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

情報源符号化の有限長解析と漸近解析
―オーバーフロー確率に着目したアプローチ―

Non-Asymptotic and Asymptotic Analyses of Source Coding: An Approach from the Viewpoint of the Overflow Probability

申　請　者

齋藤	翔太
Shota	SAITO

数学応用数理専攻　情報理論研究

2018 年 2 月

現代社会には，多種多様な膨大な量のデータが存在している．このような世の中にあって，情報理論は，デジタルデータの収集，蓄積，伝送，解析のために，欠くことのできない基礎理論として重要な役割を果たしている．

本論文では，情報理論において主にデータの圧縮に関わる情報源符号化問題を扱っており，これは次のように数理モデル化される．まず，有限集合 $\mathcal{X}$ に値をとる離散確率変数 X_i から成る確率過程 $\{X_i\}_{i=1}^{\infty}$ を，情報源と呼ぶ．可変長無歪み情報源符号化の枠組みでは，時点 n までの情報源系列 $X_1, X_2, \ldots, X_n$（以下 X^n と表す）の実現値 $x_1, x_2, \ldots, x_n$（以下 x^n と表す）が，符号化関数 $f_n : \mathcal{X}^n \to \mathcal{U}^\star$ により変換される．ただし，$\mathcal{U} := \{0, 1, \cdots, K-1\}$（$K$ は 2 以上の任意の整数）であり，$\mathcal{U}^\star$ は $\mathcal{U}$ の要素からなる系列の集合を表す．このとき，系列 $f_n(x^n)$ を符号語という．また，その長さを符号語長と呼び，$\ell(f_n(x^n))$ と表す．そして，可変長無歪み情報源符号化においては，符号語は復号化関数 $g_n : \mathcal{U}^\star \to \mathcal{X}^n$ によって $\mathbb{P}[X^n \neq g_n(f_n(X^n))] = 0$ となるように元の情報源系列 x^n に復元される．

情報源符号化の研究では，数理モデルが定められた後，ある評価基準のもとで理論限界が導出される．可変長無歪み情報源符号化における代表的な評価基準として，情報源出力文字 1 個あたりの平均符号語長 $\frac{1}{n}\sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n)\ell(f_n(x^n))$ がある．ただし，$p_{X^n}(x^n)$ は X^n の確率関数である．例えば，定常無記憶情報源に対して，$n \to \infty$ のとき，情報源出力文字 1 個あたりの平均符号語長の理論限界は，シャノンエントロピー $H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_K p_X(x)$ であることが知られている．このように，ある評価基準の理論限界を，情報源の確率分布から定義される量を用いて特徴付けることが，情報理論の一つの研究目標であり，漸近解析（$n \to \infty$ となる条件での解析）と，近年では有限長解析（n が有限である場合の解析）も行われている．

理論限界が解明された後は，その限界に近づく符号化，復号化関数の組 (f_n, g_n)（これを符号と呼ぶ）を構成し，その性能評価が行われる．従来，情報源の確率分布が未知であったとしても，平均符号語長が理論限界（シャノンエントロピー）に近づく符号（ユニバーサル符号）が提案されており，その中でも，平均符号語長が理論限界に最も速い速度で収束する符号の一つにベイズ符号がある．情報源系列 x^n の確率関数が，未知パラメータ $\theta_*^k \in \Theta^k \subset \mathbb{R}^k$ によって $p_{\theta_*^k}(x^n)$ と表されており，確率関数族 $\{p_{\theta^k} : \theta^k \in \Theta^k \subset \mathbb{R}^k\}$ とパラメータ θ^k の事前確率密度関数 $w(\theta^k)$ が既知であるとき，ベイズ符号は，情報源系列 x^n の生起確率を $\int_{\Theta^k} w(\theta^k) p_{\theta^k}(x^n) d\theta^k$ と推定し，これを用いて算術符号化等を行う．ベイズ符号は，定常エルゴードマルコフ情報源等に対して，平均符号語長が定数項まで精密に評価される等，平均符号語長に関する研究が盛んに行われてきた．

以上のように，情報源符号化の主要な研究テーマとしては，ある評価基準のもとでの理論限界の導出と，具体的な符号の評価の二つが挙げられる．本論文では，符号語長がしきい値 $R \geq 0$ を超過する確率，すなわち，$\mathbb{P}\left[\frac{1}{n}\ell(f_n(X^n)) > R\right]$ と定義されるオーバーフロー確率という評価基準のもとで，これらの二つの研究課題にアプローチしている．

本論文では，まず第 1 章にて，研究背景と研究目的を述べている．次に，第 2 章では，情報源符号化のいくつかの代表的な数理モデルと，評価基準を説明している．

本論文の第 3 章から第 6 章では，論文の一つ目の研究テーマであるオーバーフロー確率という評価基準のもとでの理論限界の有限長解析を行っている．情報源系列長 n が有限の場合，符号語長は平均値に収束するとは限らないため，平均符号語長のみならず，符号語長の分布自体を考慮することが重要であり，符号語長の確率分布の裾を表すオーバーフロー確率は，有限長解析を行う際の一つの重要な評価基準である．従来，いくつかの研究により，オーバーフロー確率に関する理論限界が，自己情報量 $\frac{1}{n}\log_K \frac{1}{p_{X^n}(X^n)}$ を用いて特徴付けられていた．これに対して，本論文では，いくつかの代表的な問題設定に対する理論限界が，すべて smooth 最大エントロピーやそれに基づく量を用いて特徴付けられることを示している．ここで，smooth 最大エントロピー $H^\delta(X)$ は，

$$H^\delta(X) = \min_{\substack{\mathcal{Z} \subset \mathcal{X}: \\ \mathbb{P}[X \in \mathcal{Z}] \geq 1-\delta}} \log_K |\mathcal{Z}|$$

と定義される．

第 3 章では可変長無歪み情報源符号化を扱っており，この章における証明のアイデアが，後の第 4 章や 5 章における証明のアイデアの基礎となっている．主要な結果の一つとして，$\mathbb{P}[\ell(f(X)) > R] \leq \delta$ を満たす符号が存在するようなしきい値 $R \geq 0$ の下限として定義される $R^*(\delta)$ を解析し，

$$H^\delta(X) - 1 < R^*(\delta) \leq \lfloor H^\delta(X) \rfloor$$

という結果を導いている．この結果は，最小しきい値 $R^*(\delta)$ が smooth 最大エントロピー $H^\delta(X)$ から（$K = 2$ の場合）1 ビット以内の範囲に入っていることを明らかにしている．また，不等式 $R^*(\delta) \leq \lfloor H^\delta(X) \rfloor$ の証明を，ランダム符号化の議論を用いるのではなく，明示的な符号を構成することで行っており，この考え方が第 4 章，5 章において主結果を導く際の基礎になっている．

第 4 章では，第 3 章にて扱った復号誤り確率が零という条件を拡張し，復号誤り確率を許容した可変長情報源符号化を扱っている．この章では，主要な結果の一つとして，$\mathbb{P}\left[\ell(f(X)) > R\right] \leq \delta$ かつ $\mathbb{P}[X \neq g(f(X))] \leq \epsilon$ を満たす符号が存在するようなしきい値 $R \geq 0$ の下限として定義される $R^*(\epsilon, \delta)$ を解析し，

$$H^{\epsilon+\delta}(X) - 1 < R^*(\epsilon, \delta) \leq \lfloor H^{\epsilon+\delta}(X) \rfloor$$

という結果を導いている．この結果は，$R^*(\epsilon, \delta)$ が δ と ϵ の和で定まる $H^{\epsilon+\delta}(X)$ で特徴付けられることを明らかにしている．

上述のように，第 4 章では復号誤り確率を扱っているが，これを一般化した問題設定として，第 5 章では可変長有歪み情報源符号化を扱っている．特に，本論文では，情報源シンボル $x \in \mathcal{X}$ と復号後のシンボル $y \in \mathcal{Y}$ との間の違いを測る歪み測度を $d : \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$ とするとき，歪み超過確率 $\mathbb{P}[d(X, g(f(X))) > D]$ を扱っている．主要な結果の一つとして，$\mathbb{P}\left[\ell(f(X)) > R\right] \leq \delta$ かつ $\mathbb{P}[d(X, g(f(X))) > D] \leq \epsilon$ を満たす符号が存在するようなしきい値 $R \geq 0$ の下限として定義される $R^*(D, \epsilon, \delta)$ を解析し，

$$G^{\delta}_{D,\epsilon}(X) - 1 < R^*(D, \epsilon, \delta) \leq \lfloor G^{\delta}_{D,\epsilon}(X) \rfloor$$

という結果を導いている．ただし，$G^{\delta}_{D,\epsilon}(X)$ は smooth 最大エントロピーをもとに定義される量である．この結果は，$D \geq 0$ が与えられたとき，$R^*(D, \epsilon, \delta)$ が δ と ϵ の和で定まる $G^{\delta}_{D,\epsilon}(X)$ で特徴付けられることを明らかにしている．

第 3 章から第 5 章までは，オーバーフロー確率のもとでの可変長情報源符号化を扱っている．この問題と，固定長情報源符号化（符号語長が一定である情報源符号化）との間には密接な関係があることが知られている．そこで，第 6 章では固定長情報源符号化に着目し，固定長情報源符号化における代表的な問題の一つである Slepian-Wolf 情報源符号化（相関を持つ 2 つの情報源からの出力系列を 2 つの符号器により符号化し，1 つの復号器で復号するような情報源符号化）の解析を行っている．この問題に対しては，復号誤り確率が漸近的に ϵ 以下となるような符号化レートの組として定義される ϵ-達成可能レート領域を求めることが一つの研究課題である．本論文では，2 次の ϵ-達成可能レート領域と呼ばれる達成可能レート領域を評価し，smooth 最大エントロピーと関係のある量を用いてこれを特徴付けている．

本論文の第 7 章と第 8 章では，論文の二つ目の研究テーマであるオーバーフロー確率のもとでのベイズ符号の漸近解析を行っている．ベイズ符号は平均符号語長という評価基準のもとでは，様々な評価が行われてきたが，オーバーフロー確率という評価基準のもとでの性能は従来明らかではなかった．そこで，本論文では，オーバーフロー確率という評価基準のもとで，ベイズ符号を解析し，ベイズ符号の新たな性能を明らかにしている．

第 7 章では，ベイズ符号のオーバーフロー確率が与えられた定数 ϵ を超えないという条件のもとで，最小しきい値を解析している．ベイズ符号の符号語長を $\ell_B(\cdot)$ と表し，この最小しきい値を $R^*_{\ell_B}(n, \epsilon, \theta^k_*)$ と表すとき，

$$R^*_{\ell_B}(n, \epsilon, \theta^k_*) \leq H_{\theta^k_*}(\mathbf{X}) + \frac{\sigma_{\theta^k_*}(\mathbf{X})}{\sqrt{n}} Q^{-1}(\epsilon) + \frac{k}{2n} \ln n + O\left(\frac{1}{n}\right),$$
$$R^*_{\ell_B}(n, \epsilon, \theta^k_*) \geq H_{\theta^k_*}(\mathbf{X}) + \frac{\sigma_{\theta^k_*}(\mathbf{X})}{\sqrt{n}} Q^{-1}(\epsilon) + \frac{k}{2n} \ln n + \frac{C_l(n)}{n} + O\left(\frac{1}{n}\right)$$

が成り立つことを示している．ただし，$H_{\theta^k_*}(\mathbf{X}) := \lim_{n\to\infty}(1/n) E_{p_{\theta^k_*}}\left[\ln(1/p_{\theta^k_*}(X^n))\right]$，$\sigma_{\theta^k_*}(\mathbf{X}) := \sqrt{\lim_{n\to\infty}(1/n) V_{p_{\theta^k_*}}\left[\ln(1/p_{\theta^k_*}(X^n))\right]}$（$E_{p_{\theta^k_*}}[\cdot]$ と $V_{p_{\theta^k_*}}[\cdot]$ はそれぞれ $p_{\theta^k_*}$ による期待値と分散を表す）であり，$Q^{-1}(z)$ は $Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$ の逆関数，$C_l(n)$ は $o(\ln n)$ であり $\frac{1}{\sqrt{n}}\{C_l(n) + O(\ln \ln n)\} \to -0$ を満たす負の項である．この結果から明らかになった興味深い知見の一つは，ベイズ符号の最小しきい値の上界と下界の第 1，2 項，すなわち，$H_{\theta^k_*}(\mathbf{X})$ と $\frac{\sigma_{\theta^k_*}(\mathbf{X})}{\sqrt{n}} Q^{-1}(\epsilon)$ は，オーバーフロー確率が最小となる非ユニバーサル符号の上界と下界の第 1，2 項と一致することである．ベイズ符号は，ベイズ基準のもとで平均符号語長を最小にするよう設計された符号であり，オーバーフロー確率が最小になるように設計された符号ではない．それにも関わらず，この結果は，ベイズ符号がオーバーフロー確率最小の符号と比較しても $O(1/\sqrt{n})$ までは同様の性能を示すことを明らかにしている．

第 8 章では，オーバーフロー確率が漸近的に零となる条件のもとで，ベイズ符号のオーバーフロー確率を評価している．この章で得られた主要な結果の一つとしては以下が挙げられる．まず，$\{R_n\}_{n=1}^\infty$ を $R_n = (1/n) H_{\theta^k_*}(X^n) + \tau_n$ なる数列とする．ただし，$\{\tau_n\}_{n=1}^\infty$ は，$\lim_{n\to\infty} \tau_n = +0$ かつ $\lim_{n\to\infty} \sqrt{n}\tau_n = \infty$

を満たす数列である．このとき，本論文では，ベイズ符号のオーバーフロー確率を解析し，定常無記憶情報源に対して，

$$\lim_{n\to\infty} \frac{\ln \mathbb{P}_{\theta_*^k}[\ell_B(X^n) > nR_n]}{n\tau_n^2} = -\frac{1}{2\sigma^2_{\theta_*^k}(X)}$$

が成り立つことを明らかにしている．また，この結果と，従来知られているオーバーフロー確率が最小となるように設計された符号に対する同様の解析結果を比較することで，ベイズ符号がオーバーフロー確率最小の符号と同様の性能を示すことを明らかにしている．

最後に第 9 章において，本論文のまとめと今後の課題を述べている．

以上を総括すると，本論文では，まず，情報源符号化におけるいくつかの代表的な問題設定に対して，オーバーフロー確率という評価基準のもとでの，有限長の理論限界の導出に成功している．有限長の理論限界の解明の研究は，今後の技術の改善の余地を明らかにするという工学的な意義がある．次に，本論文では，オーバーフロー確率という評価基準のもとでのベイズ符号の漸近評価を行っている．ベイズ符号は理論評価のみならず，効率的なアルゴリズムも提案されている工学的にも重要な符号である．本論文では，ベイズ符号がオーバーフロー確率最小の符号と比較しても遜色のない性能を示すことを明らかにしており，ベイズ符号に関する新たな知見を得ることに成功している．以上のように，本論文では，オーバーフロー確率という評価基準のもとで，有限長解析と漸近解析の双方から解析を行い，これらのどちらの成果も理論面から評価でき，また工学的に意味をもつ成果と言える．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．

2017 年 11 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	（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
	早稲田大学教授	博士（数理科学）	（東京大学）	清水　泰隆
	早稲田大学名誉教授	工学博士	（大阪大学）	平澤　茂一