Graduate School of Fundamental Science and Engineering
Waseda University

博 士 論 文 概 要
Doctoral Thesis Synopsis

論 文 題 目
Thesis Theme

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

申 請 者
(Applicant Name)

Shota	SAITO
齋藤	翔太

Department of Pure and Applied Mathematics, Research on Information Theory

October, 2017

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，現代社会において欠くことのできない基礎理論である．

本論文では，情報理論における研究分野の中の情報源符号化問題を扱う．これは，主にデータの圧縮に関わる問題であり，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．確率過程にどのような確率構造を仮定するかによって，様々な情報源が定義される．例えば，$\{X_i\}_{i=1}^{\infty}$ が定常性とエルゴード性を満たす場合，定常エルゴード情報源と呼ばれる．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 $g_n : \mathcal{U}^\star \to \mathcal{X}^n$ によって誤りなく（すなわち，$\mathbb{P}[X^n \neq g_n(f_n(X^n))] = 0$ 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．理論限界導出の一つの意義は，限界が明らかになることによって，現在の技術に改善の余地があるのかどうかが明らかとなることである．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る．また，このように，理論限界の研究においては，その黎明期から漸近解析（$n \to \infty$ となる条件での解析）が盛んに行われてきた．しかしながら，現実的には，情報源系列 n の長さは有限であるため，近年，有限長解析（n が有限である場合の解析）も行われるようになっている．さらに，平均符号語長だけではなく，従来，問題設定に応じて様々な評価基準が提案され，様々な量を用いて理論限界が明らかにされている．例えば，ある問題設定に関しては，レニーエントロピー $H_\alpha(X^n) = \frac{1}{1-\alpha}\log_K\left(\sum_{x^n \in \mathcal{X}^n}(p_{X^n}(x^n))^\alpha\right)$ $(\alpha \in (0,1) \cup (1,\infty))$ や，これを一般化した smooth レニーエントロピー $H_\alpha^\epsilon(X^n)$ によって理論限界が特徴付けられている．

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．例えば，定常無記憶情報源に対して，$n \to \infty$ のとき，平均符号語長がシャノンエントロピーに近づく符号として，シャノン符号，ハフマン符号，算術符号等が知られている．これらの符号は，情報源の確率分布が既知である場合の符号である．一方，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）も提案されている．ユニバーサル符号については，従来様々なものが提案されているが，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 $p_{\theta_*^k}(x^n)$ と表されているとする．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．ベイズ符号については，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する．

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，評価基準を説明する．

本論文の一つ目の研究テーマは，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であり，第 3 章から第 6 章までがこれに対応している．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と言える．したがって，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と考えられる．このことが，理論限界の有限長解析において，本研究がオーバーフロー確率に着目する理由の一つである．また，オーバーフロー確率を評価基準とした可変長情報源符号化と，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）の間には密接な関係があり，オーバーフロー確率に関する理論限界を明らかにすることで，固定長情報源符号化における理論限界も明らかにすることができる．

従来，様々な問題設定に対して，オーバーフロー確率に関する理論限界が導出されているが，個別の問題ごとに，様々な証明手法や様々な量を用いて理論限界の特徴付けが行われていた．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す．ここで，smooth 最大エントロピー $H^\epsilon(X^n)$ は，$\alpha \in (0,1) \cup (1,\infty)$ に関して単調非増加な関数である smooth レニーエントロピー $H_\alpha^\epsilon(X^n)$ において，$\alpha \to 0$ の極限を取った量として定義される．また，smooth 最大エントロピー $H^\epsilon(X^n)$ は，

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

と書けることが知られている．情報源の確率分布が既知の場合，有限長において，この値は生起確率の大きな情報源シンボルを数え上げることにより計算が可能であることが，従来指摘されている．

理論限界の有限長解析に関する主結果として，第 3 章では可変長無歪み情報源符号化において，$\mathbb{P}[\ell(f(X)) > R] \leq \epsilon$ 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 $R^*(\epsilon)$ を解析し，$H^\epsilon(X) \leq R^*(\epsilon) \leq \lfloor H^\epsilon(X) + 1 \rfloor$ という結果を導いている．この結果から，最小しきい値 $R^*(\epsilon)$ が $H^\epsilon(X)$ から（$K = 2$ の場合）1 ビット以内の範囲に入っていることが明らかになった．さらに，第 4 章では復号誤りを許容した可変長情報源符号化，第 5 章では復号誤りを一般化した概念である歪みを許容した可変長情報源符号化，第 6 章では符号器が 2 つ，復号器が 1 つという Slepian-Wolf 情報源符号化について，smooth 最大エントロピーやそれに基づく量を用いて理論限界を解明している．

本論文の二つ目の研究テーマは，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解析である．上述したように，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．しかしながら，平均符号語長だけでなく，他の評価基準から符号の性能を評価することは，その符号に対する知見を深めるという意味で重要であると考えられる．そこで，本論文では，定常エルゴード有限次数のマルコフ情報源に対して，オーバーフロー確率という評価基準のもとで，ベイズ符号の性能を解析し，ベイズ符号の新たな性能を明らかにする．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，オーバーフロー確率の最小しきい値 $R_B^*(n, \epsilon, \theta_*^k)$ を解析し，

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

という結果を得た．ただし，

$$H_{\theta_*^k}(\mathbf{X}) := \lim_{n\to\infty} \frac{1}{n} E_{p_{\theta_*^k}} \left[\ln \frac{1}{p_{\theta_*^k}(X^n)} \right],$$
$$\sigma_{\theta_*^k}(\mathbf{X}) := \sqrt{\lim_{n\to\infty} \frac{1}{n} V_{p_{\theta_*^k}} \left[\ln \frac{1}{p_{\theta_*^k}(X^n)} \right]}$$

（$E_{p_{\theta_*^k}}[\cdot]$ 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，オーバーフロー確率が最小となる符号と比較しても $O(1/\sqrt{n})$ までは同様の性能を示すことが明らかになった．さらに，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，このような条件でも，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論じている．

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．

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

氏 名　齋藤 翔太　　印

(As of January, 2018)

種 類 別 (By Type)	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
1.論文○	Evaluation of Overflow Probability of Bayes Code in Moderate Deviation Regime IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E100-A, no.12, pp. 2728-2731, Dec. 2017 Shota SAITO, Toshiyasu MATSUSHIMA
2.論文	Spatially ``Mt. Fuji'' Coupled LDPC Codes IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E100-A, no.12, pp. 2594-2606, Dec. 2017 Yuta NAKAHARA, Shota SAITO, Toshiyasu MATSUSHIMA
3.論文○	Second-Order Achievable Rate Region of Slepian-Wolf Coding Problem in Terms of Smooth Max-Entropy for General Sources IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E99-A, no.12, pp. 2275-2280, Dec. 2016 Shota SAITO, Toshiyasu MATSUSHIMA
4.論文○	Threshold of Overflow Probability Using Smooth Max-Entropy in Lossless Fixed-to-Variable Length Source Coding for General Sources IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E99-A, no.12, pp. 2286-2290, Dec. 2016 Shota SAITO, Toshiyasu MATSUSHIMA
5.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
6. 講演○	Variable-Length Lossy Compression Allowing Positive Overflow and Excess Distortion Probabilities 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, June 2017 Shota SAITO, Hideki YAGI, Toshiyasu MATSUSHIMA
7. 講演○	Threshold of Overflow Probability in Terms of Smooth Max-Entropy for Variable-Length Compression Allowing Errors 2016 International Symposium on Information Theory and Its Applications (ISITA), Monterey, California, USA, Oct.-Nov. 2016 Shota SAITO, Toshiyasu MATSUSHIMA

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

種 類 別 By Type	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
8. 講演○	Evaluation of Overflow Probability of Bayes Code in Moderate Deviation Regime 2016 International Symposium on Information Theory and Its Applications (ISITA), Monterey, California, USA, Oct.-Nov. 2016 Shota SAITO, Toshiyasu MATSUSHIMA
9. 講演	Spatially “Mt. Fuji” Coupled LDPC Codes 2016 International Symposium on Information Theory and Its Applications (ISITA), Monterey, California, USA, Oct.-Nov. 2016 Yuta NAKAHARA, Shota SAITO, Toshiyasu MATSUSHIMA
10. 講演○	Fundamental Limit and Pointwise Asymptotics of the Bayes code for Markov Sources 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, June 2015 Shota SAITO, Nozomi MIYA, Toshiyasu MATSUSHIMA
11. 講演○	Evaluation of the Minimum Overflow Threshold of Bayes codes for a Markov Source 2014 International Symposium on Information Theory and Its Applications (ISITA), Melbourne, Australia, Oct. 2014 Shota SAITO, Nozomi MIYA, Toshiyasu MATSUSHIMA
12. 講演	微少なアンダーフロー確率を許した可変長 intrinsic randomness 問題 電子情報通信学会情報理論研究会（IT），千葉県，2017 年 7 月 吉澤潤，齋藤翔太，松嶋敏泰
13. 講演	Variable-Length Lossy Compression Allowing Positive Overflow and Excess Distortion Probabilities 第 39 回情報理論とその応用シンポジウム予稿集（SITA2016），岐阜県，2016 年 12 月 Shota SAITO, Hideki YAGI, Toshiyasu MATSUSHIMA
14. 講演	メッセージ伝搬にもとづく疎な 2 部グラフ上のショートサイクル数え上げ法に関する研究 電子情報通信学会情報理論研究会（IT），大阪府，2016 年 1 月 中原悠太，齋藤翔太，松嶋敏泰
15. 講演	潜在変数を仮定した多次元線形回帰モデルにおけるベイズ基準のもと最適なデータ予測に関する一考察 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 潮田幹生，齋藤翔太，松嶋敏泰
16. 講演	半教師付き学習におけるベイズ基準のもと最適な予測の計算量削減方法に関する一考察 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 中野雄斗，齋藤翔太，松嶋敏泰

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

種 類 別 By Type	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
17. 講演	一般情報源に対する Smooth 最大エントロピーを用いた可変長符号化の達成可能オーバーフローしきい値について 第 38 回情報理論とその応用シンポジウム予稿集（SITA2015），岡山県，2015 年 11 月 齋藤翔太，松嶋敏泰
18. 講演	一般情報源に対する Intrinsic randomness 問題における強逆定理のバリエーション 電子情報通信学会情報理論研究会（IT），東京都，2015 年 7 月 齋藤翔太，松嶋敏泰
19. 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
20. 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
21. 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
22. 講演	消失中継通信路上での Decode-and-Forward 型通信におけるパンクチャされた空間結合 LDPC 符号のユニバーサル性 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 中原悠太，齋藤翔太，鎌塚明，松嶋敏泰
23. 講演	非線形コンバイナ型乱数生成器に対する Sum - Product Algorithm を用いる攻撃に関する一考察 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），愛知県，2014 年 11 月 久保航汰，齋藤翔太，鎌塚明，松嶋敏泰
24. 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