論　文

ブロック単位で系列を出力する情報源に対するベイズ符号とZiv-Lempel符号のユニバーサル性について

石田　崇[†a)]　後藤　正幸[†]　平澤　茂一[††]

On Universality of Both Bayes Codes and Ziv-Lempel Codes for Sources which Emit Data Sequence by Block Unit

Takashi ISHIDA[†a)], Masayuki GOTOH[†], and Shigeichi HIRASAWA[††]

あらまし　代表的なユニバーサル符号として，Ziv-Lempel（ZL）符号とベイズ符号がある．ZL符号はその改良形アルゴリズムが実際の圧縮ソフトウェアとして広く用いられている．一方ベイズ符号は一般に計算量が多く実用化が困難であるが，FSMXモデル族[9]に対しては計算量的にも実現可能なアルゴリズムが構成されている．本論文では，テキストデータなど実際のデータの確率構造を表現できるモデルとして単語単位で系列を出力する情報源を仮定し，単語長が一定（固定）の情報源クラス（ブロック単位の情報源）に対する両符号化の漸近的な圧縮性能について解析・評価を行う．その結果，この情報源クラスに対するシンボル単位の符号化アルゴリズムについて，ベイズ符号はそのままではユニバーサルとならないが，ZL78符号についてはそのままでユニバーサルとなることを明らかにする．また，ブロック単位の情報源に対するベイズ符号化法の構成法を与える．

キーワード　ベイズ符号，Ziv-Lempel符号，ユニバーサル符号，単語単位の情報源

1. まえがき

情報源の確率構造が未知である場合の情報源符号化は，一般にユニバーサル符号と呼ばれている．その代表的な手法として，Ziv-Lempel(ZL)符号[12], [13]とベイズ符号[6], [11]があげられる．ZL符号は定常（エルゴード）情報源に対して漸近最良性が保証されており，その改良形アルゴリズムであるcompress，gzip，LHAなどは，実際の圧縮用ソフトウェアとして広く用いられている．

一方，ベイズ符号はベイズ統計理論の立場から導かれた符号である．この符号を用いたデータ圧縮は，漸近最良性に加え有限長のデータ系列に対しても圧縮性能のベイズ最適性を保証している．ベイズ符号においては，算術符号に用いるための符号化確率を求めることに主眼がおかれるが，そのために情報源の確率モデル族と事前確率が陽に仮定される点が特徴である．符号化確率とは情報源系列を符号化する際に用いられるシンボルの生起確率である．

これらのユニバーサル符号において，重要な点の一つにユニバーサル性を保証する確率分布のクラスがある．ZL符号は定常（エルゴード）情報源に対して漸近最良性が保証されている[4]．また個々の系列(individual sequence)についても各系列に対する圧縮限界である経験エントロピー(empirical entropy)までの圧縮を達成し，更に定常情報源のクラスに対しては経験エントロピーの期待値が情報源のエントロピーレートに一致することが示されている[13]．

一方，ベイズ符号では仮定した確率モデル族内で漸近最良性とベイズ最適性が保証される符号である[6]．そのため，ベイズ符号では，“いかに現実のデータの確率構造をうまく反映するモデル族を設定し，計算量の少ないアルゴリズムを構成するか”が実用化に向けての大きな鍵となる．現在では，FSMXモデル族[9]に対してベイズ符号化の効率的なアルゴリズムが構成

† 早稲田大学理工学部経営システム工学科，東京都
School of Science and Engineering, Waseda University, 3–4–1 Ohkubo Shinjyuku-ku, Tokyo, 169–8555 Japan
現在，東京大学大学院工学系研究科環境海洋工学専攻，東京都
Graduate School of Engineering, University of Tokyo, 7–3–1 Hongo Bunkyo-ku, Tokyo, 113–8656 Japan
†† 早稲田大学理工学部経営システム工学科，東京都
School of Science and Engineering, Waseda University, 3–4–1 Ohkubo Shinjyuku-ku, Tokyo, 169–8555 Japan
a) E-mail: ishida@hirasa.mgmt.waseda.ac.jp

されており [5]，計算量的にもコンピュータ上での演算が可能なものとなっている．ただし，このアルゴリズムは FSMX 情報源以外に対してはそのベイズ最適性による圧縮性能の保証はない．

しかし，実際のテキストデータ，例えば C 言語のソースプログラムのファイルなどに対してこの符号化アルゴリズムを適用し実際に圧縮した際，ZL 符号化の方がよい圧縮性能を示す事例が報告されている [1], [2]．これは，実際のデータ系列はシンボルが単語単位で出力されることを仮定できる場合が多く，シンボル単位の FSMX 情報源よりも広いクラスの情報源から生起しているためと考えられる．また，圧縮の対象となるファイルは実行ファイルであったり，特定のソフトウェアの書式で記述されたバイナリーファイルであることも多い．これらのケースは，もとのテキストデータから何らかの変換が行われた後のデータ系列を圧縮するという問題に帰着する [3], [7]．したがって，実際のデータに対する ZL 符号の特性解析やベイズ符号の実用化へ向けて，このような情報源を仮定した場合の符号化アルゴリズムの有効性を議論することが重要である．

本論文では，上述した確率モデル族の最も基本的な場合として，ブロック単位で情報を発生する情報源を仮定し，この情報源に対する ZL78 符号とベイズ符号の性能について考察する．これは, [3], [7] で定義されている単語（言語）単位で情報系列を出力する情報源において，単語長を一定（固定長）とした場合に相当し，以下ではこれをブロックと呼ぶ．本研究ではブロック単位で定常である情報源モデルを考えるが，符号器においては確率構造及びブロック長を未知とする．

単語（言語）単位で情報系列を出力する情報源は従来から解析されてきたシンボル単位で定常な情報源のクラスよりも広いクラスであることが示されており [3], [7], [8]，本研究で扱うブロック単位の定常情報源クラスもこのクラスに含まれる．ただし，従来の [3] では単語単位でエルゴードを, [7], [8] では単語単位で i.i.d. を仮定して議論している．本論文では，ブロック単位に限定しているものの，定常な情報源に議論を拡張する．この情報源はブロック単位では定常であるが，1 シンボル単位で見たときには一般に非定常となっている [3]．ここでの議論は定常情報源からの出力シンボルそれぞれを，長さが未知の固定長シンボル列に変換して得られる系列，すなわち定常情報源からの系列を FF(Fix to Fix) 符号化した符号語系列を圧縮する問題と等価である．

本論文ではブロック単位の定常情報源に対して ZL78 符号とベイズ符号の性能について解析を行い，次の 2 点を示す．

（1） シンボル単位の FSMX 情報源について構成されたベイズ符号化アルゴリズムは，ブロック単位の定常情報源に対しユニバーサル性を有さないこと．

（2） 増分分解をシンボル単位で行う（これまでの）ZL78 符号は，そのままのアルゴリズムでブロック単位の定常情報源に対してユニバーサル性を有すること．

これらの結果から，更に議論を一般化することによって出力系列の単語長が可変である場合の単語単位の情報源クラスに対しても，両符号化それぞれのユニバーサル性について上記のような結果が得られるのではないかと推測され，実際のデータ系列に対する両符号化の性能について理論的な側面から一応の解釈が可能となる．

また，シンボル単位の ZL 符号化アルゴリズムは従来から研究対象とされているシンボル単位の定常情報源クラスに対してユニバーサル性を有することが示されているが，本研究によりブロック単位では定常であるがシンボル単位では非定常となる情報源のクラスに対してもユニバーサル性を有していることが示される．すなわち，ZL 符号はシンボル単位のそのままのアルゴリズムで非定常情報源クラスにもユニバーサルとなるクラスが存在することがわかる．

更に，このような情報源モデルクラスに対して，ベイズ符号を構成する方法について述べる．このときブロック長が規定値以下ならばベイズ符号化はユニバーサル性を有する．

2. 準　　備

2.1 情報源モデル

これまで ZL 符号やベイズ符号のユニバーサル性の議論は，シンボル単位で見たときの有限マルコフ情報源や定常エルゴード情報源，定常情報源を対象にしている．

しかし，現実的に圧縮対象となるデータファイルの確率的構造を考えると，過去のシンボルから次の 1 シンボルの確率が決まると仮定するより，ブロック単位あるいは単語を単位とした情報源と仮定した方が自然である場合も多い．例えば，“He is a graduate student” という文章では，“He ”, “is”, “a”, “grad-

uate", "student" などの単語単位で文章が構成されているので，これを 1 シンボルごと "H", "e", $\cdots$ のマルコフ情報源と仮定するより，単語単位のマルコフ情報源と考えた方が自然である．この情報源は各単語をあらたに一つのシンボルと考えることで（例えば，$x_1 =$ "He", $x_2 =$"is", $\cdots$），ただのマルコフ情報源とみなせる．

本論文では，最も単純な各単語長が一定（固定長），すなわちブロック単位で系列を出力する情報源を考え，ZL 符号とベイズ符号の漸近的性能を解析する．

対象となる情報アルファベットを可算有限とし，$\mathcal{X} = \{0, 1, \cdots, J-1\}$ とする．$\mathcal{X}$ 上を値域とする確率変数を X とし，その実現値を x とする．$\mathcal{X}$ の n 次の直積を $\mathcal{X}^n$ で表す．また，情報源から出力される長さ n のシンボル系列を $x_1^n = x_1 x_2 \cdots x_n \in \mathcal{X}^n$ とし，その確率分布を $P^*(x_1^n) = \Pr\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$ と記述する．

［定義 1］（ブロック単位の定常情報源）

$\mathcal{X}$ の要素を h 個 ($h \geqq 1$) 連結したシンボル列を w で表し，本論文では以下これをブロックと呼ぶ．ただし，h は未知であるが一定の値とする（すなわち，各ブロックの長さは固定長 h となる）．この情報源はブロック w について定常情報源であるとする． □

w の集合を $\mathcal{W} = \mathcal{X}^h$ と表し，$\mathcal{W}$ 上の確率変数を W で表す．また，情報源から出力される長さ n_w のブロック単位の系列を $w_1^{n_w} = w_1 w_2 \cdots w_{n_w}$ と表し，その確率分布を $P^*(w_1^{n_w}) = \Pr\{W_1 = w_1, W_2 = w_2, \cdots, W_{n_w} = w_{n_w}\}$ と記述する．

$n = n_w h$ であるとき，情報源から出力されるブロック単位の系列 $w_1^{n_w}$ を $\mathcal{X}$ のシンボルで記述すると $w_1^{n_w} = x_1^n$ となる．また，個々のブロック w_t は，

$$w_t = x_{(t-1)h+1} x_{(t-1)h+2} \cdots x_{th}$$

と表される．例えば，$h = 2$，$x_1^6 = 011001$ とすると，$w_1 = 01$，$w_2 = 10$，$w_3 = 01$ である．

この情報源は 1 シンボル単位で見たときに一般的には非定常となっているが [3]，ブロック単位で系列を発生することに着目すれば，h シンボルからなるブロックを一つのシンボルとみなした場合は定常情報源である．

h シンボル単位（長さ h のブロック単位）の定常情報源全体からなるモデルクラスを $\mathcal{M}_h$ で表す．更に，h シンボル単位の FSMX 情報源全体からなるモデルクラスを $\mathcal{F}_h$ で表す．$\mathcal{F}_h \subset \mathcal{M}_h$ である．ここで，$\mathcal{M}_1$ は従来のシンボル単位の定常情報源のクラス，$\mathcal{F}_1$ は従来のシンボル単位の FSMX 情報源のクラスに対応している．

[例 1] $\mathcal{X} = \{0, 1\}$，情報源クラスを $\mathcal{F}_2$ に含まれる 2 シンボル単位の 2 次マルコフ情報源の例を示す．過去の二つのブロック（4 シンボル）が与えられたもとで次の 1 ブロック（2 シンボル）に対して生起確率 $P^*(w_t | w_{t-2} w_{t-1})$ が与えられる．例えば，過去の 2 ブロックが $w_{t-2} = 00$, $w_{t-1} = 10$ のもとでは，

$$P^*(00|00\ 10) = p_0^*, \quad P^*(01|00\ 10) = p_1^*,$$
$$P^*(10|00\ 10) = p_2^*, \quad P^*(11|00\ 10) = p_3^*,$$
$$\text{ただし, } p_0^* + p_1^* + p_2^* + p_3^* = 1,$$

のように確率が与えられる． □

h が未知（ただし，$h \leqq h_{\max}$ となる $h_{\max}$ が与えられている）の状況でユニバーサル符号を考えるには，$\mathcal{M}_h (h = 1, 2, \ldots, h_{\max})$ の和集合である $\mathcal{M}$ に対しての符号を構成しなければならない．すなわち，本論文で対象とするモデルクラスは $\mathcal{M} = \cup_{h=1}^{h_{\max}} \mathcal{M}_h$ である．ここで，h の最大値 $h_{\max}$ は有限であるとする．同様に $\mathcal{F} = \cup_{h=1}^{h_{\max}} \mathcal{F}_h$ で定義する．$\mathcal{F}$ は，ブロック単位の FSMX 情報源クラスを表し，$\mathcal{F} \subset \mathcal{M}$ である．

いま，

$$H(\boldsymbol{X}) = \lim_{n \to \infty} \frac{1}{n} H_n(X_1^n) \tag{1}$$

$$H(\boldsymbol{W}) = \lim_{n_w \to \infty} \frac{1}{n_w} H_{n_w}(W_1^{n_w}) \tag{2}$$

が存在するとき，$H(\boldsymbol{X})$ を情報源 P^* の 1 シンボル当りのエントロピーレート，$H(\boldsymbol{W})$ を情報源 P^* のブロック単位のエントロピーレートと呼ぶ．ここで，

$$H_n(X_1^n) = -\sum_{x_1^n \in \mathcal{X}_1^n} P^*(x_1^n) \log P^*(x_1^n) \tag{3}$$

$$H_{n_w}(W_1^{n_w}) = -\sum_{\boldsymbol{w}_1^{n_w} \in \mathcal{W}_1^{n_w}} P^*(\boldsymbol{w}_1^{n_w}) \log P^*(\boldsymbol{w}_1^{n_w}) \tag{4}$$

である．また，本論文を通じて特に断わらない限り対数の底を 2 とし，符号語アルファベットの大きさも 2 とする．

［補題 1］ $\mathcal{M}$ に含まれる情報源 P^* の 1 シンボル当りのエントロピーレート $H(\boldsymbol{X})$ は常に存在し，

$$H(\boldsymbol{X}) = \frac{1}{h} H(\boldsymbol{W}) \tag{5}$$

で与えられる．

（証明） 任意の n $(n=1,2,\cdots)$ と X_1^n に対して，シンボル系列 X_1^n はブロック系列で表現できる部分 $W_1^{n_w}$ とその前後にブロックが途中で切れている部分 X_1^a, X_{n-b+1}^n とに分解することができる．

$$\begin{aligned} X_1^n &= X_1X_2\cdots X_aW_1W_2\cdots W_{n_w}X_{n-b+1}X_{n-b+2} \\ &\quad \cdots X_n \\ &= X_1^aW_1^{n_w}X_{n-b+1}^n \end{aligned} \tag{6}$$

ただし，$0 \leqq a,b < h$，$b = n-a-hn_w$ であり，$X_1^0 = X_{n+1}^n = \phi$（空系列）と定義する．すなわち，X_1^a, X_{n-b+1}^n がそれぞれ系列の最初と最後で完全なブロックとして X_1^n に含まれないシンボルの系列を表す．

情報源アルファベット $\mathcal{X}$ は有限可算アルファベットなので，$\mathcal{W}$ も有限可算となるから，$H(W_1) < \infty$ が成り立ち $H(\boldsymbol{W})$ が常に存在する [4]．よって，ブロック単位の定常情報源に対して

$$\begin{aligned} &H(\boldsymbol{X}) \\ &= \lim_{n\to\infty}\frac{1}{n}H_n(X_1^n) \\ &= \lim_{n\to\infty}\frac{1}{n}\{H_a(X_1^a) + H_{n_w}(W_1^{n_w}|X_1^a) \\ &\quad + H_b(X_{n-b+1}^n|X_1^aW_1^{n_w})\} \\ &= \lim_{n\to\infty}\frac{1}{n}H_{n_w}(W_1^{n_w}|X_1^a) \\ &= \lim_{n_w\to\infty}\frac{1}{hn_w+a+b}H_{n_w}(W_1^{n_w}|X_1^a) \\ &= \lim_{n_w\to\infty}\frac{1}{hn_w}H_{n_w}(W_1^{n_w}|X_1^a) \end{aligned} \tag{7}$$

第2式から第3式は $a,b<h$ であることによる．$W_1^{n_w}$ は n_w に関して定常情報源であるので，

$$\begin{aligned} &\lim_{n_w\to\infty}\frac{1}{hn_w}H_{n_w}(W_1^{n_w}|X_1^a) \\ &= \frac{1}{h}\lim_{n_w\to\infty}\frac{1}{n_w}H_{n_w}(W_1^{n_w}) \\ &= \frac{1}{h}H(\boldsymbol{W}) < \infty \end{aligned} \tag{8}$$

となり証明が完結する． □

もし，P^* が W に関して k 次マルコフ情報源であるときには

$$H_n(P^*) = \frac{1}{h}H(W_1|W_{-k+1}W_{-k+2}\cdots W_0) \tag{9}$$

である．ただし，

$$\begin{aligned} &H(W_1|W_{-k+1}W_{-k+2}\cdots W_0) \\ &= -\sum_{\boldsymbol{w}_{-k+1}\cdots\boldsymbol{w}_1} P^*(w_{-k+1}w_{-k+2}\cdots w_1) \\ &\quad \cdot\log P^*(w_1|w_{-k+1}w_{-k+2}\cdots w_0) \end{aligned} \tag{10}$$

である．

2.2 ベイズ符号

情報源の確率モデル族のクラスを既知とし，パラメータは未知であるとする．逐次符号化（情報源からの出力系列 $x_1,x_2,\cdots$ の符号化が逐次的に実行される符号化法）では，情報源系列 $x_1^{t-1} = x_1x_2\cdots x_{t-1} \in \mathcal{X}^{t-1}$ が与えられた任意の時点 t において生起するシンボル $x_t \in \mathcal{X}$ を符号化する際の符号化確率 $P_c(x_t|x_1^{t-1})$ を決定するのが主問題となる．ただし，x_1^0 は空系列，$x_1^1 = x_1$ である．

現在，シンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対しては，ベイズ符号を現実的な計算量で実現するアルゴリズムが提案されている [5]．そこで，シンボル単位の FSMX 情報源のクラス $\mathcal{F}_1$ に属する情報源モデル m に対し，連続パラメータ $\theta(m)$ が定義されている場合のベイズ符号を以下に示す．すなわち，モデル m は一つのパラメトリックモデル族を表す．例えば，1次マルコフモデル，2次マルコフモデルなどを m で表し，これらは遷移確率というパラメータ $\theta(m)$ をもつパラメトリックモデル族である．

モデル m とそのパラメータ $\theta(m)$ で規定される x_1^n の確率分布を $P(x_1^n|\theta(m),m)$ とする．確率構造が未知である情報源 $P(x_1^n|\theta(m),m)$ に対し，決定関数を $AP(x_1^n)$ とする．ベイズ符号では冗長度をリスク関数 $R(m,\theta(m),AP(X_1^n))$ としたときに，モデル m とパラメータ $\theta(m)$ の事前分布 $P(m),P(\theta(m)|m)$ で平均化したベイズリスク関数 $BR(AP(X_1^n))$ を最小にする決定関数 $AP(x_1^n)$ を決定し，符号化確率とする．ただしリスク関数 $R(m,\theta(m),AP(X_1^n))$ と，ベイズリスク関数 $BR(AP(X_1^n))$ は

$$\begin{aligned} &R(m,\theta(m),AP(X_1^n)) \\ &= \sum_{x_1^n} P(x_1^n|\theta(m),m)\log\frac{P(x_1^n|\theta(m),m)}{AP(x_1^n)} \end{aligned} \tag{11}$$

$$\begin{aligned} &BR(AP(X_1^n)) \\ &= \sum_{m\in\mathcal{F}_1}\int_{\theta(m)} P(m,\theta(m))R(m,\theta(m), \\ &\quad AP(X_1^n))d\theta(m) \end{aligned} \tag{12}$$

とする．ただし，$P(m,\theta(m))=P(\theta(m)|m)P(m)$ である．

シンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対して，ベイズリスクを最小とする逐次符号の符号化確率 $AP_c(x_t|x_1^{t-1})$ は以下のように与えられる．

［補題 2］（ベイズ符号の符号化確率 [6]） 情報源クラス $\mathcal{F}_1$ に対するベイズ最適な逐次符号の符号化確率 $AP_c(x_t|x_1^{t-1})$ は

$$\begin{aligned}&AP_c(x_t|x_1^{t-1})\\&=\sum_{m\in\mathcal{F}_1}\int_{\theta(m)}P(x_t|x_1^{t-1},\theta(m),m)\\&\quad\cdot P(\theta(m)|m,x_1^{t-1})P(m|x_1^{t-1})d\theta(m)\end{aligned}\tag{13}$$

で与えられる． □

系列 x_1^n をベイズ符号で符号化したときの符号長を $L_{Bayes}(x_1^n)$ と表す．ただし，

$$\begin{aligned}L_{Bayes}(x_1^n)&=-\log AP_c(x_1^n)\\&=-\sum_{t=1}^{n}\log AP_c(x_t|x_1^{t-1})\end{aligned}\tag{14}$$

である．

2.3 ZL78 符号

ここでは Ziv-Lempel 符号の中の ZL78 符号について説明する [4], [10]．本論文を通じて $j\geqq i$ に対して x_i^j は部分系列 $x_ix_{i+1}\cdots x_j$ を表すものとする．

情報源アルファベットを $\mathcal{X}=\{0,1,\ldots,J-1\}$ とする．第 n 番目のシンボルまでの十分長いシンボル系列 $x_1^n=x_1x_2\cdots x_n$，$x_i\in\mathcal{X}$ を $p+1$ 個の空でない部分系列

$$\begin{aligned}x_1^n=x_{n(0)+1}^{n(1)}x_{n(1)+1}^{n(2)}\cdots x_{n(p)+1}^{n(p+1)}\\(n(0)=0,n(p+1)=n)\end{aligned}\tag{15}$$

に次の規則で分解する．ここで，$n(j)$ は j 番目の部分系列の末尾にあるシンボルの番号を表している．

（1） $n(1)=1$ とし，$x_{n(0)+1}^{n(1)}=x_1$ を第 1 番目の部分系列とする．

（2） 第 j 番目の部分系列 $x_{n(j-1)+1}^{n(j)}$ は最後の部分系列 $x_{n(p)+1}^{n(p+1)}$ を除きそれ以前のどの部分系列とも一致しない．すなわち，$j\leqq p$ に対し $x_{n(j-1)+1}^{n(j)}\notin\{x_{n(0)+1}^{n(1)},x_{n(1)+1}^{n(2)},\cdots,x_{n(j-2)+1}^{n(j-1)}\}$．

（3） $x_{n(j-1)+1}^{n(j)}$ はその最後 1 シンボルを除くと空系列かあるいは過去のある部分系列に一致する．すなわち

$$x_{n(j-1)+1}^{n(j)}=x_{n(r_j-1)+1}^{n(r_j)}x_{n(j)},\ x_{n(j)}\in\mathcal{X}\tag{16}$$

となる r_j が存在する．ただし，$r_j\in\{0,1,\cdots,j-1\}$ であり，$x_{n(-1)+1}^{n(0)}$ は空系列とする．

このような系列の分解を増分分解と呼ぶ．この分解アルゴリズムで得られた正整数の組 $(r_j,n(j))$ を使って，j 番目の部分系列 $x_{n(j-1)+1}^{n(j)}$ に対する符号語を

$$\varphi(x_{n(j-1)+1}^{n(j)})=r_jJ+x_{n(j)}\tag{17}$$

なる整数として定める．ただし，$x_{n(j)}\in\{0,1,\cdots,J-1\}$, $r_j<j$ より

$$0\leqq\varphi(x_{n(j-1)+1}^{n(j)})\leqq(j-1)J+(J-1)=jJ-1\tag{18}$$

である．いま，$L_j=\lceil\log_2(jJ)\rceil$ と定義する．ここで，$\lceil x\rceil$ は x 以上の最小の整数である．このとき，$\varphi(n_{n(j-1)+1}^{n(j)})$ は L_j という部分系列の番号 j だけに依存するけた数の 2 進数で表現できる．そして全体の系列 x_1^n に対する符号語は，これらの 2 進列を連接したものとする．系列 x_1^n を ZL78 符号で符号化したときの符号長を $L_{ZL}(x_1^n)$ で表す．

3. 符号化性能の解析

本章では，従来のシンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対して最適化されたベイズ符号と ZL78 符号を，前章で定義したブロック単位で出力する情報源クラス $\mathcal{M}$ に適用した場合の漸近的性能について議論する．ここで，情報源は h シンボルからなるブロック単位で系列を出力するが，符号化アルゴリズムでは 1 シンボル単位で符号化する点がポイントである．また，1 ブロックの長さ h，FSMX 情報源の確率構造は事前には未知としている．

3.1 シンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対し構成されたベイズ符号の解析

シンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対して構成されたベイズ符号を，本論文で定義したブロック単位で記号列を発生する情報源クラスに適用した場合，仮定したモデルクラスに対象モデルが含まれなければユニバーサル性を有さないことは明らかである．したがってブロック単位の記号列を発生する情報源クラスが常に $\mathcal{F}_1$ に含まれるかどうかを検証する必要があり，これについては以下の定理を導くことができる．

［定理 1］（ベイズ符号の非ユニバーサル性）

シンボル単位の FSMX 情報源 $\mathcal{F}_1$ に対して構成したベイズ符号は，ブロック単位で出力する情報源のクラス $\mathcal{M}_h$ に対し，$h \geqq 2$ のとき漸近最良性を有さない．したがって，これは情報源クラス $\mathcal{F}, \mathcal{M}$ に対してはユニバーサル符号とならない．すなわち $n \to \infty$ のとき

$$-\frac{1}{n}E^*\left[L_{Bayes}(X_1^n)\right] \to H(\boldsymbol{X}) \tag{19}$$

とならない情報源 P^* が $\mathcal{F}$ 内に存在する．ただし，E^* は P^* による期待値を表す．

（証明） 付録 1. 参照. □

この定理は，ベイズ符号を構成したクラス $\mathcal{F}_1$ よりも広いクラス $\mathcal{F}$ や $\mathcal{M}$ に対してユニバーサル性を有さないことを示しているが，ある意味で当然の結果であると言える．したがってこれは情報源クラスの設定の問題であり，ベイズ符号の適用に限界があることを示していることにはならない．

現在，効率的なアルゴリズムが提案されているベイズ符号はシンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対するものであり，もしブロック単位の情報源クラスに対して効率的なアルゴリズムを構成できればユニバーサル符号となる．その可能性については後で述べる．

3.2 ZL78 符号の解析

一方，シンボル単位の定常情報源に対しユニバーサル性が保証されている ZL78 符号が，そのままのアルゴリズムで，ブロック単位でシンボル列を発生する情報源に対してもユニバーサル性を有するならば，ZL78 符号の万能性は従来から証明されているシンボル単位の定常情報源よりも広いクラスに対して言えることになる．これに関して，情報源クラス $\mathcal{M}$ に対して，ZL78 符号化がユニバーサル性を有することを示そう．

まず，任意の $x_1^n = x_1 x_2 \cdots x_n$ に関して複雑度 $c(x_1^n)$ を以下のように定義する．

［定義 2］（系列の複雑度 [4]）

系列 $x_1^n = x_1 x_2 \cdots x_n$ を空でない相異なる部分系列に分解したときの最大個数を系列 x_1^n の複雑度 $c(x_1^n)$ とする． □

系列 x_1^n の複雑度 $c(x_1^n)$ の上界は次のように与えられることが証明されている．

［補題 3］ [4] すべての $n = 1, 2, \ldots$ とすべての x_1^n に対して，

$$c(x_1^n) < \frac{n}{(1-\varepsilon)\log_J n} \tag{20}$$

が成立する．ここで，ε は $n \to \infty$ のとき $\varepsilon \to 0$ を満たす． □

まず以下では議論を容易に進めるために，情報源クラス $\mathcal{M}$ から出力される系列がブロックが n_w だけ連なった系列の場合，すなわち

$x_1^n = x_1 x_2 \cdots x_n = w_1 w_2 \cdots w_{n_w} = w_1^{n_w}$ $(w_t = x_{(t-1)h+1} x_{(t-1)h+2} \cdots x_{th}, n = n_w h)$ の場合を考えよう．この関係を $x_1^n = w_1^{n_w}$ とも記述する．

いま，系列 x_{-kh+1}^n を固定して考える．系列 x_1^n が c 個の相異なる部分系列に

$x_1^n = x_{n(0)+1}^{n(1)} x_{n(1)+1}^{n(2)} \cdots x_{n(c-1)+1}^{n(c)}$ $(n(0) = 0, n(c) = n)$ と分解されたとする．この際，ブロック単位による並び $w_1 w_2 \cdots w_{n_w}$ の切れ目ではないところに部分系列の区切り目がくる場合がある．そのような場合を考慮し，各部分系列 $x_{n(j-1)+1}^{n(j)}$ に対して状態 $s_{n(j-1)+1}$ を以下のように定義する．ここで，$\alpha = n(j-1) \bmod h$ とする．$\alpha \in \{0, 1, \ldots, h-1\}$ であり，ブロック w の頭から数えて $\alpha+1$ 番目のシンボルが部分系列の一番最初のシンボルとなっていることを表している．この α を用いて，状態を $s_{n(j-1)+1} = x_{n(j-1)-\alpha-kh+1}^{n(j-1)}$ と定義する．これは，部分系列 $x_{n(j-1)+1}^{n(j)}$ の直前 α シンボルとブロック k 個の系列を表している．

ここで，w は k 次のマルコフ情報源から出力されるものとすると，状態 $s_{n(j-1)+1}$ により，$x_{n(j-1)}^{n(j)}$ の条件付確率は完全に規定される．また，ブロック $w_t = (x_{(t-1)h+1} x_{(t-1)h+2} \cdots x_{th})$ の途中に部分系列の区切れがある場合，ブロックの生起確率を以下のように分解する．シンボル $x_{(t-1)h+i}$ と $x_{(t-1)h+i+1}$ $(i < h)$ の間に区切れがあるとすると

$$\begin{aligned} &P^*(x_{(t-1)h+1} x_{(t-1)h+2} \cdots x_{th} | x_{(t-k-1)h+1}^{(t-1)h}) \\ &\quad = P^*(x_{(t-1)h+1} \cdots x_{(t-1)h+i} | x_{(t-k-1)h+1}^{(t-1)h}) \\ &\qquad \cdot P^*(x_{(t-1)h+i+1} \cdots x_{th} | x_{(t-k-1)h+1}^{(t-1)h+i}) \end{aligned} \tag{21}$$

とできる．ただし

$$\begin{aligned} &P^*(x_{(t-1)h+1} \cdots x_{(t-1)h+i} | x_{(t-k-1)h+1}^{(t-1)h}) \\ &\quad = \sum_{x_{(t-1)h+i+1} \cdots x_{th}} P^*(w_t | x_{(t-k-1)h+1}^{(t-1)h}) \end{aligned} \tag{22}$$

を周辺分布とし，

$$P^*(x_{(t-1)h+i+1} \cdots x_{th} | x_{(t-k-1)h+1}^{(t-1)h+i})$$

$$= \frac{P^*(x_{(t-1)h+1}\cdots x_{th}|x_{(t-k-1)h+1}^{(t-1)h})}{P^*(x_{(t-1)h+1}\cdots x_{(t-1)h+i}|x_{(t-k-1)h+1}^{(t-1)h})} \tag{23}$$

と定義する．

そして，$l=1,2,\ldots$ と，s に対して，$c_{l,s}$ を c 個の部分系列のうちで，長さが l でそれに対する状態が $s_{n(j-1)+1}=s$ であるような部分系列の個数であるとする．すると，

$$\sum_{l,s} c_{l,s} = c \tag{24}$$

$$\sum_{l,s} l\ c_{l,s} = n \tag{25}$$

が成り立っている．

ここで，次の補題が成り立つ．

［補題 4］（Ziv の不等式の一般化） $h \geqq 2$ のブロック単位で k 次マルコフ情報源である P^* から出力される $x_1^n = w_1^{n_w}$ の関係をみたす系列 $x_1x_2\cdots x_n$ について，相異なる部分系列への任意の分解は

$$\log P^*(x_1,x_2,\cdots,x_n|s_1) \leqq -\sum_{l,s} c_{l,s}\log c_{l,s} \tag{26}$$

をみたす．ただし $s_1 = x_{-kh+1}^0$ は任意である．

（証明） ブロック単位で定常な情報源はシンボル単位で見たときには一般には定常とはならないが，h シンボル（ブロックの長さ）ごとの周期をもつことから，任意の $k \in \{0,1,2,\cdots h-1\}$，$i=1,2,\cdots$ について，状態 s が同じであれば $(X_{k+1},X_{k+2},\cdots,X_{k+l})$ と $(X_{k+1+ih},X_{k+2+ih},\cdots,X_{k+l+ih})$ の状態 s のもとでの条件付確率分布は同一である．したがって，$^\forall l \geqq 1$，$^\forall k \geqq 0$ に対して

$$\sum_{x_{k+1}^{k+l}\in\mathcal{X}_1^l} P^*(x_{k+1}x_{k+2}\cdots x_{k+l}|s) = 1 \tag{27}$$

が成り立つ．したがって，部分系列 $x_{n(j-1)+1}^{n(j)}$ の条件付確率が状態 $s_{n(j-1)+1}$ によって完全に規定され，部分系列 $x_{n(j-1)+1}^{n(j)}$ がすべて異なることから

$$\sum_{\substack{j\,:\,n(j)-n(j-1)=l \\ s_{n(j-1)+1}=s}} P^*(x_{n(j-1)+1}^{n(j)}|s_{n(j-1)+1}) \leqq 1 \tag{28}$$

が成り立ち，文献 [4],p.192 補題 6.2 の証明と全く同様の議論から補題 4 が成り立つ． □

以上の用意から複雑度とエントロピーレートの関係について次のことが言える．

［補題 5］ ブロック単位での k 次マルコフ情報源 $\{W_i : i=1,2,\ldots\}$ について，情報源から出力される長さ n の系列が $X_1^n = W_1^{n_w}$ である場合，その複雑度 $c(X_1^n)$ は，

$$\limsup_{n\to\infty}\ E^*\left[\frac{c(X_1^n)\log c(X_1^n)}{n}\right] \leqq \frac{1}{h}H(W_1|W_{-k+1}W_{-k+2}\cdots W_0) \tag{29}$$

を満たす．ただし，E^* は W の真の確率分布 P^* に関する期待値を表す．

（証明） 付録 2. 参照． □

以上は $X_1^n = W_1^{n_w}$ と限定した場合であったが，以下ではこれらの結果を用いることにより，一般的に X_1^a, X_{n-b+1}^n がそれぞれ系列の最初と最後で完全なブロックとして X_1^n に含まれないような一般的な系列

$$\begin{aligned} X_1^n &= X_1X_2\cdots X_aW_1W_2\cdots W_{n_w}X_{n-b+1}X_{n-b+2} \\ &\quad \cdots X_n \\ &= X_1^aW_1^{n_w}X_{n-b+1}^n \end{aligned} \tag{30}$$

を ZL 符号化した符号長を考えよう．ただし，$0 \leqq a,b<h$，$b=n-a-hn_w$ であり，$X_1^0 = X_{n+1}^n = \phi$（空系列）と定義する．

このとき，$c(W_1^{n_w})$ を系列 $W_1^{n_w}$ をシンボル単位で増分分解したときの複雑度とすると，$W_1^{n_w}$ は X_1^n に完全に含まれる部分系列であり，$0 \leqq a,b<h$ であるので

$$c(W_1^{n_w}) \leqq c(X_1^n) \leqq c(W_1^{n_w}) + 2h \tag{31}$$

の関係が成り立つ．この関係より，任意の系列 x_1^n に対する ZL 符号化について，以下の補題が成り立つ．

［補題 6］ ブロック単位での k 次マルコフ情報源 $\{W_i : i=1,2,\ldots\}$ はエントロピーレート $\frac{1}{h}H(W_1|W_{-k+1}W_{-k+2}\cdots W_0)$ をもち，任意の長さ n の系列 $X_1^n = X_1^aW_1^{n_w}X_{n-b+1}^n$ をこの情報源から出力される n_w 個のブロックとそれらの前後にある長

さ a，b のシンボルからなる系列を表す確率変数とする．増分分解法による符号化で達成される符号語長を $L_{ZL}(x_1^n)$ とするとき，情報源出力 1 シンボル当りの平均符号長 $(1/n)E^*\left[L_{ZL}(X_1^n)\right]$ は

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(X_1^n)\right]\\&=\limsup_{n\to\infty}\frac{1}{n}E^*\left[c(X_1^n)\log c(X_1^n)\right]\\&=\limsup_{n\to\infty}\frac{1}{n}E^*\left[c(W_1^{n_w})\log c(W_1^{n_w})\right]\\&=\frac{1}{h}H(W_1|W_{-k+1}W_{-k+2}\cdots W_0)\end{aligned}\tag{32}$$

を満たす．

（証明） 付録 3. 参照． □

補題 6 において $W_1^{n_w}$ は n_w について定常で $H(W_1)<\infty$ なので，$k\to\infty$ のとき

$$\frac{1}{h}H(W_1|W_{-k+1}W_{-k+2}\cdots W_0)\to\frac{1}{h}H(\boldsymbol{W})\tag{33}$$

となる [4]．したがって，ZL78 符号は情報源クラス $\mathcal{M}$ に対してユニバーサルであることが示され，以下の定理を得る．

［定理 2］（ZL78 符号のユニバーサル性）

ZL78 符号は，ブロック単位で出力する情報源のクラス $\mathcal{M}$ に対して，漸近最良性を有するユニバーサル符号である．すなわち，$n\to\infty$ のとき

$$\frac{1}{n}E^*\left[L_{ZL}(X_1^n)\right]\to H(\boldsymbol{X})\tag{34}$$

が成り立つ． □

ZL78 符号は従来のシンボル単位のアルゴリズムそのままで，情報源 $\mathcal{M}$ に対してもユニバーサル符号となっていることが明らかとなった．情報源 $\mathcal{M}$ はシンボル単位で見たときに定常とはならないことから，今まで知られている定常情報源よりも少しではあるが広い非定常の情報源クラスに対しても ZL78 符号のユニバーサル性が示されたことになる．もし，h が既知であればブロックを 1 シンボルとみなすことによって，更に性能の良い ZL78 符号を構成することができる．すなわち，ブロック単位のアルゴリズムで符号化することによって平均符号長がより早くエントロピーレートに収束すると考えられる．

4. ブロック単位の FSMX 情報源クラス $\mathcal{F}$ に対するベイズ符号の構成法

本章ではベイズ符号について ${}^\forall h\leqq h_{\max}$ となる $h_{\max}$ が与えられたとき，単語長未知のブロック単位の FSMX 情報源クラス $\mathcal{F}$ に対して，ベイズ符号を構成する方法について述べる．

モデルクラス $\mathcal{F}_h$ のみを対象とする場合には，h シンボルからなるブロックを 1 シンボルとみなしてベイズ符号化アルゴリズム [5] を適用することによってベイズ最適性とユニバーサル性が保証される．そこで，h が未知（ただし，$h\leqq h_{\max}$ となる $h_{\max}$ は既知）の場合は h シンボルからなる単語を出力する情報源クラス $\mathcal{F}=\cup_{h=1}^{h_{\max}}\mathcal{F}_h$ に対してベイズ符号を構成すればよい．ベイズ最適な逐次符号の符号化確率 AP_c を計算するには，まずモデルクラス $\mathcal{F}_h$ ごとに次の h シンボル $(h=1,2,\cdots,h_{\max})$ の出現確率を計算し，それから各 $\mathcal{F}_h$ ごとに計算された出現確率の混合を取る必要がある．いま，h_{LCM} を $h=1,2,\cdots,h_{\max}$ の最小公倍数とすると，各 $\mathcal{F}_h$ における次の h シンボル出現確率の混合を取って符号化確率を計算するためには h_{LCM} シンボルごとでなければ計算できない．したがって，符号化確率 AP_c は

$$\begin{aligned}&AP_c\left(x_{t\cdot h_{LCM}+1}^{(t+1)\cdot h_{LCM}}|x_1^{t\cdot h_{LCM}}\right)\\&=\sum_{h=1}^{h_{\max}}\sum_{m\in\mathcal{F}_h}\int_{\theta(m)}P\left(x_{t\cdot h_{LCM}+1}^{(t+1)\cdot h_{LCM}}|x_1^{t\cdot h_{LCM}},\theta(m),m\right)\\&\quad\cdot P\left(\theta(m)|m,x_1^{t\cdot h_{LCM}}\right)P\left(m|\mathcal{F}_h,x_1^{t\cdot h_{LCM}}\right)\\&\quad\cdot P\left(\mathcal{F}_h|x_1^{t\cdot h_{LCM}}\right)d\theta(m)\end{aligned}\tag{35}$$

と計算することができる．

式 (35) によるベイズ符号は，モデルクラス $\mathcal{F}$ に対するユニバーサル符号となっている．

5. 考　　察

3. より，ブロック単位で出力されるデータ系列に対しては，ZL78 符号はシンボル単位のそのままのアルゴリズムでユニバーサルであることがわかった．つまり，ZL78 符号のアルゴリズムはシンボル単位で処理を行っているが，このアルゴリズムがユニバーサル性を保証しているシンボル単位の定常情報源クラスよりも広いクラスであるシンボル単位で非定常のクラス（ただし，ブロック単位では定常）でもユニバーサル性を保証することが言えた．一方，ベイズ符号はシンボル単位のアルゴリズムそのままではユニバーサルとならないが，**4.** で述べた方法によって FSMX 情報源クラス $\mathcal{F}$ に対してベイズ符号が構成できる．

この結果は，テキストデータのような実際のデータに対して従来のシンボル単位のベイズ符号化アルゴリズムを適用した場合，圧縮性能が ZL 符号に劣ることがあることについて一応の解釈を与えたと考える．すなわち，実際のデータでは，シンボル単位のベイズ符号化アルゴリズムがベイズ最適性を保証している FSMX 情報源のクラスよりも広いクラスから生起している場合があると考えられる．

次に，individual sequence に対する ZL78 符号の結果との比較を行う．文献 [13] では，個々の無限系列 (individual sequence) x^∞ に対して圧縮限界である経験エントロピー (empirical entropy) $\hat{H}(x^\infty)$ が定義され，ZL 符号は各系列 x^∞ に対して $\hat{H}(x^\infty)$ までの圧縮を保証している ([13]，Theorem 3)．すなわち，

$$\frac{1}{n}L_{ZL}(x_1^n) \to \hat{H}(x^\infty) \tag{36}$$

が成り立つ．更にエントロピーレート $H(\boldsymbol{X})$ をもつ定常エルゴード情報源から（確率 1 で）出力される系列に対して，文献 [13]，Theorem 4 より

$$\Pr\{\hat{H}(X^\infty) = H(\boldsymbol{X})\} = 1 \tag{37}$$

また，エントロピーレート H をもつ定常情報源から出力される系列に対しては文献 [13]，Corollary 3 より

$$E[\hat{H}(X^\infty)] = H(\boldsymbol{X}) \tag{38}$$

が示されている．

本研究の定理 2 と式 (36) より，

$$\frac{1}{n}E^*[L_{ZL}(X_1^n)] \to E^*[\hat{H}(X^\infty)] = H(\boldsymbol{X}) \quad (n \to \infty) \tag{39}$$

が言える．したがって，ブロック単位の定常情報源から出力される系列に対しても文献 [13]，Corollary 3 と同様の性質が示された．

6. むすび

本論文では系列が単語単位で出力されている情報源を考慮し，そのような情報源を表す最も基本的な確率モデルとして，ブロック（固定長の単語）を出力する情報源モデルを定義した．ただし，符号器はブロック長は未知としている．このブロック単位の定常情報源に対して，従来から提案されているシンボル単位の FSMX 情報源に対するベイズ符号化法と ZL 符号化法の符号化性能の解析を行った．

解析の結果，ブロック単位の情報源に対して

① シンボル単位の FSMX 情報源に対して構成されたベイズ符号化法ではユニバーサル性を有さないこと．

② ZL78 符号化法はシンボル単位で行うアルゴリズム（増分分解）そのままでも，ユニバーサル性を有すること．すなわち，非定常のクラスに対してもエントロピーレートが存在し，ZL78 符号がユニバーサルとなるクラスが存在すること．

が明らかとなった．

更に，ブロック長の最大値 $h_{\max}$ が既知であるとき，ブロック単位の定常情報源クラスに対してベイズ符号を構成する方法について述べた．

今後は本論文の考察をもとに，更に実データに近い情報源モデルとして，可変長の単語を出力する情報源モデル [3], [7] を単語単位で定常のモデルまで拡張し，考察を行う必要がある．シンボル単位のベイズ符号化法に関しては，最も簡単なモデルである，単語が固定長のモデルによって，ユニバーサル性を“ 有さない ”ことが言えたので，単語が可変長のモデルについてもユニバーサル性を有さないことは明らかである．ZL78 符号化法に関しては，単語が可変長のモデルに対してもユニバーサル性を“ 有している ”ことが想像されるが，今後理論的に考察すべき課題と言える．また，ZL78 符号がユニバーサル性をもつ非定常情報源のクラスの性質を明らかにすることも今後の課題である．

謝辞　筆者は，本研究を進めるに当り御討論，御助言を頂きました早稲田大学 小林学氏に深く感謝致します．また，有益な御指摘を頂きました査読者に心より感謝致します．

文　献

[1] 番場正和，松嶋敏泰，平澤茂一，“ベイズ符号のデータ圧縮に対する性能評価，” 第 19 回情報理論とその応用シンポジウム予稿集 (SITA96), pp.625–628. 1996.

[2] 番場正和，“ベイズ符号の実データに対する圧縮性能評価，” 早稲田大学大学院修士論文．1997.

[3] 後藤正幸，松嶋敏泰，平澤茂一，“単語単位で系列を出力する情報源，” 第 22 回情報理論とその応用シンポジウム予稿集 (SITA99), pp.359–362. 1999.

[4] 韓　太舜，小林欣吾，情報の符号化の数理，岩波講座応用数学，岩波書店，東京，1994.

[5] T. Matsushima and S. Hirasawa, “A Bayes coding algorithm using context tree,” Proc. Int. Symp. on Information Theory, pp.386, Trondheim, Norway, June–July 1994.

[6] T. Matsushima, H. Inazumi, and S. Hirasawa, “A class of distortionless codes designed by Bayes de-

cision theory," IEEE Trans. Inf. Theory, vol.IT-37, no.5, pp.1288–1293, Sept. 1991.
[7] 西新幹彦，森田啓義，"言語アルファベット情報源の漸近等分割性について，" 第 20 回情報理論とその応用シンポジウム予稿集 (SITA97), pp.345–348, 1997.
[8] M. Nishiara and H. Morita, "On the AEP of word-valued sources," IEEE Trans. Inf. Theory, vol.IT-46, no.3, pp.1116–1120, May 2000.
[9] J. Rissanen, "Complxity of strings in the class of Markov sources," IEEE Trans. Inf. Theory, vol.IT-32, no.4, pp.526–532, July 1986.
[10] 情報理論とその応用学会，情報源符号化-無歪みデータ圧縮，情報理論とその応用学会シリーズ 1-I，培風館，東京，1998.
[11] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens, "The context-tree weighting method : Basic properties," IEEE Trans. Inf. Theory, vol.IT-41, no.3, pp.653–664, May 1995.
[12] J. Ziv and A. Lempel, "A universal algorithm for data compression," IEEE Trans. Inf. Theory, vol.IT-23, no.3, pp.337–343, May 1977.
[13] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," IEEE Trans. Inf. Theory, vol.IT-24, no.5, pp.530–536, Sept. 1978.

付　　録

1. 定理 1 の証明

ここでは，最も簡単な場合として，$\mathcal{X}=\{0,1\}, h=2, k=0$ で与えられる情報源（固定長 2 の単語を i.i.d. で出力する情報源）を考える．この情報源から出力されるデータ系列に対して，シンボル単位の FSMX 情報源 $\mathcal{F}_1$ について構成されたベイズ符号化アルゴリズム（従来法）を適用した際に，漸近最良性を有さないことを示す．これにより，ブロック単位で生起する情報源のクラス $\mathcal{F}, \mathcal{M}$ の中に圧縮限界まで到達できない情報源の存在が示されるので，情報源のクラス $\mathcal{F}, \mathcal{M}$ に対して一般にユニバーサル性を有さないことが示される．

いま，真の情報源のパラメータ $P^*=(p_0^*, p_1^*, p_2^*, p_3^*)$ を

$$\begin{aligned} p_i^* &= \Pr\{W=w_i\} \\ &= P^*(w_i) \end{aligned} \tag{A·1}$$

と表すことにする．すなわち，

$$P^*(00)=p_0^*, \quad P^*(01)=p_1^*,$$
$$P^*(10)=p_2^*, \quad P^*(11)=p_3^*$$

である．

アルゴリズムでは任意の最大深さ d のシンボル単位の FSMX 情報源 $\mathcal{F}_1$（過去の d シンボルの記憶をもつ情報源）のクラスを仮定し，ベイズ符号で用いるモデルのパラメータを $\boldsymbol{\theta}=(\theta_{s_0}, \theta_{s_1}, \cdots, \theta_{s_{2^d-1}})$ と表す．ここで，$s=x_1^d \in \mathcal{X}^d$ は，過去の系列 X_{-d+1}^0 によって決まる状態を表している．また，

$$\theta_s = P_\theta(0|s)$$

とする．

もし $\boldsymbol{\theta}$ が与えられれば，系列 x_{-d+1}^0 が出現したもとで（状態 s のもとで）次に生起するブロック $(w_1=x_1x_2)$ に対して，以下の理想符号長 $L(w_1|s)$ で Shannon 符号化を行うことができる．

$$\begin{aligned} L_\theta(w_1|s) &= -\log P_\theta(w_1|s) \\ &= -\log P_\theta(x_1x_2|x_{-d+1}^0) \\ &= -\log P_\theta(x_1|x_{-d+1}^0) P_\theta(x_2|x_{-d+2}^1) \end{aligned} \tag{A·2}$$

すると 1 ブロック当りの平均符号長 $\overline{L(\boldsymbol{\theta})}$ は以下のように求まる．

$$\begin{aligned} \overline{L(\boldsymbol{\theta})} &= E^*\left[L_\theta(w_1|s)\right] \\ &= \sum_s Q^*(s) \sum_{\boldsymbol{w}_1} P^*(w_1)\cdot L(w_1|s) \\ &= \sum_{x_{-d+1}^0} P^*(x_{-d+1}^0) \sum_{x_1^2} P^*(x_1^2) \\ &\quad \cdot \Big(-\log P_\theta(x_1|x_{-d+1}^0) P_\theta(x_2|x_{-d+2}^1)\Big) \end{aligned} \tag{A·3}$$

ただし，E^* は $P^*(\cdot)$ による期待値を表し，$Q^*(s)$ は状態 s の定常分布を表す．いま，情報源は i.i.d. なので，$Q^*(s)=P^*(x_{-d+1}^0)$ である．

いま，$E^*[L_\theta(w_1|s)]$ の下限を与える最適パラメータを $\tilde{\boldsymbol{\theta}}$ とする．すなわち，ベイズ符号の 1 ブロック当りの平均符号長は

$$\overline{L(\tilde{\boldsymbol{\theta}})} = H(P^*) + D(P^* \| P_{\tilde{\theta}}) \tag{A·4}$$

で下界される．$H(P^*)$ は情報源 P^* の 1 単語当りのエントロピー，$D(P^*\|P_{\tilde{\theta}})$ はダイバージェンスを表し，

以下の式で与えられる．

$$H(P^*) = -\sum_{x_1^2} P^*(x_1^2)\log P^*(x_1^2) \tag{A·5}$$

$$D(P^*\|P_{\tilde{\theta}}) = -\sum_{x_{-d+1}^0} P^*(x_{-d+1}^0)\sum_{x_1^2} P^*(x_1^2) \cdot \log\frac{P^*(x_1^2)}{P_{\tilde{\theta}}(x_1|x_{-d+1}^0)P_{\tilde{\theta}}(x_2|x_{-d+2}^1)} \tag{A·6}$$

$D(P^*\|P_{\tilde{\theta}}) \geqq 0$ で，等号は任意の $w_1 = x_1^2, x_{-d+1}^0$ に対し，$P^*(w_1) = P_{\tilde{\theta}}(x_1|x_{-d+1}^0)P_{\tilde{\theta}}(x_2|x_{-d+2}^1)$ が成立するときに限られる．

いま，$w_1 = 00$ とするとダイバージェンスが 0 となる条件は，任意の $w_1 = x_1^2, x_{-d+1}^0$ に対して

$$P^*(00) = P_{\tilde{\theta}}(0|x_{-d+1}^0)\cdot P_{\tilde{\theta}}(0|x_{-d+2}^0\ 0) \tag{A·7}$$

が成り立つときで，このときに限られる．式 (A·7) において，x_{-d+1} が 0 と 1 の場合で

$$\begin{aligned} P^*(00) &= P_{\tilde{\theta}}(0|0\ x_{-d+2}^0)\cdot P_{\tilde{\theta}}(0|x_{-d+2}^0\ 0) \\ &= P_{\tilde{\theta}}(0|1\ x_{-d+2}^0)\cdot P_{\tilde{\theta}}(0|x_{-d+2}^0\ 0) \end{aligned} \tag{A·8}$$

となり，式 (A·8) より，任意の x_{-d+2}^0 について

$$P_{\tilde{\theta}}(0|0\ x_{-d+2}^0) = P_{\tilde{\theta}}(0|1\ x_{-d+2}^0) \tag{A·9}$$

が言え，これを $P_{\tilde{\theta}}(0|x_{-d+2}^0)$ と表すことにすると式 (A·7) は

$$P^*(00) = P_{\tilde{\theta}}(0|x_{-d+2}^0)\cdot P_{\tilde{\theta}}(0|x_{-d+3}^0\ 0) \tag{A·10}$$

となる．

次に x_{-d+2} が 0 と 1 の場合で同様の議論を行うことにより，

$$P_{\tilde{\theta}}(0|0\ x_{-d+3}^0) = P_{\tilde{\theta}}(0|1\ x_{-d+3}^0) \tag{A·11}$$

が言える．

この議論を繰り返すと最終的に

$$P_{\tilde{\theta}}(0|0) = P_{\tilde{\theta}}(0|1) \tag{A·12}$$

となる．以上から，ダイバージェンスが 0 となる条件はパラメータ $P_{\tilde{\theta}}(0|s)$ が状態 s に依存せず一定の値をとる場合で，

$$\bar{\theta} \stackrel{\text{def}}{=} \tilde{\theta}_{s_0} = \tilde{\theta}_{s_1} = \cdots = \tilde{\theta}_{s_{2^d-1}} \tag{A·13}$$

となるときに限られる．このとき，$w = 01, 10, 11$ についても任意の x_{-d+1}^0 に対して，$P^*(w_1) = P_{\tilde{\theta}}(x_1|x_{-d+1}^0)P_{\tilde{\theta}}(x_2|x_{-d+2}^1)$ が成立することがわかる．

すなわち，P^* と $P_{\tilde{\theta}}$ が．

$$p_0^* = \bar{\theta}^2, \quad p_1^* = p_2^* = \bar{\theta}(1-\bar{\theta}), \quad p_3^* = (1-\bar{\theta})^2 \tag{A·14}$$

のような関係を満たすときに限りダイバージェンスは 0 となるが，それ以外の場合には正の値をとる．

以上より，このような情報源のクラスに対して冗長度が正となる情報源モデルが存在することが言えたので，シンボル単位の FSMX 情報源クラス $\mathcal{F}_1$ に対して構成されたベイズ符号が情報源クラス $\mathcal{F}$, $\mathcal{M}$ に対しユニバーサルでないことが証明された．

2. 補題 5 の証明

証明の流れは文献 [4],p.192 定理 6.1 の証明とほぼ同様の流れとなるので簡単に示す．

補題 3 及び補題 4 から

$$\frac{c(x_1^n)\log c(x_1^n)}{n} \leqq -\frac{1}{n}\log P^*(x_1^n|s_1) + \delta(n) \tag{A·15}$$

が言える．ただし，$\delta(n) \to 0\ \ (n \to \infty)$ である．

ここで，$s_1 = w_{-k+1}^0$ と $n = hn_w$ であることに注意し，$x_{-kh+1}^n = w_{-k+1}^{n_w}$ は，情報源 $\mathcal{M}$ からの出力系列である確率変数ベクトル $W_{-k+1}^{n_w}$ の実現値であるから式 (A·15) は

$$\begin{aligned} &\frac{c(X_1^n)\log c(X_1^n)}{n} \\ &\quad \leqq -\frac{1}{hn_w}\log P^*(W_1^{n_w}|W_{-k+1}^0) + \delta(n) \end{aligned} \tag{A·16}$$

と書ける．両辺の期待値をとって，

$$\begin{aligned} &\frac{1}{n}E^*\left[c(X_1^n)\log c(X_1^n)\right] \\ &\quad \leqq \frac{1}{h}E^*\left[-\log P^*(W_1|W_{-k+1}^0)\right] + \delta(n) \end{aligned} \tag{A·17}$$

を得る．ところで，

$$E^*\left[-\log P^*(W_1|W_{-k+1}^0)\right] = H(W_1|W_{-k+1}^0) \tag{A·18}$$

が成立するので，

$$\begin{aligned} &\frac{1}{n}E^*\left[c(X_1^n)\log c(X_1^n)\right] \\ &\quad \leqq \frac{1}{h}H(W_1|W_{-k+1}^0) + \delta(n) \end{aligned} \tag{A·19}$$

を得る．よって，

$$\limsup_{n\to\infty}\frac{1}{n}E^*\left[c(X_1^n)\log c(X_1^n)\right] \leqq \frac{1}{h}H(W_1|W_{-k+1}^0) \quad \text{(A·20)}$$

が帰結する． □

3. 補題 6 の証明

証明の流れは文献 [4],p.195 定理 6.2 の証明とほぼ同様の流れであるが，$c(X_1^n)$ と $c(W_1^{n_w})$ の取扱いに注意する．補題 5 と ZL78 符号の符号長の関係から，$X_1^n = W_1^{n_w}$ $(n = hn_w)$ の関係を保ったまま $n \to \infty$ を考えると，文献 [4] の議論より

$$\limsup_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(W_1^{n_w})\right] = \limsup_{n\to\infty}\frac{1}{n}E^*\left[c(W_1^{n_w})\log c(W_1^{n_w}))\right] \leqq \frac{1}{h}H(W_1|W_{-k+1}^0) \quad \text{(A·21)}$$

が成立することがわかる．

一方，$n = hn_w + a + b$ $(a, b < h)$ をみたす一般の長さの系列 x^n については，

$$c(W_1^{n_w}) \leqq c(X_1^n) \leqq c(W_1^{n_w}) + 2h \quad \text{(A·22)}$$

の関係と $x\log x$ が単調増加関数であることから，

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(X_1^n)\right] &\leqq \limsup_{n\to\infty}\frac{1}{n}E^*\left[c(X_1^n)\log c(X_1^n))\right] \\ &\leqq \limsup_{n\to\infty}\frac{1}{n}E^*\left[\{c(W_1^{n_w})+2h\}\log\{c(W_1^{n_w})+2h\}\right] \\ &\leqq \limsup_{n\to\infty}\frac{1}{n}E^*\left[c(W_1^{n_w})\log c(W_1^{n_w})+C\right] \\ &\leqq \frac{1}{h}H(W_1|W_{-k+1}^0) \quad \text{(A·23)}\end{aligned}$$

が導かれる．ただし，C は適当な有界の定数である．

一方，増分分解に基づく符号は語頭条件を満たすので Kraft の不等式を満たす．よって，

$$\frac{1}{n}E^*\left[L_{ZL}(W_1^{n_w})\right] \geqq \frac{1}{n}H(W_1^{n_w}|W_{-k+1}^0) \quad \text{(A·24)}$$

が成立しなければならない．ここで，$n_w \to \infty$ のとき

$$\frac{1}{n_w}H(W_1^{n_w}|W_{-k+1}^0) \to H(W_1|W_{-k+1}^0) \quad \text{(A·25)}$$

であるから，$n = n_w h + a + b$ $(a, b < h)$ より

$$\liminf_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(W_1^{n_w})\right] \geqq \frac{1}{h}H(W_1|W_{-k+1}^0) \quad \text{(A·26)}$$

$c(W_1^{n_w}) \leqq c(X_1^n)$ より，

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(X_1^n)\right] &\geqq \liminf_{n\to\infty}\frac{1}{n}E^*\left[L_{ZL}(W_1^{n_w})\right] \\ &\geqq \frac{1}{h}H(W_1|W_{-k+1}^0) \quad \text{(A·27)}\end{aligned}$$

結局，式 (A·23)，(A·27) より，式 (32) を得る． □

（平成 12 年 3 月 29 日受付，12 月 25 日再受付）

石田　崇

平 11 早大・理工・経営システム工学卒．平 13 同大大学院修士課程了．現在，東京大学大学院工学系研究科博士課程在学中．情報源符号化，統計的学習理論，統計的モデル選択，ベイズ統計応用などに興味をもつ．

後藤　正幸　（正員）

平 4 武蔵工大・工・経営工学卒．平 6 同大大学院修士課程了．平 6，早大・理工学研究科・博士後期課程入学．平 8-11，同大・理工・経営システム工学科助手．早稲田大学，東京都立短期大学，武蔵工業大学の非常勤講師などを経て，現在，東京大学大学院工学系研究科環境海洋工学専攻助手．博士（工学）．情報源符号化，統計的学習理論，統計的モデル選択，ベイズ統計応用などの研究に従事．IEEE, 情報理論とその応用学会，人工知能学会，日本経営工学会各会員．

平澤　茂一　（正員）

昭 36 早大・理工・数学卒．昭 38 同大電気通信卒．同年三菱電機入社．昭 56 早大・理工・工業経営学科（現在経営システム工学科）教授，現在に至る．情報理論とその応用，データ伝送方式，並びに計算機応用システムの開発などの研究に従事．工博．昭 54UCLA 計算機科学科客員研究員．昭 60 ハンガリー科学アカデミー，昭 61 伊トリエステ大学客員研究員．平 5 電子情報通信学会 小林記念特別賞，業績賞受賞．平 8 情報理論とその応用学会会長．IEEE, 情報理論とその応用学会，人工知能学会，情報処理学会，OR 学会，日本経営工学会等各会員．