テクニカルノート

$\log^* n$ より短い符号語の割合が 1 に漸近する正整数符号について

中村博文† 村島定行††

先に，正整数の従来符号の構成要素の MSB 列を，単進符号と CBT 符号を組み合わせた符号で置換して，符号語長の上界の伸びを緩和した符号を作成する方法 U_m を提案した．また U_m を Elias の符号 ω，Bentley&Yao の探索木をもとにした符号 ω^*，Stout の符号 R_l と S_l，Even&Rodeh の符号 C_{ER} の各従来符号に適用した符号作成例を示した．本テクニカルノートではそれら作成符号が，符号化する正整数 n を大きくしていくとき $\log^* n$ より短い符号語の割合が 1 に近づくという性質を有することを示す．ここで $\log^* n = \log_2 n + \log_2 \log_2 n + \log_2 \log_2 \log_2 n + \cdots$ である．

On Positive Integer Codes that Ratios of Codewords Shorter than $\log^* n$ Tend to One

Hirofumi Nakamura† and Sadayuki Murashima††

A code construction method U_m has been proposed. It encodes the MSBs string of recursively prefixed length information of original codewords by the unary code and the CBT code. As examples, U_m was applied to the code of Elias, the code based on the search tree of Bentley&Yao, the codes of Stout and the code of Even&Rodeh. Following fact is shown in this paper: these codes made by U_m have the property that the codeword lengths of positive integers (assume n be one of them) shorter than $\log^* n$ tends to 1 in probability, where $\log^* n = \log_2 n + \log_2 \log_2 n + \log_2 \log_2 \log_2 n + \cdots$.

1. はじめに

正整数 n を，n の 2 進数表現と，これに再帰的に前置した桁数情報の 2 進数表現（構成要素と呼ぶ）の並びを用いて表す符号化法が，多数提案されている[1)~4)]．また筆者らは，正整数の従来符号の構成要素の MSB 列を，単進符号と CBT（完全 2 分木）符号を組み合わせた符号で置換することによって，n に対する符号語長の上界の増加を緩和した符号を作成する方法 U_m $(m \geq 2)$ を提案した[5)]．そしてこの U_m を，Elias[1)] の符号 ω，Bentley&Yao[2)] の探索木をもとにした符号 ω^*，Stout[3)] の符号 R_l と S_l，Even&Rodeh[4)] の符号（文献 6) にならって C_{ER} と表す．R_3 と等しい）の各従来符号に適用した符号作成例（$U_{m\omega}$，$U_{m\omega^*}$，U_{mR_l}，U_{mS_l}，$U_{mC_{ER}}$）を示した．

以下で，従来符号 ω，ω^*，R_l，S_l，C_{ER} のいずれにもあてはまる議論ではこれらの従来符号を C と表すことにする．任意の符号 X による n の符号語を $X(n)$ のように，符号語長を $|X(n)|$ のように表す．$C(n)$ の構成要素の数を $F_C(n)$，C に U_m を適用して作成した符号を U_{mC} と表す．$\log_2$ を log，log の i 回の合成関数を $\log^i$，$\log n + \log^2 n + \log^3 n + \cdots$ を $\log^* n$ と表す．$\log^*$ は対数スター関数と呼ばれ，0 以上の項のみ足し合わせる．

山本[7),8)] によって符号 $C_{a,f}$ が提案され，その $C_{a,f}$ について，正整数 n を大きくしていくとき $\log^* n$ より短い符号語の割合が 1 に近づくという性質のあることが示されている．

本テクニカルノートでは，U_{mC} もこの性質を有することを示す．

2. $C(n)$ と $U_{mC}(n)$ について

Elias の符号 ω による正整数 n の符号語 $\omega(n)$ は

$$\omega(n) = \omega_B(n)0.$$

$$\omega_B(n) = \begin{cases} \lambda & (n = 1) \\ \omega_B(\lfloor \log n \rfloor) B(n, \lfloor \log n \rfloor + 1). & (n \geq 2) \end{cases} \quad (1)$$

† 都城工業高等専門学校機械工学科（電子計算機センター）
Department of Mechanical Engineering (Computer Center), Miyakonojo National College of Technology
†† 鹿児島大学工学部情報工学科
Department of Information and Computer Science, Faculty of Engineering, Kagoshima University

と表せる．ここで，λ は空列を，$B(i,j)$ は非負整数 i の j ビットでの 2 進数表現を表す．ピリオドは符号語の構造が分かりやすいように文献 9) に似せて付加したものであり，実際の符号語には不要である．たとえば，$\omega(2^{16}) = 10.100.10000.10^{16}.0.$，$F_\omega(2^{16}) = 5$ である．ここで用いているように，ビット 0 と 1 については繰り返し i 回現れるときそれぞれ 0^i，1^i と表すことがある．

$U_{mC}(n)$ は，$C(n)$ の構成要素の MSB 列を，単進符号 u と横尾の CBT 符号[10]（CBT と表す）を組み合わせた符号 u_m の符号語で置き換えた構造である．

J 個の可能性の中から $0 \le j \le J-1$ なる j を表現する符号 CBT の符号語 $CBT(j,J)$ を

$$CBT(j,J) = \begin{cases} B(j, \lfloor \log J \rfloor) \\ \quad (0 \le j \le 2^{\lceil \log(J+1) \rceil} - J - 1) \\ B(2^{\lceil \log(J+1) \rceil} - J + j, \lfloor \log J \rfloor + 1) \\ \quad (2^{\lceil \log(J+1) \rceil} - J \le j \le J-1) \end{cases}$$

とし[10]，$u(i)$ $(i \ge 1)$ を

$$u(i) = 1^{i-1}0$$

とするとき，$u_m(i)$ $(i \ge 1)$ を次のように定義した[5]．

$$u_m(i) = u\Big(\Big\lfloor \frac{i-1}{m} \Big\rfloor + 1\Big).CBT((i-1) \mod m, m).$$

たとえば，$u_2(1) = 0.0.$，$u_2(2) = 0.1.$，$u_2(3) = 10.0.$，$u_2(4) = 10.1.$，$u_2(5) = 110.0.$，$u_2(6) = 110.1.$ である．

ω に U_m を適用した $U_{m\omega}$ の符号語 $U_{m\omega}(n)$ は

$$U_{m\omega}(n) = u_m(F_\omega(n)) : U_{m\omega B}(n). \quad (n \ge 1)$$

$$U_{m\omega B}(n) = \begin{cases} \lambda & (n = 1) \\ U_{m\omega B}(\lfloor \log n \rfloor) B'(n, \lfloor \log n \rfloor). & (n \ge 2) \end{cases}$$

である[5]．ここで，$B'(i, \lfloor \log i \rfloor)$ は i の 2 進数表現の MSB を除いた $\lfloor \log i \rfloor$ ビットを表す．ピリオドは式 (1) での表記と対応するように挿入した．コロンも符号語の構造が分かりやすいように付加したものであり，実際の符号語には不要である．たとえば，$U_{2\omega}(2^{16}) = 110.0. : 0.00.0000.0^{16}..$ である．

他の C や U_{mC} の定義の記述は紙面の都合で参考文献に譲る．

3. $|U_{mC}(n)| < \log^* n$ を満たす n の割合について

構成要素数が同じになる正整数ごとに $U_{mC}(n)$ が $\log^* n$ ビットより短くなる割合を求めていく．文献 7), 8), 11) の議論の進行を参考にした．本テクニカルノートの解析手法は，特に文献 11) の 3.2 節とほぼ同じである．

$C(n)$ の各構成要素には $C(n)$ の語尾から語頭に向けて 0 で始まる番号が振ってあるものと考える．特定の構成要素を指し示すときに用いる．

C で符号化したときに，$F_C(n)$ が K であり，かつ，$C(n)$ の構成要素のうちで 0 番目（符号語の終わりを表す 0）〜2 番目以外の構成要素のビットがすべて 1 となる（符号語の体裁は，$1\cdots1.$ $\ldots$ $.1\cdots1.$ $1\{0|1\}\cdots\{0|1\}.$ $1\{0|1\}\cdots\{0|1\}.0.$ である．ここで，$\{0|1\}$ はビットが 0 または 1 であることを表す）ような正整数の集合を $S_C(K)$ と表す．$S_C(K)$ は $K \ge 4$（ω^* では $K \ge 5$）で定義する．$S_C(K)$ の要素数を $\#(S_C(K))$ と表す．本テクニカルノートでは最終的に K が大きい場合を論じるので，K が小さい場合については言及しない．

3.1 $\#(S_C(K))$ について

$F_C(n)$ が K となる最小の n の値を $E_C(K)$ と表す．

$n \in S_C(K)$，かつ，$C(n)$ の 2 番目の構成要素が $1\cdots1$（すべてのビットが 1）となる（符号語の体裁は，$1\cdots1.$ $\ldots$ $.1\cdots1.11\cdots1.1\{0|1\}\cdots\{0|1\}.0.$）ような正整数 n は，$E_C(K+1)/2 \le n \le E_C(K+1)-1$ の範囲内の値であり，該当する正整数の個数は $E_C(K+1)/2$ 個である．同様に，$C(n)$ の 2 番目の構成要素が $1\cdots10$（最後のビットだけが 0）になる（符号語の体裁は，$1\cdots1.$ $\ldots$ $.1\cdots1.11\cdots10.1\{0|1\}\cdots\{0|1\}.0.$）ような正整数の個数は $E_C(K+1)/4$ 個である．一般に，$C(n)$ $(n \in S_C(K))$ の 2 番目の構成要素を 2 進数と見なした値が $E_C(K) - i$ になる正整数の個数は $E_C(K+1)/2^i$ 個である．

$Q(j)$ $(j \ge 1)$ を

$$Q(j) = \sum_{i=1}^{j} \frac{1}{2^i}$$

と定義するとき，$\#(S_C(K))$ は

$$\begin{aligned} \#(S_C(K)) &= \sum_{i=1}^{\frac{1}{2}E_C(K)} \frac{1}{2^i} E_C(K+1) \\ &= Q\left(\frac{1}{2}E_C(K)\right) E_C(K+1) \quad (2) \end{aligned}$$

と表せる．

3.2 $|C(n)|$ $(n \in S_C(K))$ について

$C(n)$ $(n \in S_C(K))$ は，0〜2 番目の構成要素を除き，各構成要素のビットがすべて 1 である．$F_C(n)$ が K であるような最大の n を $E'_C(K)$ と表すとき，$C(E'_C(K))$ は語尾以外の各構成要素のビットがすべ

表 1 従来符号の符号語の例
Table 1 Examples of codewords of known codes.

Notation	Bit String
$\omega(E_\omega(5)) = \omega(2^{16})$	$10.100.10000.10^{16}.0.$
$\omega(E'_\omega(4)) = \omega(2^{16}-1)$	$11.1111.1^{16}.0.$
$\omega(E'_\omega(2)) = \omega(3)$	11.0.
$\omega^*(E_{\omega^*}(6)) = \omega^*(2^{127})$	$1.100.1000.10000000.10^{127}.0.$
$\omega^*(E'_{\omega^*}(5)) = \omega^*(2^{127}-1)$	$1.111.1111111.1^{127}.0.$
$\omega^*(E'_{\omega^*}(3)) = \omega^*(7)$	1.111.0.
$R_3(E_{R_3}(5)) = R_3(2^{127})$	$100.1000.10000000.10^{127}.0.$
$R_3(E'_{R_3}(4)) = R_3(2^{127}-1)$	$111.1111111.1^{127}.0.$
$R_3(E'_{R_3}(2)) = R_3(7)$	111.0.
$S_2(E_{S_2}(5)) = S_2(2^{66})$	$00.100.1000000.10^{66}.0.$
$S_2(E'_{S_2}(4)) = S_2(2^{66}-1)$	$11.111111.1^{66}.0.$
$S_2(E'_{S_2}(2)) = S_2(3)$	11.0.

$^*R_3 = C_{ER}$

て 1 である．そこで，まず，$|C(E'_C(K))|$ の上界を明らかにしておく．

$E'_C(K)$ と $E_C(K+1)$ とは

$$E'_C(K) = E_C(K+1) - 1 \tag{3}$$

の関係にある．

$K \geq 2$（ω^* では $K \geq 3$）では，$C(E_C(K+1))$ の語頭と語尾以外（ω^* ではさらに語頭の直後の構成要素も除く）の全構成要素に対して，1 ビット短い構成要素を $C(E'_C(K))$ 中に必ず見出すことができる（**表 1** に例を示す）．よって

$$|C(E'_C(K))| \leq |C(E_C(K+1))| - F_C(E_C(K+1)) + 3$$

である．文献 5) の式 (5) で，一般に正整数 n について

$$|C(n)| \leq \log^* n + F_C(n) + O(1)$$

であることが示されている．これを用いて，

$$|C(E'_C(K))| \leq \log^*(E_C(K+1)) + O(1)$$

である．また一般に $\log^k(n+1) - \log^k n$ は k の増加にともなって等比数列よりも早く 0 に近づくので，$\log^*(n+1) - \log^* n$ は定数で抑えられ，$E'_C(K)$ と $E_C(K+1)$（式 (3) 参照）について

$$\log^* E_C(K+1) - \log^* E'_C(K) \leq O(1)$$

である．よって，$K \geq 2$（ω^* では $K \geq 3$）で，

$$|C(E'_C(K))| \leq \log^*(E'_C(K)) + O(1) \tag{4}$$

が成立する．続いて，この結果を用いて $C(n)$ $(n \in S_C(K))$ の上界を求める．

$C(n)$ $(n \in S_C(K))$ から 1 番目と 2 番目の構成要素を除いたビット列は $C(E'_C(K-2))$ と同じである（表 1 に例を示す）．よって，$n(\in S_C(K))$ について

$$\begin{aligned}|C(n)| &= |C(E'_C(K-2))| \\ &\quad + |B(\lfloor \log n \rfloor + a, \lfloor \log(\lfloor \log n \rfloor + a) \rfloor + 1)| \\ &\quad + |B(n, \lfloor \log n \rfloor + 1)| \\ &= |C(E'_C(K-2))| \\ &\quad + \lfloor \log(\lfloor \log n \rfloor + a) \rfloor + \lfloor \log n \rfloor + 2\end{aligned}$$

である．ここで，a は，ω では 0，ω^* と R_l では 1，S_l $(l \geq 2)$ では $-l$ である．この式に関して

$$\lfloor \log(\lfloor \log n \rfloor + a) \rfloor \leq \log\log n + O(1)$$
$$\lfloor \log n \rfloor \leq \log n + O(1)$$

が成り立つ．これらと式 (4) を用いて，n $(\in S_C(K))$ について

$$|C(n)| \leq \log^* E'_C(K-2) + \log\log n + \log n + O(1)$$

である．$\log^k E'_C(K-2)$ $(= \log^k(\lfloor \log(\lfloor \log n \rfloor + a) \rfloor + a))$ と $\log^k(\log\log n)$ $(= \log^{k+2} n)$ との差は，k の増加にともなって等比数列よりも早く 0 に近づく．よって，n $(\in S_C(K))$ について

$$|C(n)| \leq \log^* n + O(1) \tag{5}$$

が成立する．

3.3 $|U_{mC}(n)| < \log^* n$ となる条件について

$U_{mC}(n)$ は，$C(n)$ の $F_C(n)$ 個あるいは $F_C(n)-1$ 個の構成要素（この個数を i と表す）の MSB を，$u_m(i)$ で置き換えた形の符号語である．$C(n)$ の i ビットが

$$|u_m(i)| \leq \frac{i}{m} + \log m + O(1)$$

ビット[5] で置き換わる．$n \in S_C(K)$ なる n の符号化では $|C(n)|$ として式 (5) を用いてよく，

$$|U_{mC}(n)| \leq \log^* n - \left(1 - \frac{1}{m}\right) F_C(n) + \log m + O(1)$$

である．これは，有限の適当な定数 A を用いて

$$|U_{mC}(n)| \leq \log^* n - \left(1 - \frac{1}{m}\right) F_C(n) + \log m + A$$

と表せる．U_{mC} $(m \geq 2)$ について，定数 F'_{mC} を

$$F'_{mC} = \frac{\log m + A}{1 - \frac{1}{m}} \tag{6}$$

と定義する．$m \geq 2$ で F'_{mC} は有限の値である．U_{mC} $(m \geq 2)$ について，$n \in S_C(F_C(n))$ かつ $F_C(n) > F'_{mC}$ であるような n では必ず $|U_{mC}(n)| < \log^* n$ である．

3.4 $|U_{mC}(n)| < \log^* n$ の割合の漸近的性質

$F_C(n)$ が K である正整数全体の中で，$U_{mC}(n)$ が $\log^* n$ より短くなる割合を $P_C(K)$ と表す．$S_C(K)$

が定義され（すなわち，$K \geq 4$, ω^* では $K \geq 5$），かつ，$K > F'_{mC}$ では，

$$P_C(K) > \frac{\#(S_C(K))}{E_C(K+1) - E_C(K)}$$

である．式 (2) を用いて

$$P_C(K) > \frac{Q\left(\frac{1}{2}E_C(K)\right)}{1 - \frac{E_C(K)}{E_C(K+1)}}$$

である．この式に関して，

$$\lim_{K\to\infty} Q\left(\frac{1}{2}E_C(K)\right) = 1$$
$$\lim_{K\to\infty} \frac{E_C(K)}{E_C(K+1)} = 0$$

が成立する．また，$P_C(K)$ は 1 より大きくなることはない．よって

$$\lim_{K\to\infty} P_C(K) = 1 \qquad (7)$$

が得られる．

n を大きくしていくとき，式 (7) の意味で，$|U_{mC}(n)| < \log^* n$ である割合は 1 に近づくといえる．すなわち，正整数の従来符号 C に，符号語の構成要素の MSB 列を単進符号と CBT 符号を組み合わせた符号で置換する方法 U_m $(m \geq 2)$ を適用して作成した符号 U_{mC} は，符号化する正整数 n を大きくしていくとき符号語長が $\log^* n$ より短い割合が 1 に近づく．

4. おわりに

正整数の従来符号 ω, ω^*, R_l, S_l, C_{ER} に符号語の構成要素の MSB 列を単進符号と CBT 符号を組み合わせた符号で置換する方法 U_m $(m \geq 2)$ を適用して作成した符号 $U_{m\omega}$, $U_{m\omega^*}$, U_{mR_l}, U_{mS_l}, $U_{mC_{ER}}$ には，山本の符号 $C_{a,f}$ と同様に，符号化する正整数 n を大きくしていくとき符号語長が $\log^* n$ より短い割合が 1 に近づく性質のあることを示した．

参考文献

1) Elias, P.: Universal codeword sets and representations of the integers, *IEEE Trans. Inform. Theory*, Vol.IT-21, No.2, pp.194–203 (1975).
2) Bentley, J.L. and Yao, A.C.: An almost optimal algorithm for unbounded searching, *Inform. Proc. Letter*, Vol.5, No.3, pp.82–87 (1976).
3) Stout, Q.F.: Improved Prefix Encoding of the Natural Numbers, *IEEE Trans. Inform. Theory*, Vol.IT-26, No.5, pp.607–609 (1980).
4) Even, S. and Rodeh, M.: Economical encoding of commas between strings, *Comm. ACM*, Vol.21, No.4, pp.315–317 (1978).
5) 中村博文，村島定行：従来符号の長さ情報の MSB 列を符号化することによる新たな正整数符号の構成法，情報処理学会論文誌，Vol.40, No.4, pp.1745–1753 (1999).
6) 韓　太舜，小林欣吾：整数のユニバーサル符号化，情報と符号化の数理 第 5 章，pp.193–234，培風館 (1999).
7) 山本博資：正整数の新しい再帰的なユニバーサル表現，第 20 回情報理論とその応用シンポジウム予稿集，pp.497–500 (1997).
8) Yamamoto, H.: A new recursive universal code of the positive integers, *IEEE Trans. Inform. Theory*, Vol.46, No.2, pp.717–723 (2000).
9) Bell, T.C., Cleary, J.G. and Witten, I.H.: *Text Compression*, Prentice-Hall, New Jersey (1990).
10) 横尾英俊：ユニヴァーサル情報源符号化のための修正 Ziv-Lempel 符号，信学論（A），Vol.J68-A, No.7, pp.664–671 (1985).
11) 中村博文，村島定行：再帰的に定義される長さ情報の MSB 省略を用いる正整数符号，信学論（A），Vol.J83-A, No.4, pp.432–438 (2000).

(平成 12 年 5 月 8 日受付)
(平成 12 年 7 月 5 日採録)

中村　博文（正会員）

昭和 36 年生．昭和 58 年豊橋技科大学情報学部卒業．昭和 60 年同大学院工学研究科修士課程修了．昭和 60 年（株）日立製作所神奈川工場．昭和 61 年都城工業高等専門学校機械工学科助手，現在同助教授．電子計算機センター配属．情報源符号化の研究に従事．

村島　定行

昭和 17 年生．昭和 40 年九州工業大学電気工学科卒業．昭和 40 年京都工芸繊維大学電気工学科助手，昭和 45 年鹿児島大学電子工学科講師，昭和 49 年同助教授，現在同情報工学科教授．昭和 49 年京都大学工学博士．昭和 55 年情報処理学会創立 20 周年記念論文賞受賞．数値解析，情報理論，データ圧縮，神経回路網の研究に従事．著書「代用電荷法とその応用」（森北出版，昭和 58 年）．