分布を表すパラメータとすることで，この区分定常情報源は表現される．

統計学の変化点検出問題においても，この種の確率過程を用いて重要な研究が行なわれているが，ここでも情報源符号化問題の場合はデータを圧縮する為の符号化確率を求めることが主眼であるため，ベイズ符号では，変化点を特定することはせず，可能性のある変化点全てを考慮して重み付け和をとることで確率を求めている．さらに，一般に重み付け和の計算量は符号語長の指数オーダとなるが，多項式オーダの効率的符号化アルゴリズムも提案されている．

また，インターネット上でやり取りされるファイルの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，その理論的な性能評価は十分でなかった．

本研究では，情報源にある種の階層的モデルクラスを仮定した場合，ブロックソート法の特徴的前処理である BW 変換後の系列が区分定常情報源からの出力系列と見なせることを利用し，区分定常情報源のベイズ符号のアルゴリズムを応用することにより，実用性を失うこと無く理論的にも明解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．

最後に，第 5 章で結論を述べ論文をまとめている．

以上を総括すると，本論文は，ユニバーサル符号において，従来から有用性が理論的に評価されていたベイズ符号の一部の仮定が崩れた場合の漸近評価を行い，その場合においてもモデルのクラスを仮定したベイズ符号が有用であることを示した点と，実装されているデータ圧縮アルゴリズムの処理の本質部分がベイズ符号で最適化できることに着目し，理論的に明解で実用性の高いアルゴリズムを提案した点は，理論面と実用面双方から高く評価できる．よって，本論文は博士（工学）の学位として価値あるものと認める．

2016 年 6 月
審査員

（主査）	早稲田大学教授	博士（工学）	（早稲田大学）	松嶋　敏泰
	早稲田大学教授	工学博士	（大阪大学）	谷口　正信
	早稲田大学教授	工学博士	（早稲田大学）	大石　進一
	早稲田大学教授	博士（工学）	（早稲田大学）	柏木　雅英
	早稲田大学名誉教授	工学博士	（大阪大学）	平澤　茂一