論　文

副情報源を伴う情報源の増分分解に基づくユニバーサル符号化法の漸近最良性

植松　友彦[†]　　前田　浩次[†]

Asymptotical Optimality for Universal Data Compression Algorithm with Side Information Based on Incremental Parsing

Tomohiko UYEMATSU[†] and Koji MAEDA[†]

あらまし　副情報源を伴う情報源の符号化問題とは、情報源と相関を有する情報源を副情報源として符号器および復号器で参照しながら符号化するという問題である。この符号化問題について、Slepian と Wolf は、情報源1文字あたりの符号長の限界が条件付きエントロピーレートによって与えられていることを示した。他方、Ziv は、現在多くのデータ圧縮法で利用されている増分分解法を利用することで、副情報源を伴う情報源の具体的かつ実用的なユニバーサル符号化法を提案した。しかしながら、Ziv の提案した符号化法の漸近最良性は未知であった。本論文では、Ziv によるユニバーサル符号化法が定常エルゴード情報源に対して漸近最良性を有していること、すなわち入力系列が長くなると1文字あたりの符号長が条件付きエントロピーレートに確率1で収束することを明らかにしている。

キーワード　情報源符号化, 副情報源, 増分分解法, ユニバーサル符号化, 定常エルゴード情報源

1. まえがき

副情報源を利用した情報源の符号化とは、主情報源系列 x と、それに伴う副情報源系列 y に対して、符号器、復号器がともに副情報源を参照しながら入力系列の符号化・復号化を行うことである (図1)。ここで、復号化によって得られる系列は符号化された入力系列と完全に一致すると仮定する。この問題は、Slepian と Wolf[1] によって始められた相関を有する情報源の符号化問題の特殊な場合であり、平均符号長の限界は条件付きエントロピーレート $H_{X|Y}$ で与えられることが知られている。符号器と復号器が共に副情報源を参照できる場合の符号化法としては、Ziv の固定長符号化法[2] と増分分解[3] を利用した可変長符号化法[4]、村松の符号化法[5]、Subrahmanya と Berger の LZ77 符号[6] に基づいた符号化法[7]、Yang らによる文法に基づく符号化と多重パターンマッチング (Multiple Pattern Matching) を組み合わせた符号化法[8] などが知られている。特に、Ziv の固定長符号は独立系列に対して、村松と Yang らの符号化法は定常エルゴード情報源に対して、それぞれ漸近最良性が証明されている。しかしながら、これらの符号化法の中で実用的かつ理論的に興味深い、2種の Lempel-Ziv 符号を用いた符号化法の漸近的最良性はまだ示されていない。

小文では、符号器と復号器が共に副情報源を参照できる場合について、情報源の統計的性質に依存しないユニバーサル符号化法として、増分分解を利用した Ziv の符号化法を取り上げる。Ziv の提案した符号化法は、有限状態通信路に対するユニバーサルな復号法を提案する際の副産物であったため、符号化法の漸近的最良性については明らかにされていない。そこで、この Ziv による符号化法が定常エルゴード情報源に対して漸近的に最良であること、すなわち、入力系列が長くなると、1文字あたりの符号長が条件付きエントロピーレートに確率1で収束することを示す。

2. 符号化アルゴリズムとその漸近的最良性

Ziv による具体的な符号化アルゴリズムを述べる前に増分分解[3] について説明する。以下では、主情報源のアルファベットを $\mathcal{X}$、副情報源のアルファベットを $\mathcal{Y}$ とし、共に有限であると仮定する。長さ n の主

[†] 東京工業大学 大学院 理工学研究科 集積システム専攻, 東京都
Dept. of Communications and Information Systems, Tokyo Institute of Technology, Tokyo, 152-8552 Japan

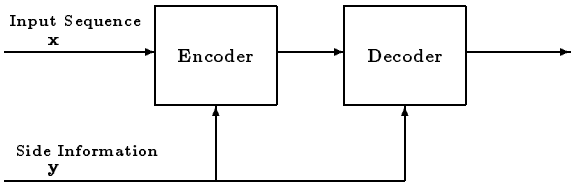

図1　副情報源を伴う情報源の符号化
Fig. 1　Source coding with side information.

情報源系列を $x_1^n = x_1 x_2 \dots x_n (\in \mathcal{X}^n)$ によって、対応する副情報源系列を $y_1^n = y_1 y_2 \dots y_n (\in \mathcal{Y}^n)$ によってそれぞれ表す。主情報源系列と副情報源系列を併せて得られる同時系列を $(xy)_1^n \in (\mathcal{X} \times \mathcal{Y})^n$ によって表す。また、部分列 $x_i x_{i+1} \dots x_j$ 等を記号 x_i^j によって表す。

同時系列 $(xy)_1^n \in (\mathcal{X} \times \mathcal{Y})^n$ の増分分解

$$(xy)_1^n = (xy)_{n_0+1}^{n_1} (xy)_{n_1+1}^{n_2} \dots (xy)_{n_{p-1}+1}^{n_p}$$

とは、次の条件を満足する空でない系列への分解である。但し、 $n_0 = 0$、$n_p = n$ とする。

（1） 最後の系列 $(xy)_{n_{p-1}+1}^{n_p}$ を除いた、$p-1$ 個の部分列 $(xy)_{n_{i-1}+1}^{n_i} (1 \leqq i < p)$ は全て相異なる。

（2） $n_i - n_{i-1} \geqq 2$ を満足する全ての $i = 1, 2, \dots, p$ に対して、$(xy)_{n_{j-1}+1}^{n_j} = (xy)_{n_{i-1}+1}^{n_i - 1}$ となる正整数 $j (< i)$ が存在する。

次に、Ziv による符号化アルゴリズム [4] について説明する。

符号化アルゴリズム　(Z アルゴリズム)

1. 同時系列 $(xy)_1^n \in (\mathcal{X} \times \mathcal{Y})^n$ を増分分解し、分解後の系列を $(xy)_1^n = (xy)_{n_0+1}^{n_1} (xy)_{n_1+1}^{n_2} \dots (xy)_{n_{p-1}+1}^{n_p}$ とする。但し、$n_0 = 0$、$n_p = n$ である。ここで $i = 0$ とおく。

2. これから符号化する部分列の長さ $(n_{i+1} - n_i)$ を Elias の ω^* 符号 [9] で2元表示する。

3. $(xy)_1^n$ の部分列 $(xy)_{n_i+1}^{n_{i+1}}$ の最後の1文字を除いた系列 $(xy)_{n_i+1}^{n_{i+1}-1}$ と一致する系列を $(xy)_{n_j+1}^{n_{j+1}} (1 \leqq j \leqq i-1)$ から探す (増分分解の性質上、一致する系列が必ず存在する)。

4. $y_{n_i+1}^{n_{i+1}-1}$ と一致する系列を $y_{n_j+1}^{n_{j+1}} (1 \leqq j \leqq i-1)$ から探し、その総数を q とする。そのとき、手順 3. で一致した系列が、これらの系列のうちの前から何番目に出現しているのかを、$\lceil \log q \rceil$ ビットで2元表示する。但し、$\lceil x \rceil$ は x 以上の最小整数を表す記号である。

5. 主情報源の最後の1文字 $x_{n_{i+1}}$ を $\lceil \log |\mathcal{X}| \rceil$ ビットで表示する。

6. $i < p-1$ である場合、$i = i+1$ として 2. に戻る。$i = p-1$ である場合、符号化を終了する。　□

この符号化アルゴリズムは、情報源の統計的性質を一切利用しておらず、ユニバーサルな符号である。

次に復号アルゴリズムを示す。これにより上記の符号が瞬時符号であることがわかる。

復号アルゴリズム

1. $n_0 = 0$、$i = 0$ とする。

2. Elias の ω^* 符号の復号アルゴリズムによって、これから復号する系列の長さ L を得る。$n_{i+1} = n_i + L$ とする。

3. 副情報源系列 $y_{n_i+1}^{n_{i+1}-1}$ を取り出し、$y_{n_i+1}^{n_{i+1}-1}$ がそれ以前に復号した部分列 $y_{n_j+1}^{n_{j+1}} (1 \leqq j \leqq i-1)$ と一致する回数 q を求める。次に、符号語列から $\lceil \log q \rceil$ ビットを取り出し、復号する系列が同一の副情報源系列を持つ中で前から何番目に出現していたのかという情報を得た後、主情報源系列 $x_{n_i+1}^{n_{i+1}-1}$ を復号する。

4. 符号語列から $\lceil \log |\mathcal{X}| \rceil$ ビットを取り出し、復号することにより、主情報源部分列の最後の1文字を得る。

5. 符号語列がまだある場合は、$i = i+1$ として 2. に戻る。無い場合は復号を終了する。　□

例 1: 上で述べた符号化の例を挙げる。

情報源アルファベットを $\mathcal{X}, \mathcal{Y} = \{0, 1\}$ とし、主情報源系列 x_1^{18} とそれに対応する副情報源系列 y_1^{18} を

$$x_1^{18} = 001010101000000001$$
$$y_1^{18} = 100010110101011011$$

とする。この同時系列を増分分解すると、

$$x_1^{18} = 0 \;\; 0 \;\; 1 \;\; 01 \;\; 010 \;\; 10 \;\; 00 \;\; 000 \;\; 001$$
$$y_1^{18} = 1 \;\; 0 \;\; 0 \;\; 01 \;\; 011 \;\; 01 \;\; 01 \;\; 011 \;\; 011$$

が得られる。ここでは、最後の部分列 (001) を符号化することを考える。このとき、

(1) 系列長は 3 なので、まず 3 を ω^* 符号で符号化して 10 を得る。

(2) 最後の1文字を除いた副情報源系列は 01 となる。

(3) 符号化を行なっている部分列以前の副情報源部分列の中に 01 は 3 回現れているので、$\lceil \log 3 \rceil = 2$ ビットで主情報源一致位置を表す。

(4) 主情報源一致位置は 3 番目であるので、$3-1$ の 2 元表示 10 を得る。

(5) 最後の 1 文字 1 を 1 ビットで符号化して 1 を得る。

最終的な符号語はこれらを連接した 10101 になる。 □

以下では長さ n の系列を Z アルゴリズムで符号化することを写像 $\varphi_n : (\mathcal{X}\times\mathcal{Y})^n \to \{0,1\}^*$ によって表す。次の定理は、この符号が漸近的に最良なユニバーサル符号であることを表している。但し、$l(\cdot)$ は 2 元系列の長さを表す関数とする。

［定理 1］ 定常エルゴード情報源 (X,Y) からの出力系列 (x,y) に対して、Z アルゴリズムの符号長は

$$\lim_{n\to\infty}\frac{l(\varphi_n((x,y)_1^n))}{n} = H_{X|Y} \qquad a.s. \qquad (1)$$

を満足する。但し、$H_{X|Y}$ は情報源 (X,Y) の条件付きエントロピーレートであり、

$$H_{X|Y} = \lim_{n\to\infty}\frac{1}{n}H(X_1^n|Y_1^n)$$

によって定義される。 □

3. 証　　明

ここでは Ornstein と Weiss による成果 [10] を利用して定理を証明する。尚、別証明が文献 [11] に述べられている。

3.1 定義と補題

この節では定理の証明に用いる幾つかの定義と補題を述べる。

［定義 1］ 証明に用いる記号を以下のように定義する。

- $c((xy)_1^n)$: 同時系列 $(xy)_1^n$ を増分分解したときに得られる部分列の個数
- $c(y_1^n)$: 部分列 $y_{n_{i-1}+1}^{n_i}(1\leqq i\leqq c((xy)_1^n))$ の中で相異なるものの個数
- $y(l)$: 増分分解後の部分列 $y_{n_{i-1}+1}^{n_i}(1\leqq i\leqq c((xy)_1^n))$ の中で、第 l 番目 $(1\leqq l\leqq c(y_1^n))$ に初めて出現したもの
- $c_l(x_1^n|y_1^n)$: $y_{n_{i-1}+1}^{n_i} = y(l)$ であるような $x_{n_{i-1}+1}^{n_i}(1\leqq i\leqq c(y_1^n))$ の個数

□

例 2: $\mathcal{X}=\mathcal{Y}=\{0,1\}$ とし、

$$x_1^{19} = 0110000100101001010$$
$$y_1^{19} = 1010100111111010101$$

とするとき、同時系列 $(xy)_1^{19}$ を増分分解すると、

$$x_1^{19} = 0\ 1\ 1\ 0\ 00\ 01\ 00\ 10\ 10\ 01\ 010$$
$$y_1^{19} = 1\ 0\ 1\ 0\ 10\ 01\ 11\ 11\ 10\ 10\ 101$$

となり、この場合、

$c((xy)_1^{19}) = 11$　$c(y_1^{19}) = 6$

$y(1)=1$　$y(2)=0$　$y(3)=10$

$y(4)=01$　$y(5)=11$　$y(6)=101$

$c_1(x_1^{19}|y_1^{19})=2$　$c_2(x_1^{19}|y_1^{19})=2$

$c_3(x_1^{19}|y_1^{19})=3$　$c_4(x_1^{19}|y_1^{19})=1$

$c_5(x_1^{19}|y_1^{19})=2$　$c_6(x_1^{19}|y_1^{19})=1$

となる。 □

［定義 2］ (強被覆 [12]) $\mathcal{X}$ を有限アルファベットとする。$\mathcal{X}^m$ の部分集合 C^m、長さ $n(>m)$ の系列 $x_1^n\in\mathcal{X}^n$ と $\delta>0$ が与えられたとき、$x_i^{i+m-1}\notin C^m$ $(i=1,2,\ldots,n-m+1)$ であるインデックス i の個数が δn 個以下であるとき、すなわち、

$$\left|\left\{i\in\{1,2,\ldots,n-m+1\} : x_i^{i+m-1}\notin C^m\right\}\right| \leqq \delta n$$

であるとき、x_1^n は C^m によって $(1-\delta)$ 強被覆される ($(1-\delta)$-strongly-covered by C^m) という。 □

［補題 1］ (Ornstein and Weiss [10])

同時エントロピー H_{XY} を持つ定常エルゴード情報源 (X,Y) と、任意の $\epsilon>0$ が与えられたとする。同時系列 $xy\in(\mathcal{X}\times\mathcal{Y})^\infty$ に対して、$(xy)_1^n = (xy)_{\tilde{n}_0+1}^{\tilde{n}_1}(xy)_{\tilde{n}_1+1}^{\tilde{n}_2}\cdots(xy)_{\tilde{n}_{p-1}+1}^{\tilde{n}_p}$ を互いに異なる語への分割とするとき、ある $N(\epsilon,xy)$ が存在して、$n\geqq N(\epsilon,xy)$ ならば、

$$\sum_{i:\tilde{n}_i-\tilde{n}_{i-1}<\frac{\log n}{H_{XY}+\epsilon}}(\tilde{n}_i-\tilde{n}_{i-1}) \leqq \epsilon n$$

が確率 1 で成り立つ。 □

［補題 2］ $(xy)_1^n\in(\mathcal{X}\times\mathcal{Y})^n$ を Z アルゴリズムによって符号化したときの符号長は、

$$l(\varphi_n((xy)_1^n)) \leqq \sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)[\log c_i(x_1^n|y_1^n) + \log|\mathcal{X}| + \log\ell(y(i)) + 2\log\log\ell(y(i)) + 9] \quad (2)$$

を満足する。但し、$\ell(\cdot)$ は $\mathcal{Y}$ 上の系列長を表す関数である。 □

[補題 2 の証明]　符号化アルゴリズムの手順 2. で用いた Elias の ω^* 符号の整数 i に対する符号の符号長は、

$$\log i + 2\log\log i + 7$$

以下である [9]。従って、手順 2. の部分の符号長の合計 $l_0((xy)_1^n)$ は、

$$\begin{aligned} l_0((xy)_1^n) \leqq & \sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)\{\log \ell(y(i)) \\ & +2\log\log \ell(y(i)) + 7\} \end{aligned}$$

によって上から抑えられる。

次に、手順 4. の部分の符号長の合計 $l_1((xy)_1^n)$ は、

$$\begin{aligned} l_1((xy)_1^n) &= \sum_{i=1}^{c(y_1^n)} \sum_{q=1}^{c_i(x_1^n|y_1^n)} c_i(x_1^n|y_1^n)\lceil \log q \rceil \\ &\leqq \sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)\{\log c_i(x_1^n|y_1^n) + 1\} \end{aligned}$$

によって上から抑えられる。最後に手順 5. の部分の符号長の合計 $l_2((xy)_1^n)$ は $c((xy)_1^n)\lceil \log|\mathcal{X}| \rceil$ ビットとなる。

ここで、

$$\sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n) = c((xy)_1^n)$$

が成り立つことに着目すれば、

$$\begin{aligned} & l(\varphi_n((xy)_1^n)) \\ &= l_0((xy)_1^n) + l_1((xy)_1^n) + l_2((xy)_1^n) \\ &\leqq \sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)[\log c_i(x_1^n|y_1^n) + \log|\mathcal{X}| \\ &\qquad + \log \ell(y(i)) + 2\log\log \ell(y(i)) + 9] \end{aligned}$$

となり、題意は示された。　(証明終)

［補題 3］ (Ziv [4])

任意の同時系列 $(xy)_1^n \in (\mathcal{X}\times\mathcal{Y})^n$ に対し、

$$\frac{1}{n}\sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)\log \ell(y(i)) \leqq O\left(\frac{\log\log n}{\log n}\right)$$

が成り立つ。　□

以上の事実を用いて、次節では定理 1 の証明を行なう。

3. 2　定理の証明

まず、式 (1) の片方の不等式

$$\liminf_{n\to\infty} \frac{l(\varphi_n((x,y)_1^n))}{n} \geq H_{X|Y} \quad a.\,s.$$

は次の補題から直ちに示される。

［補題 4］ (Muramatsu [5])

$\mathcal{X}\times\mathcal{Y}$ 上の定常エルゴード情報源 (X,Y) に対する副情報源を伴う任意の 2 元無損失符号の系列 $\{(\psi_n, \psi_n^{-1})\}$ は、

$$\liminf_{n\to\infty} \frac{1}{n} l(\psi_n((x,y)_1^n)) \geq H_{X|Y} \quad a.\,s.$$

を満足する。但し、

$$\psi_n : (\mathcal{X}\times\mathcal{Y})^n \to \{0,1\}^*,$$

$$\psi_n^{-1} : \{0,1\}^* \times \mathcal{Y}^n \to \mathcal{X}^n$$

である。　□

読者の便宜のため、補題 4の証明を付録に示しておく。

従って、以下では、式 (1) のもう一方の不等式

$$\limsup_{n\to\infty} \frac{l(\varphi_n((x,y)_1^n))}{n} \leq H_{X|Y} \quad a.\,s. \qquad (3)$$

を文献 [5] の補題 4.3.2. と同様に示す。尚、便宜上 $x_1^m = x_1x_2\ldots x_m$ などを x^m と書くことにする。

ある $0<\epsilon<1$ と正の整数 m を固定し、確率分布 $\mu_{XY}(x,y)$ に従う定常エルゴード情報源 (X,Y) に対し、集合 $C_\epsilon^m \subset \mathcal{X}^m\times\mathcal{Y}^m$ を

$$\begin{aligned} C_\epsilon^m = \Bigg\{ & (x,y)^m \in \mathcal{X}^m\times\mathcal{Y}^m : \\ & \left| -\frac{1}{m}\log\mu_{XY}^m((x,y)^m) - H_{XY} \right| \leqq {\delta_\epsilon}^2/8 \\ & \text{and } \left| -\frac{1}{m}\log\mu_Y^m(y^m) - H_Y \right| \leqq {\delta_\epsilon}^2/8 \Bigg\} \end{aligned}$$

によって定める。但し、H_{XY} と H_Y はそれぞれ情報源 (X,Y) および Y のエントロピーレートを表し、

$$\delta_\epsilon = \min\left\{ 2h^{-1}(\epsilon/8), \frac{\epsilon}{4\log|\mathcal{X}|} \right\}$$

とする。また、h^{-1} は、2 元エントロピー関数

$$\begin{aligned} h(\alpha) = & -\alpha\log_2\alpha - (1-\alpha)\log_2(1-\alpha), \\ & 0 \leqq \alpha \leqq 1/2 \end{aligned}$$

の逆関数である。このとき、十分大きな m に対して

$$\mu_{XY}^m(C_\epsilon^m) \geqq 1-{\delta_\epsilon}^2/4 \tag{4}$$

が成り立つことは、よく知られている [13]。以下では、式 (4) が成り立つように m を選ぶ。

集合 C_ϵ^m を用いて、$y^m \in \mathcal{Y}^m$ によって定まる $\mathcal{X}^m$ の部分集合 $T_\epsilon^m(y^m)$ を

$$T_\epsilon^m(y^m) \triangleq \{x^m \in \mathcal{X}^m : (x^m,y^m) \in C_\epsilon^m\}$$

によって定義すると、C_ϵ^m の定義から、

$$\begin{aligned}
&|T_\epsilon^m(y^m)|2^{-m\left(H_{X|Y}+\frac{{\delta_\epsilon}^2}{4}\right)} \\
&\qquad = \sum_{x^m\in T_\epsilon^m(y^m)} 2^{-m\left(H_{X|Y}+\frac{{\delta_\epsilon}^2}{4}\right)} \\
&\qquad = \sum_{x^m\in T_\epsilon^m(y^m)} \frac{2^{-m\left(H_{XY}+\frac{{\delta_\epsilon}^2}{8}\right)}}{2^{-m\left(H_Y-\frac{{\delta_\epsilon}^2}{8}\right)}} \\
&\qquad < \sum_{x^m\in T_\epsilon^m(y^m)} \frac{\mu_{XY}^m((x,y)^m)}{\mu_Y^m(y^m)} \\
&\qquad = \sum_{x^m\in T_\epsilon^m(y^m)} \mu_{X|Y}^m(x^m|y^m) \\
&\qquad \leqq 1
\end{aligned}$$

が成り立つ。従って、$T_\epsilon^m(y^m)$ の要素数は $y^m \in \mathcal{Y}^m$ によらず

$$\begin{aligned}
|T_\epsilon^m(y^m)| &\leqq 2^{m\left(H_{X|Y}+\frac{{\delta_\epsilon}^2}{4}\right)} \\
&\leqq 2^{m\left(H_{X|Y}+\frac{\epsilon}{8}\right)}
\end{aligned} \tag{5}$$

で上から抑えられる。

次に、任意の $k(>m)$ および $(x^k,y^k) \in \mathcal{X}^k \times \mathcal{Y}^k$ に対して、インデックスの集合を、

$$\begin{aligned}
J_\epsilon^m(x^k,y^k) \triangleq \Big\{&j : 1 \leqq j \leqq k-m+1, \\
&x_j^{j+m-1} \in T_\epsilon^m(y_j^{j+m-1})\Big\}
\end{aligned}$$

によって定義する。このとき、$y^k \in \mathcal{Y}^k$ に対し、$\mathcal{T}_\epsilon^k(y^k)$ を、

$$\mathcal{T}_\epsilon^k(y^k) \triangleq \left\{x^k \in \mathcal{X}^k : \frac{\left|J_\epsilon^m(x^k,y^k)\right|}{k} \geqq 1-\frac{\delta_\epsilon}{2}\right\}$$

によって定義する。すなわち、$T_\epsilon^m(y^m)$ によって $(1-\frac{\delta_\epsilon}{2})$ 強被覆されるような $\mathcal{X}^k$ の系列の集合が $\mathcal{T}_\epsilon^k(y^k)$ である。ここで、$x^k \in \mathcal{T}_\epsilon^k(y^k)$ に対して、$j_v(v=1,2,\ldots)$ を次のように定義する。

$$j_1 \triangleq \min\left\{j : 1 \leqq j \leqq k, j \in J_\epsilon^m(x^k,y^k)\right\}$$

$$j_{v+1} \triangleq \min\left\{j : j_v+m \leqq j \leqq k, j \in J_\epsilon^m(x^k,y^k)\right\}$$

$$(v=2,3,\ldots)$$

また、v_{max} によって、j_v が定義できる最大の v を表す。このとき、$x^k \in \mathcal{T}_\epsilon^k(y^k)$ は、v_{max} 個の $T_\epsilon^m(y_j^{j+m-1})$ $(1 \leqq j \leqq k-m+1)$ に属する部分列と、$k-v_{max}m$ 個の $T_\epsilon^m(y_j^{j+m-1})$ に属さない記号に分割することができる。$\mathcal{T}_\epsilon^k(y^k)$ の定義から、明らかに

$$1 \leqq v_{max} \leqq \frac{k}{m}$$

かつ、

$$0 \leqq k-v_{max}m \leqq \frac{\delta_\epsilon k}{2} < \frac{k}{2}$$

であり、これらの分割の組合せの総数 $\binom{v_{max}+k-v_{max}m}{k-v_{max}m}$ は $\binom{k}{\delta_\epsilon k/2}$ という上界を持つ。このことと、式 (5) から、集合 $\mathcal{T}_\epsilon^k(y^k)$ の要素数に関して、以下の不等式が確率 1 で成立する。

$$\begin{aligned}
&|\mathcal{T}_\epsilon^k(y^k)| \\
&\leqq \sum_{i=0}^{\lfloor\delta_\epsilon k/2\rfloor}\binom{k}{\delta_\epsilon k/2}|\mathcal{X}|^i\left[\max_{y^m\in\mathcal{Y}^m}|T_\epsilon^m(y^m)|\right]^{\frac{k-i}{m}} \\
&\leqq \sum_{i=0}^{\lfloor\delta_\epsilon k/2\rfloor}\binom{k}{\delta_\epsilon k/2}|\mathcal{X}|^{\frac{\delta_\epsilon k}{2}}\left[2^{m[H_{X|Y}+\epsilon/8]}\right]^{\frac{k}{m}} \\
&< 2^{k[h(\delta_\epsilon/2)+\frac{\delta_\epsilon\log|\mathcal{X}|}{2}+H_{X|Y}+\frac{\epsilon}{8}]} \\
&< 2^{k[H_{X|Y}+\frac{3\epsilon}{8}]}
\end{aligned} \tag{6}$$

他方、エルゴード定理 [12] から、ほとんど全ての系列 $(xy) \in (\mathcal{X}\times\mathcal{Y})^\infty$ に対して、

$$\lim_{n\to\infty}\frac{1}{n-m+1}\sum_{i=1}^{n-m+1}\chi((x,y)_i^{i+m-1}) = \mu(C_\epsilon^m)$$

が成り立つ。但し、関数 $\chi : (\mathcal{X}\times\mathcal{Y})^m \to \{0,1\}$ は、

$$\chi((x,y)^m) = \begin{cases} 1 & \text{if } (x,y)^m \in C_\epsilon^m \\ 0 & \text{otherwise} \end{cases}$$

によって定義される。これと式 (4) を組み合わせれば、ある正の整数 $N(xy)$ が存在して $n \geqq N(xy)$ のとき、$(xy)_1^n$ が C_ϵ^m によって $(1-\delta_\epsilon^2/2)$ 強被覆されることが分かる。このとき、系列 $(xy)_1^n$ が

$$(xy)_1^n = (xy)_{\tilde{n}_0+1}^{\tilde{n}_1}(xy)_{\tilde{n}_1+1}^{\tilde{n}_2}\dots(xy)_{\tilde{n}_{p-1}+1}^{\tilde{n}_p} \quad (7)$$

と互いに異なる語に分割されるならば、部分列 $(xy)_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}, (i=1,2,\dots,p)$ のうち、C_ϵ^m によって $(1-\delta_\epsilon/2)$ 強被覆されていないものの合計長に関して

$$\begin{aligned}
&\sum_{i:(xy)_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}\text{-is-not-covered by } C_\epsilon^m} (\tilde{n}_i - \tilde{n}_{i-1}) \\
&\leqq \sum_{i:(xy)_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}\text{-is-not-covered by } C_\epsilon^m} \frac{2A_i}{\delta_\epsilon} \\
&\leqq \frac{2}{\delta_\epsilon}\sum_i^p A_i \\
&\leqq \delta_\epsilon n \\
&\leqq \frac{\epsilon n}{4\log|\mathcal{X}|} \\
&\leqq \frac{\epsilon n}{2} \qquad (8)
\end{aligned}$$

が成り立つ。但し、

$$A_i = \left|\left\{w \in \{\tilde{n}_{i-1},\dots,\tilde{n}_i - m + 1\} : x_w^{w+m-1} \notin C_\epsilon^m\right\}\right|$$

であり、最初の不等式は、$\delta_\epsilon(\tilde{n}_i - \tilde{n}_{i-1})/2 \leqq A_i$ から、3 番目の不等式は、$(xy)_1^n$ が C_ϵ^m によって $(1-\frac{{\delta_\epsilon}^2}{2})$ 強被覆されることから、

$$\sum_{i=1}^p A_i \leqq \frac{{\delta_\epsilon}^2 n}{2}$$

が成り立つことを用いた。これから、$(xy)_1^n$ が C_ϵ^m によって $(1-\frac{{\delta_\epsilon}^2}{2})$ 強被覆され、かつ式 (7) のように互いに異なる語に分割されるなら、C_ϵ^m によって $(1-\frac{\delta_\epsilon}{2})$ 強被覆されない語 $(xy)_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}$ の長さは合計で高々 $\epsilon n/2$ しかないことがわかる。すなわち、$x_{\tilde{n}_{i-1}+1}^{\tilde{n}_i} \in \mathcal{T}_\epsilon^{\tilde{n}_i-\tilde{n}_{i-1}}(y_{\tilde{n}_{i-1}+1}^{\tilde{n}_i})$ である $x_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}$ の合計長は少なくとも $(1-\frac{\epsilon}{2})n$ ある。

この結果と補題 1 を組み合わせると、$\ell(x_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}) > \frac{\log n}{H_{XY}+\epsilon/8}$ でない、または $x_{\tilde{n}_{i-1}+1}^{\tilde{n}_i} \notin \mathcal{T}_\epsilon^{\tilde{n}_i-\tilde{n}_{i-1}}(y_{\tilde{n}_{i-1}+1}^{\tilde{n}_i})$ である $x_{\tilde{n}_{i-1}+1}^{\tilde{n}_i}$ の合計長は高々 $5\epsilon n/8$ しかない。従って、補題 2 と補題 3から

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{l(\varphi_n((xy)_1^n))}{n} \\
&\leqq \limsup_{n\to\infty}\left\{\frac{\sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)\log c_i(x_1^n|y_1^n)}{n} + O\left(\frac{\log\log n}{\log n}\right)\right\} \\
&\leqq \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{c(y_1^n)} c_i(x_1^n|y_1^n)\log c_i(x_1^n|y_1^n)
\end{aligned}$$

が得られる。ここで、補題 1 と式 (8) を用いると、

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{l(\varphi_n((xy)_1^n))}{n} \\
&\leqq \limsup_{n\to\infty}\frac{1}{n}\left\{\sum_{\substack{i=1;\\ \ell(y(i))>\frac{\log n}{H_{XY}+\epsilon/8}}}^{c(y_1^n)} c_i(x_1^n|y_1^n) \times \log\max_{y(i)}\left|\mathcal{T}_\epsilon^{\ell(y(i))}(y(i))\right| + C_{(xy)^{5\epsilon n/8}}\log\frac{5\epsilon n}{8}\right\}
\end{aligned}$$

が得られる。但し、$C_{(xy)^{5\epsilon n/8}}$ は長さ $5\epsilon n/8$ の系列 $(xy)^{5\epsilon n/8}$ を互いに異なる部分列に分割したときの部分列の最大の個数であり、

$$\frac{C_{(xy)^{5\epsilon n/8}}\log(5\epsilon n/8)}{5\epsilon n/8} \leqq \frac{\log|\mathcal{X}||\mathcal{Y}|}{1-\eta_n} \qquad (9)$$

が成立する [14]。但し、$n\to\infty$ のとき $\eta_n \to 0$ である。式 (6) と式 (9) を代入すると、

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{l(\varphi_n((xy)_1^n))}{n} \\
&\leqq \limsup_{n\to\infty}\frac{1}{n}\left\{\sum_{\substack{i=1;\\ \ell(y(i))>\frac{\log n}{H_{XY}+\epsilon/8}}}^{c(y_1^n)} c_i(x_1^n|y_1^n)\ell(y(i)) \times(H_{X|Y}+\frac{3\epsilon}{8}) + \frac{5\epsilon n\log|\mathcal{X}||\mathcal{Y}|}{8(1-\eta_n)}\right\} \\
&\leqq \limsup_{n\to\infty}\frac{1}{n}\left\{n(H_{X|Y}+\frac{3\epsilon}{8}) + \frac{5\epsilon n\log|\mathcal{X}||\mathcal{Y}|}{8(1-\eta_n)}\right\} \\
&\leqq H_{X|Y}+\varepsilon
\end{aligned}$$

が確率 1 で成立する。但し、$\varepsilon = \epsilon/\log|\mathcal{X}||\mathcal{Y}|$ とした。$\epsilon > 0$ は任意なので、これで式 (3) が示された。

(証明終)

4. むすび

本稿では、任意の副情報源を伴う情報源を増分分解に基づいて符号化するZivによるアルゴリズムを述べた後、情報源が定常エルゴード情報源の場合、Zivの符号化アルゴリズムの漸近最良性を示した。今後の課題としてはZivの符号化アルゴリズムが条件付きエントロピーレートに収束する速度を調べること、適応型算術符号を利用した、副情報源を伴う符号化法を検討すること等が挙げられる。

謝辞

本研究の一部は日本学術振興会未来開拓学術研究推進事業JSPS-RFTF97P00601の援助により行なわれた。

文献

[1] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. on Inform. Theory*, vol.IT-19, pp.471-480, Jul. 1973.

[2] J. Ziv, “Fixed-Rate encoding of individual sequences with side information,” *IEEE Trans. on Inform. Theory*, vol.IT-30, pp.348-352, March 1984.

[3] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. on Inform. Theory*, vol.IT-24, no.5, pp.530-536, Sep. 1978.

[4] J. Ziv, “Universal decoding for finite-state channels,” *IEEE Trans. on Inform. Theory*, vol.IT-31, pp.453-460, Jul. 1985.

[5] J. Muramatsu, “Universal data compression algorithms for stationary ergodic sources based on the complexity of sequences,” Ph.D.Thesis, Nagoya University, Nagoya, Japan, 1998.

[6] J. Ziv and A. Lempel, “A Universal algorithm for sequential data compression,” *IEEE Trans. on Inform. Theory*, vol.IT-23, no.3, pp.337-343, May 1977.

[7] P. Subrahmanya and T. Berger, “A sliding window Lempel-Ziv algorithm for differential layer encoding in progressive transmission,” *Proc. of 1995 IEEE Int. Symp. on Inform. Theory*, p.266, Sep. 1995.

[8] E.-H. Yang, A. Kaltchenko and J. C. Kieffer, “Universal lossless data compression with side information by using a conditional MPM grammer transform,” *Proc. of 2000 IEEE Int. Symp. on Inform. Theory*, p.298, June 2000.

[9] P. Elias, “Universal codeword sets and representations of the integers,” *IEEE Trans. on Inform. Theory*, vol.IT-21, pp.194-203, 1975.

[10] D. S. Ornstein and B. Weiss, “Entropy and data compression schemes,” *IEEE Trans. on Inform. Theory*, vol.39, pp.78-83, Jan. 1993.

[11] 前田 浩次, 植松 友彦, “副情報源を伴う情報源の増分分解に基づくユニバーサル符号化,” 第23回情報理論とその応用シンポジウム予稿集, pp.503-506, Oct. 2000.

[12] P. C. Shields, “The Ergodic Theory of Discrete Sample Paths,” American Mathmatical Society, 1996.

[13] T. M. Cover, “A proof of the data compression theorem of Slepian-Wolf for ergodic sources,” *IEEE Trans. on Inform. Theory*, vol.IT-21, pp.226-228, March 1975.

[14] J. Ziv and A. Lempel, “On the complexity of finite sequences,” *IEEE Trans. on Inform. Theory*, vol.IT-22, pp.75-81, Sep. 1976.

付録

補題4の証明

定常エルゴード情報源Yに対し、漸近的最良な無損失の可変長符号の系列が存在する[3]。すなわち、

$$\tilde{\psi}_n : \mathcal{Y}^n \to \{0,1\}^*$$

$$\tilde{\psi}_n^{-1} : \tilde{\psi}_n(\mathcal{Y}^n) \to \mathcal{Y}^n$$

であり、

$$\limsup_{n\to\infty} \frac{1}{n} l(\tilde{\psi}_n(y_1^n)) \leqq H_Y$$

かつ

$$\tilde{\psi}_n^{-1}(\tilde{\psi}_n(y_1^n)) = y_1^n \quad \forall y_1^n \in \mathcal{Y}^n$$

を満足する符号列$\{(\tilde{\psi}_n, \tilde{\psi}_n^{-1})\}$が存在する。

さて、与えられた$\{(\psi_n, \psi_n^{-1})\}$は副情報源を伴う無損失符号であるから、

$$\psi_n^{-1}(\psi_n(x_1^n, y_1^n), y_1^n) = x_1^n, \quad \forall x_1^n \in \mathcal{X}^n, \forall y_1^n \in \mathcal{Y}^n$$

を満足する。符号(ψ_n, ψ_n^{-1})と$(\tilde{\psi}_n, \tilde{\psi}_n^{-1})$を用いて情報源$(X,Y)$に対する符号$(\phi_n, \phi_n^{-1})$

$$\phi_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{0,1\}^*$$

$$\phi_n^{-1} : \phi_n(\mathcal{X}^n \times \mathcal{Y}^n) \to \mathcal{X}^n \times \mathcal{Y}^n$$

を以下のように構成する。

$$\phi_n(x_1^n, y_1^n) \stackrel{\triangle}{=} \psi_n(x_1^n, y_1^n) * \tilde{\psi}_n(y_1^n)$$

$$\phi_n^{-1}(b^* * \bar{b}^*) \stackrel{\triangle}{=} (\psi_n^{-1}(b^*, \tilde{\psi}_n^{-1}(\bar{b}^*)), \tilde{\psi}_n^{-1}(\bar{b}^*))$$

但し、$*$は文字列の連接を表し、

$$b^* \stackrel{\triangle}{=} \psi_n(x_1^n, y_1^n)$$

$$\bar{b}^* \triangleq \widetilde{\psi}_n(y_1^n)$$

とする。このとき、ψ_n と $\widetilde{\psi}_n$ は語頭符号なので、ϕ_n もまた語頭符号である。更に、ϕ_nと$\widetilde{\psi}_n$ は単射なので、

$$\begin{aligned}
&\phi_n^{-1}(\phi_n(x_1^n, y_1^n)) \\
&= (\psi_n^{-1}(\psi_n(x_1^n, y_1^n), \widetilde{\psi}_n^{-1}(\widetilde{\psi}_n(y_1^n))), \widetilde{\psi}_n^{-1}(\widetilde{\psi}_n(y_1^n))) \\
&= (\psi_n^{-1}(\psi_n(x_1^n, y_1^n), y_1^n), y_1^n) \\
&= (x_1^n, y_1^n)
\end{aligned}$$

であり、(ϕ_n, ϕ_n^{-1}) が無損失符号であることが分る。従って、概収束符号化逆定理 [12] より、

$$\liminf_{n\to\infty} \frac{1}{n} l(\phi_n(x_1^n, y_1^n)) \geqq H_{XY}$$

が成立する。これから直ちに、

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} l(\psi_n(x_1^n, y_1^n)) + H_Y \\
&\geqq \liminf_{n\to\infty} \frac{1}{n} l(\psi_n(x_1^n, y_1^n) + \limsup_{n\to\infty} \frac{1}{n} l(\widetilde{\psi}_n(y_1^n)) \\
&\geqq \liminf_{n\to\infty} \frac{1}{n} [l(\psi_n(x_1^n, y_1^n) + l(\widetilde{\psi}_n(y_1^n))] \\
&= \liminf_{n\to\infty} \frac{1}{n} l(\phi_n(x_1^n, y_1^n)) \\
&\geqq H_{XY}
\end{aligned}$$

が確率 1 で成り立つので、

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} l(\psi_n(x_1^n, y_1^n)) &\geqq H_{XY} - H_Y \\
&= H_{X|Y}
\end{aligned}$$

が成り立ち、補題 4は証明された。 (証明終)

（平成 13 年?月?日受付，?月?日再受付）

植松　友彦 （正員）

昭 57 東工大・工・電気電子卒．昭 59 同大大学院修士課程了．同年同大・工・電気電子工学科助手，同講師を経て平 3 同助教授．平 4 北陸先端大・情報科学研究科助教授．平 9 東工大・工・電気電子工学科助教授．工博．情報理論，特にシャノン理論の研究に従事．昭 63 年度本会篠原記念学術奨励賞受賞．平 4 年度ならびに平 7 年度本会論文賞受賞．著書「文書データ圧縮アルゴリズム入門」,「よくわかる通信工学」,「現代シャノン理論」など．IEEE，情報理論とその応用学会各会員．

前田　浩次

平 11 東工大・工・電気電子卒．平 13 同大・大学院修士課程修了．同年株式会社 NTT ドコモ勤務．在学中，情報理論，特に情報源符号化の研究に従事．