Lempel-Ziv 符号と情報理論

植松　友彦[†a)]

Lempel-Ziv Coding as a Tool for Information Theory

Tomohiko UYEMATSU[†a)]

あらまし　情報源の統計的性質をあらかじめ知ることなく符号化・復号化が行え，かつ入力列が長くなるにつれて，平均符号長が情報源のエントロピーレートに漸近する符号をユニバーサル符号という．本論文では，代表的なユニバーサル符号として，Ziv と Lempel によって提案された増分分解に基づく Lempel-Ziv（LZ78）符号を取り上げる．まず，LZ78 符号の漸近的最良性と平均冗長度について詳細に述べた後，LZ78 符号が情報源の符号化のみならず情報理論における各種問題に対する最適解を具体的なアルゴリズムとともに与えることを示し，ユニバーサル符号の情報理論における重要性を明らかにする．

キーワード　Lempel-Ziv 符号，ユニバーサル符号，漸近的最良性，平均冗長度，シャノン理論

1. まえがき

情報源の統計的性質をあらかじめ知ることなく符号化・復号化が行え，かつ入力列が長くなるにつれて，平均符号長が情報源のエントロピーレートに漸近する符号をユニバーサル符号という．ユニバーサル符号の研究は，ロシアの研究者 Fitingof [1] によって組合せ論的な考え方を用いて創始された [2]．これと同時期に米国の Lynch と Davisson が，2 元の定常エルゴード情報源についてやや実用的なユニバーサル符号を提案した [3], [4]．70 年代に入ると，これらの研究者による啓蒙活動が原動力となり，ユニバーサル符号は活発に研究されるようになり，Ziv と Lempel による 2 種の Lempel-Ziv（LZ）符号の提案 [5], [6] によって一つの頂点に達したと考えられる．

本論文では，代表的なユニバーサル符号として，Ziv と Lempel によって提案された増分分解に基づいた Lempel-Ziv（LZ78）符号 [5] を取り上げる．まず，LZ78 符号の漸近的最良性と平均冗長度の上界について詳細に述べる．次に，LZ78 符号が，情報源の符号化のみならず，乱数の生成，仮説検定，通信路のユニバーサル復号法などの情報理論の各種問題に対する最適解を具体的アルゴリズムとともに与えることを示し，ユニバーサル符号の情報理論における重要性の一端を明らかにする．

以下では，情報源アルファベット $\mathcal{X}$ は有限とし，$\mathcal{X}$ 上の無限系列を $x(= x_1x_2\cdots)$ によって，また x の部分列 $x_ix_{i+1}\cdots x_j$ を x_i^j によって表す．特に断らない限り，本論文で取り扱う符号は，2 元の符号アルファベット $\{0,1\}$ を有する可変長符号であり，語頭条件を満足するものとする．また，$\{0,1\}$ 上の有限系列の集合を $\{0,1\}^*$ によって表す．なお，log の底は 2 とする．

2. LZ78 符号とその漸近的最良性

本章では，代表的なユニバーサル符号である LZ78 符号について，符号化アルゴリズムとその漸近的最良性について述べる．

2.1 LZ78 符号

LZ78 符号では，符号化する系列 $x_1^n \in \mathcal{X}^n$ を，次の 2 条件を満足する部分列に分解することが基本になっている [5]．

（条件 1）　最後の部分列を除いたすべての部分列は互いに異なっている．

（条件 2）　j 番目の部分列は，それまでに出現した $i(< j)$ 番目の部分列に最後の 1 記号が付加されたものである．

この条件を満足する分解のことを増分分解（incremental parsing）と呼ぶ．例えば，系列 $x_1^{14} =$

† 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 Integrated Systems, Tokyo Institute of Technology, Tokyo, 152–8552 Japan
a) E-mail: uematsu@it.ss.titech.ac.jp

01101011101001 は，増分分解によって

$$x_1^{14} = 0|1|10|101|11|01|00|1 \qquad (1)$$

のように分解される．増分分解によって系列 x_1^n を分解したとき，j 番目の部分列に対し，最後の 1 記号を除いて一致する部分列を表すインデックス i と最後の 1 記号を合計 $\lceil \log j|\mathcal{X}| \rceil$ ビットの 2 元系列で表したものが LZ78 符号である(注 1)．

増分分解は,「LZ 分解木」と呼ばれる木によって容易に行える．系列 $x_1^n \in \mathcal{X}^n$ が与えられたとき，LZ 分解木は次のアルゴリズムによって構成される．

[LZ 分解木の構成アルゴリズム]

Step.1　根だけからなる木を作り，

$$i \longleftarrow 1, \quad j \longleftarrow 1$$

とし，ポインタを根におく．

Step.2　ポインタのある節点から x_j と同じラベルを有する枝をたどり，その枝の先の節点にポインタを移動する．もし，該当する枝がない場合は，Step.5 へ行く．

Step.3　$j = n$ ならば，x_i^j を最後の部分列としてアルゴリズムを終了する．

Step.4　$j \longleftarrow j+1$ として，Step.2 へ戻る．

Step.5　x_i^j を新たな部分列とする．ポインタの示す節点から新しい枝を 1 本伸ばし，その枝にラベル x_j を付加する．$j = n$ ならば，アルゴリズムを終了する．それ以外の場合は，

$$i \longleftarrow j+1, \quad j \longleftarrow j+1$$

として，再びポインタを根に戻した後，Step.2 に戻る．　□

一例として，式 (1) の増分分解に対応する LZ 分解木を図 1 に示す．

LZ 分解木は次の性質を有する [7].

（1）　木の枝数 e は，

$$e = \begin{cases} t, & \text{すべての部分列が異なるとき} \\ t-1, & \text{上記以外のとき} \end{cases}$$

を満足する．ただし，t は増分分解によって得られる部分列の個数である．

（2）　各々の内部節点からはたかだか $|\mathcal{X}|$ 個の枝が生えており，これらの枝は，ラベルとして $\mathcal{X}$ 上の相異なる記号を有している．

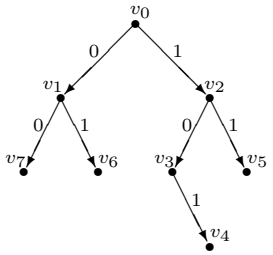

図 1　LZ 分解木の例
Fig. 1　An example of LZ parsing tree.

（3）　根から出発するパス（経路）と増分分解によって得られる相異なる部分列との間には 1 対 1 の対応がある．例えば，図 1 において根 v_0 から節点 v_j までのパスは，j 番目の部分列に対応している．

2.2　LZ78 符号の漸近最良性

次の定理は，LZ78 符号の定常エルゴード情報源 [8] に対する漸近最良性を表している．

[定理 1]（Ziv-Lempel [5]）　エントロピーレート $H(X)$ を有する定常エルゴード情報源 X からの長さ n の出力列 X_1^n を LZ78 符号によって符号化したときの符号長を $L_{LZ78}(X_1^n)$ とする．このとき，1 記号当りの符号長がエントロピーレートに確率 1 で収束する，すなわち，

$$\lim_{n\to\infty} \frac{1}{n} L_{LZ78}(X_1^n) = H(X) \quad a.s.$$

が成り立つ．

この定理の証明法としては，Ziv と Lempel による方法 [5]，Cover と Thomas による方法 [9, ch.12]，Ornstein と Weiss による方法 [10] がある．ここでは，増分分解以外の分解法にも適用可能な Ornstein と Weiss による証明法を示す．まず，証明に用いる定理と補題を述べる．

[定理 2]（The distinct-words theorem [10]）　定常エルゴード情報源 X と任意の $\epsilon > 0$ に対し，X からの出力列 $x \in \mathcal{X}^\infty$ の部分列 x_1^n が，相異なる部分列 $w(1), w(2), \cdots, w(t)$ によって

$$x_1^n = w(1)w(2)\cdots w(t)$$

(注 1)：$\lceil x \rceil$ は x 以上の最小整数を表し，$|\mathcal{X}|$ は集合 $\mathcal{X}$ の要素数を表す．

と分解されたとする．また $\ell(w(i))$ によって部分列 $w(i)$ の長さを表す．このとき，n が十分大きければ

$$\sum_{\ell(w(i))<\frac{\log n}{H(X)+\epsilon}} \ell(w(i)) \leqq n\epsilon \tag{2}$$

が確率 1 で成り立つ．すなわち，$w(i)$ $(i=1,2,\cdots,t)$ の中で，長さが $\log n/(H(X)+\epsilon)$ 未満のものの総長はたかだか $n\epsilon$ しかない．

［補題 1］（Ziv-Lempel [5]） $C(x_1^n)$ によって，$x_1^n \in \mathcal{X}^n$ を相異なる部分列に分解したときの部分列の最大数を表す．このとき，零に収束する正の数の系列 $\{\delta_n\}$ が存在して，

$$\frac{C(x_1^n)\log n}{n} \leqq (1+\delta_n)\log|\mathcal{X}|$$

が成り立つ．

以上の準備のもとで，定理 1 を証明する．増分分解によって x_1^n が $t(x_1^n)$ 個の部分列に分解されるならば，x_1^n を LZ78 符号化したときの符号長 $L_{LZ78}(x_1^n)$ について，

$$\begin{aligned} L_{LZ78}(x_1^n) &\leqq \sum_{i=1}^{t(x_1^n)} \lceil \log i|\mathcal{X}| \rceil \\ &< \sum_{i=1}^{t(x_1^n)} (\log i + \log|\mathcal{X}| + 1) \\ &\leqq t(x_1^n)(\log n + \log|\mathcal{X}| + 1) \end{aligned} \tag{3}$$

が成り立つ．他方，増分分解によって得られた部分列は，最後の部分列を除いて必ず相異なる．最後の部分列がそれ以前に出現した部分列と一致すれば，最後の部分列の長さ ℓ_t は，$\ell_t \leqq n/2$ を満足する．したがって，系列長 n が十分長ければ，系列 $x_1^{n-\ell_t}$ に定理 2 を適用することで，長さが $\log(n-\ell_t)/(H(X)+\epsilon)$ 未満の部分列の長さの総和は $(n-\ell_t)\epsilon$ 未満であることがわかる．補題 1 から，そのような部分列の総数は，たかだか $(1+\delta_n)(n-\ell_t)\epsilon\log|\mathcal{X}|/\log((n-\ell_t)\epsilon)$ 個未満である．したがって，

$$\begin{aligned} t(x_1^n) &\leqq \frac{(n-\ell_t)(H(X)+\epsilon)}{\log(n-\ell_t)} \\ &\quad + \frac{(1+\delta_n)(n-\ell_t)\epsilon}{\log((n-\ell_t)\epsilon)}\log|\mathcal{X}| + 1 \\ &\leqq \frac{n(H(X)+\epsilon)}{\log n} \\ &\quad + \frac{(1+\delta_n)n\epsilon}{\log(n\epsilon)}\log|\mathcal{X}| + 1 \end{aligned} \tag{4}$$

が成り立つ．したがって，式 (3) と式 (4) とを組み合わせることで，

$$\limsup_{n\to\infty}\frac{1}{n}L_{LZ78}(x_1^n) \leqq H(X) + \epsilon(1+\log|\mathcal{X}|)$$

が得られる．ここで，$\epsilon > 0$ は任意に小さくとれるので，

$$\limsup_{n\to\infty}\frac{1}{n}L_{LZ78}(x_1^n) \leqq H(X)$$

が確率 1 で成り立つ．

逆向きの不等式の成立は，「概収束符号化逆定理」と呼ばれる次の定理によって示される．

［定理 3］（Baron [11]） $\{f_n\}$ によって語頭符号 $f_n : \mathcal{X}^n \to \{0,1\}^*$ の系列を表す．エントロピーレート $H(X)$ を有する定常エルゴード情報源 X からの長さ n の出力列 X_1^n を符号 f_n によって符号化したときの符号長を $L_f(X_1^n)$ とする．このとき，

$$\liminf_{n\to\infty}\frac{1}{n}L_f(X_1^n) \geqq H(X) \quad a.s. \tag{5}$$

が成り立つ．

（証明） 情報源 X の確率測度を μ によって表す．各々の n に対し，$\mathcal{X}^n$ の部分集合 B_n を

$$B_n \triangleq \{x_1^n \in \mathcal{X}^n : L_f(x_1^n) + \log\mu(x_1^n) \leqq -2\log n\}$$

によって定義すれば，簡単な計算により，

$$B_n = \{x_1^n \in \mathcal{X}^n : \mu(x_1^n) \leqq 2^{-L_f(x_1^n)}n^{-2}\}$$

が成り立つ．クラフトの不等式

$$\sum_{x_1^n\in\mathcal{X}^n} 2^{-L_f(x_1^n)} \leqq 1$$

に注意すれば，

$$\begin{aligned} \sum_{n=1}^{\infty}\mu(B_n) &= \sum_{n=1}^{\infty}\sum_{x_1^n\in B_n}\mu(x_1^n) \\ &\leqq \sum_{n=1}^{\infty} n^{-2}\sum_{x_1^n\in B_n} 2^{-L_f(x_1^n)} \\ &\leqq \sum_{n=1}^{\infty} n^{-2} \\ &< \infty \end{aligned}$$

が成り立つので，Borel-Cantelli の補題 [9] から，

$$L_f(x_1^n) \geqq -\log \mu(x_1^n) - 2\log n$$

が十分大きな n について確率 1 で成り立つ．ここで，Shannon-McMillan-Breiman の定理 [9, Theorem 15.7.1] から，

$$\lim_{n\to\infty} -\frac{1}{n}\log \mu(x_1^n) = H(X)$$

が確率 1 で成り立つことに注意すれば，式 (5) の成立がわかる． (Q.E.D.)

3. LZ78 符号の平均冗長度

語頭符号 f によって系列 $x_1^n \in \mathcal{X}^n$ を符号化したときの（pointwise な）冗長度を

$$r_f(x_1^n) \stackrel{\triangle}{=} \frac{1}{n}(L_f(x_1^n) + \log \mu(x_1^n))$$

で定義する．ただし，$L_f(x_1^n)$ は符号 f によって系列 x_1^n を符号化したときの符号長を表し，$\mu(x_1^n)$ は対象となる情報源における系列 x_1^n の生起確率である．また，この冗長度の平均，すなわち

$$E[r_f] \equiv E[r_f(X_1^n)] \stackrel{\triangle}{=} \sum_{x_1^n \in \mathcal{X}^n} \mu(x_1^n) r_f(x_1^n)$$

を平均冗長度と呼ぶ．定義から明らかに

$$E[r_f] = \frac{1}{n}E[L(X_1^n)] - H(X)$$

が成り立つ．これと，可変長符号化逆定理 [8, 定理 3.9] から直ちに $E[r_f] \geqq 0$ が導ける．

LZ78 符号の平均冗長度の上界を求める研究は，80 年代後半から行われてきた．そして平均冗長度が

$$E[r_{LZ78}] \leqq O\left(\frac{\log\log n}{\log n}\right)$$

を満足することは，既に Cover らによって示されていた [9, Theorem 12.10.1]．また，Plotnik-Weinberger-Ziv [12] は，かなり広範な情報源について上記の上界式が成り立つことを示した．しかしながら，彼らは実際のシミュレーションの結果から見るとこの冗長度は大きすぎると指摘した．他方，Shtarkov と Tjalkens [13] は，無記憶情報源あるいはマルコフ情報源 [8] について LZ78 符号の平均冗長度は $O(1/\log n)$ であるとの予想を述べ，LZ78 符号の平均冗長度のオーダを決定する問題が大きく脚光を浴びるに至った．

この問題について，Louchard-Szpankowski は，2 元無記憶情報源における LZ78 符号の平均冗長度の最高次のオーダのみならず，その係数までを完全に定めた．彼らの成果を次の定理に示す．

［定理 4］（Louchard-Szpankowski [14]） 2 元無記憶情報源において記号 0 及び記号 1 が出現する確率を，それぞれ $p(>0)$ 及び $q(=1-p)$ とするとき，この情報源からの長さ n の系列を LZ78 符号によって符号化したときの平均冗長度は，

$$E[r_{LZ78}] = h(p)\frac{2-\gamma-\frac{h_2}{2h(p)}+\alpha-\delta_0(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right)$$

である．ただし，$h(x)$ はエントロピー関数であり，$\gamma = 0.577\cdots$ はオイラー定数であり，h_2 と α は，$h_2 \stackrel{\triangle}{=} p\log^2 p + q\log^2 q$ 及び

$$\alpha \stackrel{\triangle}{=} -\sum_{k=1}^{\infty}\frac{p^{k+1}\log p + q^{k+1}\log q}{1-p^{k+1}-q^{k+1}}$$

で定まる定数である．更に，関数 $\delta_0(x)$ は平均零で微小な振幅を有する振動項であり，$p=q=0.5$ のとき振幅 10^{-6} 以下，それ以外のときは $\lim_{x\to\infty}\delta_0(x)=0$ を満足する．

定理 4 は，無記憶情報源に関する LZ78 符号の性能について，最良の漸近的評価を与えている．すなわち LZ78 符号の平均冗長度が $O(\log\log n/\log n)$ ではなく，$O(1/\log n)$ で減衰することを示すとともに，平均冗長度の最高次の係数を初めて明らかにした．更に，彼らは，定常マルコフ情報源についても，平均冗長度のオーダは $1/\log n$ であると予想し，その係数についても予想を出しているが [14]，これは未解決である．なお，一般の定常エルゴード情報源については，平均冗長度が $1/\log n$ のオーダで減衰することは不可能であることが Shields によって示されている [15]．

他方，ユニフィラー情報源 [8] に対する LZ78 符号の平均冗長度の上界については次の定理が知られている．

［定理 5］（Savari [16]） 情報源アルファベット $\mathcal{X}$ と K 個の状態を有するユニフィラー情報源からの出力系列 x_1^n を LZ78 符号で符号化した際の冗長度は，

$$r_{LZ78}(x_1^n) \leqq \frac{-\log \mu^n(x_1^n)}{n\log n}\log\left(\frac{K(|\mathcal{X}|-1)}{H(X)}\right) + o\left(\frac{1}{\log n}\right) \quad (6)$$

を満足する．ただし，$H(X)$ は情報源のエントロピーレートを表す．

式 (6) の両辺の平均をとることによって，直ちに，

$$E[r_{LZ78}] \leqq \frac{h_{max}}{\log n}\log\left(\frac{K(|\mathcal{X}|-1)}{H(X)}\right)+o\left(\frac{1}{\log n}\right)$$

が得られ，ユニフィラー情報源についても LZ78 符号の平均冗長度が $1/\log n$ のオーダで減衰することがわかる．ただし，h_{max} は 1 文字当りの最大の自己情報量である．

これらの成果とは別に，Kieffer と Yang は，初等的な手法で無記憶情報源に対する LZ78 符号の冗長度を評価する方法を示した．以下では，彼らの成果とその手法について解説する．

［定理 6］（Kieffer-Yang [7]） 情報源 X が $\mathcal{X}$ 上の分布 P に従う無記憶情報源とする．このとき，$x_1^n \in \mathcal{X}^n$ について，

$$r_{LZ78}(x_1^n) \leqq \frac{8t(x_1^n)Q}{n} \tag{7}$$

が成り立つ．ただし，$t(x_1^n)$ は x_1^n を増分分解したときの部分列の数を表し，

$$Q = \max_{\substack{a\in\mathcal{X}:\\ P(a)>0}} (-\log P(a))$$

である．

この定理を用いて，LZ78 符号の無記憶情報源に対する冗長度が一様に $1/\log n$ のオーダで減衰することは次のようにして示せる．$t(x_1^n) \leqq C(x_1^n)+1$ に注意して，式 (7) の右辺に補題 1 を適用することで，

$$\max_{x_1^n\in\mathcal{X}^n} r_{LZ78}(x_1^n) \leqq O\left(\frac{1}{\log n}\right)$$

が得られる．また，この式は，定理 1 を無記憶情報源に制限したときの別証明を与えている．

定理の証明に先立ち，補題を一つ示す．

［補題 2］（Kieffer-Yang [7]） T を LZ 分解木とする．T の節点のうちで，ちょうど $|\mathcal{X}|$ 個の子を有する節点の集合を $V_0(T)$ によって表す．また，$V_0(T)$ に属していない T の節点の集合を $V_1(T)$ によって表す．このとき，もし $V_0(T)$ が空でないならば，$V_0(T)$ に属するすべての節点 v に対して，次の 2 条件を満足する T 上のパス（経路）π^v が存在する．

（1） $v'(\in V_0(T))$ と v が相異なる節点ならば，パス π^v とパス $\pi^{v'}$ は共通の枝をもたない．

（2） パス π^v は節点 v から始まり $V_1(T)$ に属する節点で終わる．

例えば，図 1 の LZ 分解木においては，$V_0(T)=\{v_0,v_1,v_2\}$，$V_1(T)=\{v_3,v_4,v_5,v_6,v_7\}$ である．また，(v_i,v_j) によって節点 v_i から節点 v_j への枝を表すと，パス $\pi^{v_0}=\{(v_0,v_1),(v_1,v_7)\}$，$\pi^{v_1}=\{(v_1,v_6)\}$，$\pi^{v_2}=\{(v_2,v_3)\}$ は補題の条件を満足する．

この補題の証明は文献 [7] を参照されたい．

（定理 6 の証明） $x_1^n \in \mathcal{X}^n$ を一つ固定し，T を x_1^n の LZ 分解木とする．また，e によって T の枝の数を表す．このとき e は増分分解によって生じる相異なる部分列の数 $t(x_1^n)$ に等しいか 1 だけ小さい．このことから，式 (3) と同様にして，

$$\begin{aligned} &L_{LZ78}(x_1^n)\\ &\quad\leqq (e+1)\log(e+1)+t(x_1^n)(\log|\mathcal{X}|+1)\\ &\quad\leqq e\log e+2+e+t(x_1^n)(\log|\mathcal{X}|+1) \end{aligned} \tag{8}$$

の成立がわかる．ただし，最初の不等式は

$$\sum_{i=1}^{t(x_1^n))} \log i \leqq t(x_1^n)\log t(x_1^n)$$

を用いた．ここで，枝の数 e と系列 x_1^n の生起確率 $\mu(x_1^n)$ の間に関係式

$$e\log e \leqq -\log\mu(x_1^n)+e(2Q+1) \tag{9}$$

が成り立つことが示されれば，式 (8) に式 (9) を代入して，不等式 $1\leqq\log|\mathcal{X}|\leqq Q$ を用いれば，式 (7) を得る．したがって，以下では，式 (9) の成立を示す．

T の根を v_0 とし，

$$\begin{aligned} V_0(T)^* &\triangleq V_0(T)-\{v_0\}\\ V_1(T)^* &\triangleq V_1(T)-\{v_0\}\\ V(T)^* &\triangleq V_0(T)^*\cup V_1(T)^* \end{aligned}$$

とする．各々の $v\in V(T)^*$ に対し，π_v によって根 v_0 から v に至る唯一のパスを表す．更に，T の任意のパス π に対し，$x(\pi)$ によってパス π をたどったときに得られる枝のラベルの系列を表す．さて，$\mathcal{X}$ 上の系列の集合 S_1 を

$$S_1 \triangleq \{x(\pi_v) : v\in V_1(T)^*\}$$

によって定義する．$V_1(T)$ の定義から，各々の $u\in S_1$ について，ある記号 $a(u)\in\mathcal{X}$ が存在し T の根から出発する系列 $u'=ua(u)$ に対応するパスは存在しない．このとき，$S_1'=\{u' : u'=ua(u), u\in S_1\}$ はプリフィックス集合なので，直ちに

$$\sum_{u'\in S_1'}\mu(u') \leqq 1$$

が得られ，$P(a(u)) \geqq 2^{-Q}$ に注意すれば，

$$\sum_{v \in V_1(T)^*} \mu(x(\pi_v)) \leqq 2^Q \qquad (10)$$

の成立がわかる．

もし $V_0(T)^*$ が空集合ならば，$|S_1| = e$ であり $V(T)^* = V_1(T)^*$ が成り立つ．したがって，対数和不等式 [8, 定理 2.3] から

$$e \log \frac{e}{\sum_{v \in V(T)^*} \mu(x(\pi_v))} \leqq \sum_{v \in V(T)^*} 1 \cdot \log \frac{1}{\mu(x(\pi_v))}$$

が成り立つ．左辺に式 (10) を代入して整理すると，

$$e \log e \leqq \left(\sum_{v \in V(T)^*} - \log \mu(x(\pi_v)) \right) + eQ \quad (11)$$

が得られる．集合 $\{x(\pi_v) : v \in V(T)^*\}$ は増分分解によって得られる相異なる部分列の集合と一致するので，

$$\sum_{v \in V(T)^*} - \log \mu(x(\pi_v)) \leqq - \log \mu(x_1^n) \qquad (12)$$

が成り立つ．この式を式 (11) に代入すると，

$$e \log e \leqq - \log \mu(x_1^n) + eQ$$

が得られ，これは式 (9) の成立を意味する．

次に，$V_0(T)^*$ が空集合でない場合を考える．補題 2 から，任意の $v \in V_0(T)^*$ に対し，節点 v から出発し，$V_1(T)^*$ に属する節点で終わるパス π^v が存在し，これらのパス $\{\pi^v : v \in V_0(T)^*\}$ は互いに共通の枝をもたない．各々の $v \in V_0(T)$ に対し，$\pi_v \pi^v$ によって，T 上でパス π_v とパス π^v を続けてたどって得られるパスを表し，$V(T)^*$ 上の関数 μ^* を

$$\mu^*(v) \triangleq \begin{cases} \mu(x(\pi_v)), & v \in V_1(T)^* \\ \mu(x(\pi_v \pi^v)), & v \in V_0(T)^* \end{cases}$$

によって定める．このときパス $\{\pi_v \pi^v : v \in V_0(T)^*\}$ は，$V_1(T)^*$ に属する相異なる節点で終わるので，

$$\sum_{v \in V_0(T)^*} \mu^*(v) \leqq \sum_{v \in V_1(T)^*} \mu(x(\pi_v))$$

が成り立つ．この式と式 (10) を組み合わせることで，

$$\begin{aligned} \sum_{v \in V(T)^*} \mu^*(v) &= \sum_{v \in V_0(T)^*} \mu^*(v) + \sum_{v \in V_1(T)^*} \mu^*(v) \\ &\leqq 2 \sum_{v \in V_1(T)^*} \mu^*(v) \\ &\leqq 2^{Q+1} \end{aligned}$$

が得られる．この式と，$|V(T)^*| = e$ から，式 (11) の導出と同様にして，

$$e \log e \leqq \left(\sum_{v \in V(T)^*} - \log \mu^*(v) \right) + e(Q+1)$$

が得られる．ここで，パス $\{\pi^v : v \in V_0(T)^*\}$ が共通の枝をもたないことから，

$$\begin{aligned} \sum_{v \in V_0(T)^*} - \log \mu^*(v) \leqq & \sum_{v \in V_0(T)^*} - \log(x(\pi_v)) \\ & + \sum_{e \in E(T)} - \log P(l(e)) \end{aligned}$$

が成り立つ．ただし，$E(T)$ は T の枝の集合を表し，$l(e)$ は枝 e のラベルを表す．この式の最後の項は eQ によって上から抑えられる．この結果と，

$$\sum_{v \in V_1(T)^*} - \log \mu^*(v) = \sum_{v \in V_1(T)^*} - \log \mu^*(x(\pi_v))$$

及び式 (12) を用いれば，式 (9) の成立がわかる．

(Q.E.D.)

なお，Kieffer-Yang が文献 [7] で指摘しているように，定理 6 は無記憶情報源のみならず，より一般的な情報源についても成立する．例えば，定理 6 がユニフィラー情報源についても成り立つことを示すのは良い演習問題である．

4. LZ78 符号の応用

本章では，LZ78 符号に代表されるユニバーサル符号の情報理論における各種の応用について述べる．

4.1 乱数生成への応用

もし情報源符号化によって系列中の冗長性が取り除かれるのならば，符号語中に出現する記号 0 と 1 はランダムに発生するはずである．このことは，情報源符号化によって真の乱数が生成できることを意味している．本節では，(ユニバーサルな) 可変長符号による乱数生成について，Visweswariah-Kulkarni-Verdú [17] 及び韓 [18] による成果を示す．なお，固定長符号による乱数生成については，文献 [18] を参照されたい．

まず，レート R の乱数生成器を定義する．

[定義 1] (Visweswariah-Kulkarni-Verdú [17]) 定常エルゴード情報源 X に対し，写像 $f_n : \mathcal{X}^n \to \{0,1\}^*$ の列がレート R の乱数生成器であるとは，入力系列長 n に依存する正整数の集合 G_n が存在して，次の条件を満足することである．

(i) $\lim_{n\to\infty} \Pr\{L_f(X^n) \in G_n\} = 1$

(ii) $\liminf_{n\to\infty} \frac{1}{n} \sum_{r\in G_n} r\Pr\{L_f(X^n) = r\} = R$

(iii) $\limsup_{n\to\infty} \max_{r\in G_n} \frac{1}{r} D(f_{n,r}(X^n) \parallel B^r) = 0$

ただし，$L_f(X^n)$ は，系列 X^n を f_n によって写像して得られる列の長さを，また $D(\cdot \parallel \cdot)$ はダイバージェンスを表す．$f_{n,r}(X^n)$ は，入力列を $\{x_1^n : L_f(x^n) = r\}$ に制限したとき符号 f_n によって得られる各々の符号語の生起確率を，B^r は $\{0,1\}^r$ 上の一様分布をそれぞれ表す．

条件 (i) は G_n に属さない符号長が現れないこと，条件 (ii) は入力 1 記号当りの出力乱数列の平均長がレート R になること，条件 (iii) は得られた乱数列が一様分布に従うことを意味している．このとき，次の定理が成り立つ．

［定理 7］（Visweswariah-Kulkarni-Verdú [17]） もし，定常エルゴード情報源 X と任意の $\delta > 0$ に対し，語頭符号 $f_n : \mathcal{X}^n \to \{0,1\}^*$ が，

$$\lim_{n\to\infty} \Pr(|L_f(X^n) - H(X^n)| \leqq H(X^n)\delta) = 1 \tag{13}$$

を満足するならば，語頭符号 f_n はレート $H(X)$ の乱数生成器である．

Visweswariah らは，Shannon 符号，Huffman 符号及び LZ78 符号が条件 (13) を満足することを示した [17]．

他方，次の定理は，情報源 X から取り出せる乱数のレートが $H(X)$ で制限されることを表している．

［定理 8］（Visweswariah-Kulkarni-Verdú [17]） 定常エルゴード情報源 X に対し，任意の乱数生成器のレート R は常に $R \leqq H(X)$ を満足する．

これらの定理から，LZ78 符号は，入力される定常エルゴード情報源に依存せず，漸近的に最大のレートを達成するユニバーサルな乱数生成器であることが結論される．

この成果とは別に，韓は，語頭符号が，規格化された条件付きダイバージェンスを零にするという意味での乱数生成になっていることを示した．

［定理 9］（韓 [18, 定理 2.19]） $\mathcal{X}$ を有限あるいは可算無限アルファベットとする．もし，定常エルゴード情報源 X に対し，語頭符号 $f_n : \mathcal{X}^n \to \{0,1\}^*$ が，

$$\lim_{n\to\infty} \frac{1}{n} E[L_f(X^n)] = H(X) \tag{14}$$

を満足するならば，

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{\infty} \Pr(I_n = r) D(f_{n,r}(X^n) \parallel B^r) = 0 \tag{15}$$

が成り立つ．ただし，I_n は $L_f(x^n) = m$ ならば $I_n = m$ なる確率変数である．

$\mathcal{X}$ が有限アルファベットならば LZ78 符号は明らかに式 (14) を満足するので，LZ78 符号は式 (15) の意味でのユニバーサルな乱数生成器となっている．

4.2 仮説検定への応用

$\mathcal{X}$ 上の二つの情報源 X_1 と X_2 があり，ある系列 $x_1^n \in \mathcal{X}^n$ が与えられたとき，この系列が二つの情報源のどちらから出力されたものかを判定するのが仮説検定である [18]．実際の仮説検定では，受容域と呼ばれる $\mathcal{X}^n$ の部分集合 A_n を定め，与えられた系列 x_1^n が $x_1^n \in A_n$ を満足するときは情報源 X_1 から出力されたと判定し，そうでないときは情報源 X_2 から出力されたとする．受容域は，判別関数 $h : \mathcal{X}^n \to R$ を用いて，$A_n = \{x_1^n \in \mathcal{X}^n : h(x_1^n) > 0\}$ として常に書くことができる．

さて，情報源 X_1 及び X_2 が，それぞれ分布 P_1 及び P_2 に従う定常エルゴード情報源であるとき，

$$P_2(A_n) \leqq 2^{-\lambda n}$$

という条件下で $P_1(A_n)$ を最大にする問題を考えよう．この問題は，X_2 から系列 x_1^n が出力されたとき，X_1 から出力されたと誤って判定する確率を $2^{-\lambda n}$ 以下に抑えたまま，X_1 から x_1^n が出力されたときに，正しく判定する確率を最大にすることを意味する．仮説検定について，最も一般的な成果が次の定理である．

［定理 10］（韓 [18, 定理 4.1]） $\mathcal{X}$ を有限あるいは可算無限アルファベットとする．任意の $\epsilon > 0$ に対し，ある受容域 A_n が存在して，$P_2(A_n) \leqq 2^{-\lambda n}$ かつ

$$\lim_{n\to\infty} P_1(A_n) \geqq 1 - \epsilon$$

が成り立つための必要十分条件は，$\underline{D}(P_1 \parallel P_2) > \lambda$ である．ただし，

$$\underline{D}(P_1 \parallel P_2) \triangleq \text{p-}\liminf_{n\to\infty} \frac{1}{n} \sum_{x_1^n \in \mathcal{X}^n} P_1(x_1^n) \log \frac{P_1(x_1^n)}{P_2(x_1^n)}$$

であり，p-liminf は確率的下極限 [18] を表す．また，そのような受容域は，判別関数

$$h(x_1^n) \stackrel{\triangle}{=} \frac{1}{n}\log\frac{P_1(x_1^n)}{P_2(x_1^n)} - \lambda$$

によって定められる．

さて，Ziv は，分布 P_2 が既知であるが，分布 P_1 が未知の場合の仮説検定を考え，次の定理を得た．

［定理 11］（Ziv [19]） 判別関数を

$$h'(x_1^n, \lambda) \stackrel{\triangle}{=} \frac{1}{n}[-\log P_2(x_1^n) - L_{LZ78}(x_1^n)] - \lambda$$

によって定めれば，受容域 $A'_n \stackrel{\triangle}{=} \{x_1^n \in \mathcal{X}^n : h'(x_1^n, \lambda) > 0\}$ は，分布 P_1 に依存しない．このとき，与えられた $\lambda > 0$ について，

$$\underline{D}(P_1 \parallel P_2) > \lambda$$

を満足するならば，任意の $\epsilon > 0$ について $P_2(A'_n) \leqq 2^{-n\lambda}$ かつ

$$\lim_{n\to\infty} P_1(A'_n) \geqq 1 - \epsilon$$

が成り立つ．

（注意） Ziv は，P_1 と P_2 が fading memory condition [19, 式 (1)] を満足する定常エルゴード情報源について定理 11 を証明したが，定理 10 を利用すれば，そのような条件は不要である．

更に，Ziv は，分布 P_1 及び P_2 がともに未知だが，X_2 からのトレーニング系列が利用できる場合についてもユニバーサル符号を用いて仮説検定が行えることを示しているが，やや複雑な定義が必要なので割愛する．興味のある方は文献 [19] を参照されたい．

4.3 通信路符号化への応用

Ziv は，ユニフィラーな有限状態通信路において，通信路の統計的性質に依存せず，しかもランダム符号化による信頼性関数を達成するユニバーサルな復号法を増分分解に基づいて構成した [20]．更に，Lapidoth-Ziv は，上記のユニバーサル復号法が必ずしもユニフィラーではない有限状態通信路についても適用できることを明らかにした [21]．本節では，これらのユニバーサル復号法について述べる．

まず，有限状態通信路を定義する．

［定義 2］（有限状態通信路） $\mathcal{X}$ と $\mathcal{Y}$ をそれぞれ有限の入出力アルファベット，S を状態を表す有限集合とする．このとき，有限状態通信路は，遷移確率

$$W(y, s'|x, s) \quad y \in \mathcal{Y},\ x \in \mathcal{X},\ s, s' \in S$$

によって完全に定められる．特に，初期状態 $s_0 \in S$ が与えられたとき，通信路に $x_1^n \in \mathcal{X}^n$ を入力して $y_1^n \in \mathcal{Y}^n$ が出力される確率は，

$$W(y_1^n|x_1^n, s_0) = \sum_{s_1^n \in S^n} \prod_{i=1}^{n} W(y_i, s_i|x_i, s_{i-1})$$

によって与えられる．

Ziv の復号法は次のように述べられる [20]．今，符号語集合 $C(\subset \mathcal{X}^n)$ に属する符号語 x_1^n と受信語 $y_1^n \in \mathcal{Y}^n$ が与えられたとき，$\{(x_i, y_i)\}_{i=1}^n$ を $\mathcal{X} \times \mathcal{Y}$ 上の系列と考えて増分分解する．例えば，$\boldsymbol{x} = 010001$, $\boldsymbol{y} = 010101$ は，

$$\boldsymbol{x} = 0|1|00|01$$
$$\boldsymbol{y} = 0|1|01|01$$

と分解される．さて，$c(y_1^n)$ によって y_1^n の部分列の総数を表し，$y(l)$ によって y_1^n の相異なる部分列で l 番目に出現するものを表す．更に，$c_l(x_1^n|y_1^n)$ によって $y(l)$ の出現回数を表す．先の例では，$y(1) = 0$，$y(2) = 1$，$y(3) = 01$ であり，$c_1(x_1^n|y_1^n) = 1$，$c_2(x_1^n|y_1^n) = 1$，$c_3(x_1^n|y_1^n) = 2$ である．このとき，復号関数を

$$u(x_1^n, y_1^n) \stackrel{\triangle}{=} \frac{1}{n}\sum_{l=1}^{c(y_1^n)} c_l(x_1^n|y_1^n)\log c_l(x_1^n|y_1^n)$$

によって定義し，受信語 y_1^n を $u(x_1^n, y_1^n)$ を最小にする符号語 $x_1^n \in C$ に復号する．この復号法が通信路に依存しないのは明らかである．

Ziv の復号法について，次の定理が成り立つ．

［定理 12］（Lapidoth-Ziv [21]） W を有限状態通信路とする．$\overline{P}^n_{W,ML}(\text{error})$ によって，復号法として最ゆう復号法を用い，$\mathcal{X}$ 上の一様分布に基づくランダム符号化によって符号語を定めた際の平均復号誤り率を表す．また，$\overline{P}^n_{W,Z}(\text{error})$ によって，Ziv の復号法を用いた際，同様のランダム符号化による平均復号誤り率を表す．このとき，すべての W について

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{\overline{P}^n_{W,Z}(\text{error})}{\overline{P}^n_{W,ML}(\text{error})} = 0$$

が成り立つ．更に，符号長 n の符号語集合の系列 $\{C_n\}$ が存在して，すべての W について

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{P^n_{W,Z}(\text{error}|C_n)}{\overline{P}^n_{W,ML}(\text{error})} = 0$$

が成り立つ．ただし，$P^n_{W,Z}(\text{error}|C_n)$ は，符号 C_n を Ziv の復号法で復号したときの平均復号誤り率を表す．

この定理は，ランダム符号化の信頼性関数が通信路に依存しないユニバーサル復号法によっても達成できることを示すとともに，実際にユニバーサルな通信路符号の存在を示している．

5. むすび

本論文では，LZ78 符号の漸近的最良性及び平均冗長度の上界を明らかにするとともに，LZ 符号が各種の情報理論の問題の解法に有効であることを明らかにした．紙面の都合で述べられなかったユニバーサル符号の応用として，補助情報源を伴う情報源のユニバーサル符号化 [22]，ひずみを許したユニバーサル符号化 [23], [24]，ユニバーサルギャンブリング [25]，ユニバーサル予測 [26]，通信路の遅延推定 [27] などがあり，これらについては，いずれ機会を見つけて解説したい．

謝辞　このような論文執筆の機会をお与え頂いた特集号編集委員会，並びに，この論文のもとになる招待講演を催して頂いた情報理論研究専門委員会に深謝する．また，文献 [7] の成果の公表を快諾して頂いた J. C. Kieffer 教授並びに E.-h. Yang 教授に感謝する．

文　献

[1] B.M. Fitingof, “Optimal coding in the case of unknown and changing message statistics,” Probl. Inform. Transm., vol.2, no.2, pp.3–11, 1966.

[2] L.D. Davisson, “Universal noiseless coding,” IEEE Trans. Inf. Theory, vol.IT-19, no.6, pp.783–795, Nov. 1973.

[3] T.J. Lynch, “Sequence time coding for data compression,” Proc. IEEE, vol.54, pp.1490–1491, Oct. 1966.

[4] L.D. Davisson, “Comments on ‘Sequence time coding for data compression’,” Proc. IEEE, vol.54, p.2010, Dec. 1966.

[5] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” IEEE Trans. Inf. Theory, vol.IT-27, pp.230–237, March 1981.

[6] J. Ziv and A. Lempel, “A universal algorithm for data compression,” IEEE Trans. Inf. Theory, vol.IT-23, pp.337–343, 1977.

[7] J.C. Kieffer and E.-H. Yang, “A simple technique for bounding the pointwise redundancy of the 1978 Lempel-Ziv algorithm,” Proc. DCC 1999, pp.434–441, 1999.

[8] 韓　太舜，小林欣吾，情報と符号化の数理，培風館，1999.

[9] T.M. Cover and J.A. Thomas, Elements of information theory, John Wiley, 1991.

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

[11] A. Baron, Logically smooth density estimation, Ph.D. Thesis, Dept. of Elec. Eng., Stanford Univ., 1985.

[12] E. Plotnik, J. Weinberger, and J. Ziv, “Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm,” IEEE Trans. Inf. Theory, vol.38, pp.66–72, 1992.

[13] Y.M. Shtarkov and T.J. Tjarkens, “The redundancy of the Ziv-Lempel algorithm for memoryless sources,” ITW ’90, Einthoven, the Netherlands, 1990.

[14] G. Louchard and W. Szpankowski, “On the average redundancy rate of the Lempel-Ziv code,” IEEE Trans. Inf. Theory, vol.43, pp.2–8, 1997.

[15] P. Shields, “Universal redundancy rates do not exist,” IEEE Trans. Inf. Theory, vol.39, pp.520–524, 1993.

[16] S.A. Savari, “Redundancy of the Lempel-Ziv incremental parsing rule,” IEEE Trans. Inf. Theory, vol.43, pp.9–21, 1997.

[17] K. Visweswariah, S.R. Kulkarni, and S. Verdú, “Source codes as random number generation,” IEEE Trans. Inf. Theory, vol.44, pp.462–471, March 1998.

[18] 韓　太舜，情報理論における情報スペクトル的方法，培風館，1998.

[19] J. Ziv, “On classification with empirically observed statistics and universal data compression,” IEEE Trans. Inf. Theory, vol.34, pp.278–286, 1988.

[20] J. Ziv, “Universal decoding for finite-state channels,” IEEE Trans. Inf. Theory, vol.IT-31, pp.453–460, July 1985.

[21] A. Lapidoth and J. Ziv, “On the universality of the LZ-based decoding algorithm,” IEEE Trans. Inf. Theory, vol.44, pp.1746–1755, Sept. 1998.

[22] 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.

[23] J. Muramatsu and F. Kanaya, “Distortion-complexity and rate-distortion function,” IEICE Trans. Fundamentals, vol.E77-A, no.8, pp.1224–1229, Aug. 1994.

[24] E.-h. Yang and C. Kieffer, “Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm,” IEEE Trans. Inf. Theory, vol.42, pp.239–245, Jan. 1996.

[25] M. Feder, “Gambling using a finite state machine,” IEEE Trans. Inf. Theory, vol.37, pp.1459–1465, Sept. 1991.

[26] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” IEEE Trans. Inf. Theory, vol.38, no.4, pp.1258–1270, July 1992.

[27] J. Stein, J. Ziv, and N. Merhav, “Universal delay estimation for discrete channels,” IEEE Trans. Inf.

Theory, vol.42, no.6, pp.2085–2093, Nov. 1996.
（平成 12 年 11 月 17 日受付，13 年 1 月 19 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，平 4 年度並びに平 7 年度本会論文賞各受賞．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」,「情報通信ネットワーク」など．IEEE，情報理論とその応用学会各会員．