PAPER

A New Class of the Universal Representation for the Positive Integers

Takashi AMEMIYA[†*], *Nonmember and* Hirosuke YAMAMOTO[†], *Member*

SUMMARY A new class of the universal representation for the positive integers is proposed. The positive integers are divided into infinite groups, and each positive integer n is represented by a pair of integers (p, q), which means that n is the q-th number in the p-th group. It is shown that the new class includes the message length strategy as a special case, and the asymptotically optimal representation can easily be realized. Furthermore, a new asymptotically and practically efficient representation scheme is proposed, which preserves the numerical, lexicographical, and length orders.

key words: *universal code, universal representation of the positive integers, message length strategy, grouping strategy*

1. Introduction

Elias[(3)] treated the coding problem of the positive integers that satisfy

$$P(n) \geqq P(n+1), \text{ for any } n \in \mathcal{N}, \tag{1}$$

where $P(n)$ is a probability distribution on the set of the positive integers $\mathcal{N} \triangleq \{1, 2, 3, \cdots\}$. Such codes can be used in various universal data compression algorithms.[(1),(2)] He introduced the notions of universality and asymptotic optimality to evaluate the performance of such codes. Let $L_\alpha(n)$ be codeword-length of n under a representation scheme α. Then the average codeword length is given by

$$E[L_\alpha(n)] = \sum_{n=1}^{\infty} P(n) L_\alpha(n). \tag{2}$$

A representation α is called *universal* if there exists a constant ρ_α, independent of $P(n)$, such that

$$\frac{E[L_\alpha(n)]}{\max\{1, H(P)\}} \leqq \rho_\alpha, \tag{3}$$

where $H(P)$ is the entropy of $P(n)$. α is called *asymptotically optimal* if the ratio of the left side of Eq.(3) tends to 1 asymptotically as follows.

$$\frac{E[L_\alpha(n)]}{\max\{1, H(P)\}} \leqq R(H(P)), \tag{4}$$

where $R(x)$ is a function such that $R(x) \to 1$ as $x \to \infty$.

Furthermore, Wang[(15)] introduced an intermediate notion between the universality and the asymptotic optimality for the case in which a representation has a parameter. Let $\alpha(t)$ be a representation with parameter t. Then, a family of representations $\alpha(t)$ is called *almost asymptotically optimal* if there exists a function $R_t(x)$ with parameter t such that

$$\frac{E[L_{\alpha(t)}(n)]}{\max\{1, H(P)\}} \leqq R_t(H(P)), \tag{5}$$

where

$$\lim_{t\to\infty} \lim_{x\to\infty} R_t(x) = 1.$$

Although many universal representations of the positive integers have been proposed, they can be divided into two classes, the message length strategy[(3)-(8)] and the flag strategy.[(9)-(18)*] In this paper, we propose another new class called *grouping strategy*.

In the grouping strategy, the positive integers are divided into countable infinite groups, and each positive integer n is represented by a pair of positive integers (p, q), which means that n is the q-th number in the p-th group. We will show that the grouping strategy coincides with the message length strategy if the positive integers are grouped by every 2^{l-1} integers, $l = 1, 2, \cdots$. Hence the grouping strategy is a wide class of the representation of the positive integers that includes the message length strategy as a special case.

The details of the grouping strategy are explained in Sect. 2. In Sect. 3, we will discuss the asymptotic performance of the grouping strategy and will show that an almost asymptotically optimal representation and an asymptotically optimal representation can easily be obtained from the grouping strategy. In Sect. 4, a new almost asymptotically optimal representation is proposed, which is almost as efficient as the best of known flag strategy schemes for large initial segments of the positive integers and satisfies the numerical, lexicographic, and length orders preserving properties.

We use the following notation for simplicity in this paper.

$\lfloor x \rfloor$: The largest integer not greater than x.
$\lceil x \rceil$: The smallest integer not less than x.

Manuscript received April 21, 1992.

[†] The authors are with the Faculty of Electro-Communications, University of Electro-Communications, Chofu-shi, 182 Japan.

[*] Presently, with the Software Department, NEC Corporation, Abiko-shi, 270-11 Japan.

[*] The comparison of these two strategies can be found in Ref.(18).

$B_\alpha(n)$: The codeword of n under the representation α.

$B_\alpha(n|N)$: The codeword of n under the representation α in the case that the maximum number of n, say N, is given.

$L_\alpha(n)$: The codeword length of n under the representation α, i.e. the length of $B_\alpha(n)$.

$[n]_-$: The binary representation of n derived by deleting the most significant bit of the conventional binary representation.

$[n]_p$: The conventional binary representation of n with p bits.
Ex. $[5]_- = 01$, $[4]_- = 00$, $[20]_- = 0100$, $[5]_4 = 0101$, $[2]_5 = 00010$.

2. Grouping Strategy

We first briefly review the message length strategy to clarify the relation between the message strategy and the grouping strategy.

In the message length strategy, the codeword $B_m(n)$ consists of two parts, *Suffix* $[n]_-$ and *Prefix* $B_\alpha(\cdot)$, which represent the value of n and the length of the suffix, respectively, as follows.

$$B_m(n) = \underbrace{B_\alpha(\lfloor \log_2 n \rfloor + 1)}_{\text{Prefix}} \quad \underbrace{[n]_-}_{\text{Suffix}}, \tag{6}$$

where α is an arbitrary representation of the positive integers.† $[n]_-$ is used in the suffix part instead of n because the most significant bit of the conventional binary representation is always one and unnecessary. The prefix part is required because $[n]_-$ does not satisfy the prefix condition.

By devising the representation α for the prefix part, we can construct many representation schemes of the positive integers. For instance, if the unary code $B_U(n)$ (which we call U scheme) is used in the prefix part, we obtain the *Single Prefix scheme* (*SP* scheme);(3)

$$B_{SP}(n) = B_U(\lfloor \log_2 n \rfloor + 1)[n]_-, \tag{7}$$

$$B_U(n) = 0^{n-1}1 \text{ or } 1^{n-1}0, \tag{8}$$

where 0^{n-1} and 1^{n-1} mean the strings of $n-1$ zeros and $n-1$ ones, respectively. If the *SP* scheme is used in the prefix part of Eq. (6), we obtain the *Double Prefix scheme* (*DP* scheme),(3)

$$\begin{aligned} B_{DP}(n) &= B_{SP}(\lfloor \log_2 n \rfloor + 1)[n]_- \\ &= B_U(\lfloor \log_2(\log_2 n + 1) \rfloor + 1)[\lfloor \log_2 n \rfloor \\ &\quad + 1]_-[n]_-. \end{aligned}$$

Furthermore, by defining the prefix part of Eq. (6) recursively, several recursive representation schemes can be obtained.(3)-(7)

We now propose the grouping strategy. Let $p \in \mathcal{N}$, and let $S(p)$ be a function of p. We first divide the positive integers into countable infinite groups such that the p-th group contains $S(p)$ positive integers. Let $T(p)$ be

$$T(p) = \sum_{j=1}^{p} S(j). \tag{9}$$

Then, the first group is $\{1, \cdots, T(1)\}$, the second group is $\{T(1)+1, \cdots, T(2)\}$, and the p-th group is $\{T(p-1)+1, \cdots, T(p)\}$.

We note that if $n \in \{T(p-1)+1, \cdots, T(p)\}$, then n can be represented by the group number p and its numerical order q in the p-th group, which satisfy

$$T(p-1) < n \leqq T(p), \tag{10}$$

$$q = n - T(p-1). \tag{11}$$

Since the maximum of q is $S(p)$, n can be represented by

$$B_g(n) = B_\alpha(p) B_\beta(q|S(p)), \tag{12}$$

where α and β are arbitrary representations for the positive integers.

As an example, let us consider the case of $S(p) = 2^{p-1}$ and $B_\beta(q|2^{p-1}) = [q-1]_{\lceil \log_2 S(p) \rceil} = [q-1]_{p-1}$. Since $T(p) = 2^p - 1$ in this case, p and q are given from Eqs. (10), (11) as follows

$$p = \lceil \log_2(n+1) \rceil = \lfloor \log_2 n \rfloor + 1,$$

$$q = n - 2^{p-1} + 1 = n - 2^{\lfloor \log_2 n \rfloor} + 1.$$

We note from these relations that $[q-1]_{p-1}$ coincides with $[n]_-$. Hence, in this case, Eq. (12) becomes

$$B_g(n) = B_\alpha(\lfloor \log_2 n \rfloor + 1)[n]_-, \tag{13}$$

which agrees with Eq. (6), i.e. the representation of the message length strategy. Therefore, the message length strategy can be considered as a special case of the grouping strategy.

We can easily prove that $B_g(n)$ defined by Eq. (12) satisfies the following properties.

Properties

1. If $B_\alpha(p)$ and $B(q|S(p))$ in Eq. (12) preserve the lexicographic order for p and q, respectively, then $B_g(n)$ preserves the lexicographic order for n, i.e. $B_g(n)$ precedes $B_g(n')$ lexicographically for any $n < n'$.
2. If $B(q|S(p))$ is a fixed length code and $B_\alpha(p)$ preserves the length order for p, then $B_g(n)$ preserves the length order for n, i.e. $L_g(n) \leqq L_g(n')$ for any $n < n'$.
3. If $B(q|S(p)) = [n]_-$ and $B_\alpha(p)$ preserves the numerical order for p, then $B_g(n)$ preserves the numerical order for n, i.e. $B_g(n)$ is smaller than $B_g(n')$ for any $n < n'$ when they are considered as

† Since α is assumed to be a representation of the positive integers and the length of $[1]_-$ is zero, one is added to the suffix length $\lfloor \log_2 n \rfloor$ in the prefix.

usual binary numbers.
Hence, by devising $B_\alpha(p)$ and $B_\beta(q|S(p))$, we can easily construct a scheme that preserves the lexicographic, numerical, and/or length orders. Since the prefix code preserving the lexicographic order can be used as a search code for the infinite set $\mathcal{N}$,(19) we can construct a universally efficient search code for the case such that the search of $n\in\mathcal{N}$ occurs with the probability satisfying Eq.(1).

It is worth noticing that Willems(20) treated a similar representation scheme to the grouping strategy. But he considered only a finite set of the positive integers $\{1, 2, 3,\cdots, 2^L-1\}$. In his scheme, say $B_W(n|2^L-1)$, if $2^{p-1}\leqq n\leqq 2^p-1$ $(p=1, 2,\cdots, L)$ and $q=n-2^{p-1}+1$, then n is encoded as

$$\begin{aligned}B_W(n|2^L-1)&=[p-1]_{\lceil\log L\rceil}[q-1]_{p-1}\\&=[p-1]_{\lceil\log L\rceil}[n]_{-}.\end{aligned}$$

In other words, he considered only fixed-length codes $[p-1]_{\lceil\log L\rceil}$ and $[q-1]_{p-1}$ as the prefix $B_\alpha(p)$ and the suffix $B_\beta(q|S(p))$, respectively, which causes that his scheme cannot be applied for the countable infinite set of the positive integers. Compared with Willems' scheme, the prefix and the suffix parts, $B_\alpha(p)$ and $B_\beta(q|S(p))$, are generalized to arbitrary representations of the positive integers in the grouping strategy.

3. Asymptotic Performance of the Grouping Strategy

In this section, we consider the asymptotic performance of the grouping strategy. The performance is determined by $S(p)$, $B_\alpha(p)$, and $B_\beta(q|S(p))$. However, since the asymptotic performance closely depends on $S(p)$, we investigate the dependency. To simplify the discussion, we consider the case that $B_\alpha(p)$ and $B_\beta(q|S(p))$ are the unary code $B_U(p)$ and the fixed-length code $[q-1]_{\lceil\log_2 S(p)\rceil}$, respectively. In this case, the grouping strategy is represented by

$$B_g(n)=B_U(p)[q-1]_{\lceil\log_2 S(p)\rceil}, \tag{14}$$

and the codeword length $L_g(n)$ is given by

$$L_g(n)=p+\lceil\log_2 S(p)\rceil. \tag{15}$$

Before going into details, we first introduce the following basic lemmas concerning the universality and the asymptotic optimality.

Lemma 1: (Ref.(12, Lemma 1))

1. If, for all n, $L_\alpha(n)\geqq n^t$, for some constant $t>0$, then the representation α is not universal.
2. If, for all n, $L_\alpha(n)\leqq K_1+K_2\log_2 n$, for some constants K_1 and $K_2\geqq 1$, then the representation α is universal.
3. If, for all n, $L_\alpha(n)\geqq K_1+K_2\log_2 n$, for some constants K_1 and $K_2>1$, then the representation α is not asymptotically optimal.

The second part of Lemma 1 can easily be extended as follows.

Lemma 2: Let $\alpha(t)$ be a representation with parameter t. If, for all n and all t, $L_{\alpha(t)}(n)\leqq K_1(t)+K_2(t)\log_2 n$, for some fixed functions $K_1(t)$ and $K_2(t)$, where $\lim_{t\to\infty} K_2(t)=1$, then the representation $\alpha(t)$ is universal and the family of $\alpha(t)$ is almost asymptotically optimal.

Lemma 3: If, for all n, $L_\alpha(n)\leqq\log_2 n+f(\log_2 n)$, for some monotonically increasing concave positive function $f(x)$ such that $\lim_{x\to\infty}(f(x)/x)=0$, then the representation α is asymptotically optimal.

We consider the performance in the cases that $S(p)$ is constant, linear, exponential, or super exponential function of p, which we call $C(a)$, $LI(b)$, $EX(c)$, $SE(c, m)$ schemes, respectively.

1. Constant case ($C(a)$ scheme)
We first treat the case where $S(p)$ is constant, i.e. $S(p)=a$, $a\in\mathcal{N}$. In this case, $T(p)$ is equal to ap. Applying Eqs.(10), (11), p and q are obtained as

$$p=\left\lceil\frac{n}{a}\right\rceil=\left\lfloor\frac{n-1}{a}\right\rfloor+1, \tag{16}$$

$$q=n-a(p-1). \tag{17}$$

By substituting Eq.(16) into Eq.(15), we can show from Lemma 1 that the $C(a)$ scheme is not universal.

2. Linear order case ($LI(b)$ scheme)
We next consider the case in which $S(p)$ is a linear function of p, i.e. $S(p)=bp$, $b, p\in\mathcal{N}$. Since $T(p)=bp(p+1)/2$ in this case, p is obtained from Eq.(10) as

$$p=\left\lceil\frac{-1+\sqrt{1+\frac{8}{b}n}}{2}\right\rceil. \tag{18}$$

From Eqs.(15), (18) and Lemma 1, the $LI(b)$ scheme is not universal, either. We can extend $S(p)$ to bp^m (or a more general polynomial of fixed degree m). However, we can easily show that the universality cannot be attained even if $S(p)$ is extended to bp^m because the order of p is given by the $(m+1)$th root instead of the square root of n.

3. Exponential order case ($EX(c)$ scheme)
Next, let $S(p)$ be an exponential function of p, i.e. $S(p)=c^{p-1}$ where c is a parameter such that $c\geqq 2$, $c\in\mathcal{N}$. As pointed out in Sect. 2, the scheme with $S(p)=2^{p-1}$, i.e. the $EX(2)$ scheme, coincides with the SP scheme. Hence, $EX(c)$ scheme can be considered as an extension of the SP scheme, which is universal. From $T(p)=(c^p-1)/(c-1)$ and Eq.(10), we get

$$\begin{aligned}p&=\lceil\log_c\{(c-1)n+1\}\rceil\\&\leqq\lceil\log_c cn\rceil\\&<\log_c n+2.\end{aligned} \tag{19}$$

By substituting Eq.(19) into Eq.(15), $L_{EX(c)}(n)$ is bounded as follows

$$L_{EX(c)}(n) \leqq p+(p-1)\log_2 c+1$$
$$<\left(1+\frac{1}{\log_2 c}\right)\log_2 n+3+\log_2 c. \qquad (20)$$

Hence, from Eq. (20) and Lemma 2, the $EX(c)$ scheme is universal and the family of $EX(c)$ schemes is almost asymptotically optimal.

4. Super exponential order case ($SE(c, m)$ scheme) The last case we treat is that $S(p)$ is a higher order function of p than the exponential order, i.e. $c^{(p^m)}$, $m, c \in \mathcal{N}$. If $m=1$, the scheme reduces to the exponential case. Hence, we assume that $m \geqq 2$. For simplicity, let $T(p)=c^{(p^m)}$ and $S(p)=c^{(p^m)}-c^{((p-1)^m)}$. From Eq. (10), we obtain

$$p=\lceil \sqrt[m]{\log_c n}\,\rceil < \sqrt[m]{\log_c n}+1. \qquad (21)$$

By substituting Eq. (21) into Eq. (15), $L_{SE(c,m)}(n)$ is bounded as follows

$$L_{SE(c,m)}(n)=p+\lceil \log_2(c^{(p^m)}-c^{((p-1)^m)})\rceil$$
$$<p+p^m\log_2 c+1$$
$$<\sqrt[m]{\log_c n}+2+(\sqrt[m]{\log_c n}+1)^m\log_2 c$$
$$=\log_2 n+\sqrt[m]{\log_c n}+2$$
$$+\log_2 c\sum_{j=1}^{m}\binom{m}{j}(\sqrt[m]{\log_c n})^{m-j}. \qquad (22)$$

Therefore, from Lemma 3, the $SE(c, m)$ scheme is asymptotically optimal.

We note that SP scheme is not asymptotically optimal, i.e. the asymptotic optimality cannot be attained when the unary code $B_U(\cdot)$ is used as the prefix part $B_\alpha(\cdot)$ of Eq. (6) in the message length strategy. However, the asymptotically optimal representation can be obtained from the grouping strategy even if the unary code $B_U(p)$ is used as the prefix part $B_\alpha(p)$ of Eq. (12).

4. A New Representation for the Positive Integers

In the previous section, we showed that $EX(c)$ and $SE(c, m)$ schemes are almost asymptotically optimal and asymptotically optimal, respectively. However, these asymptotic performances are valid only when the entropy $H(P)$ is huge. In practical cases with a small or moderate entropy, these schemes and known schemes categorized in the message strategy do not attain a good performance compared with, for instance, Capocelli's Fibonacci scheme ($Fib(f)$ scheme where f is a parameter standing for the flag length)(16) or the Yamamoto-Ochi scheme ($YO(f)$ scheme),(18) which are classified in the flag strategy and have good efficiency in large initial segments of the positive integers. In this section, we propose a new representation scheme, which is almost as efficient as the $Fib(f)$ or $YO(f)$ schemes.

Though the message strategy is not efficient, its encoding and decoding algorithms are simple since the suffix part $[n]_-$ can easily be obtained from n. So, in order to use this advantage, we define the suffix of Eq. (12) as $B_\beta(q|S(p))=[n]_-$ by letting $S(p)=2^{p-1}$ and $B_\beta(q|2^{p-1})=[q-1]_{p-1}$. However, instead of the unary code, we use the $C(a)$ scheme as the prefix of Eq. (12) to improve the performance. Furthermore, to simplify the encoding, we define the parameter a of the $C(a)$ scheme as $a=2^k$, where $k \in \mathcal{N}$. Then, this scheme, which we call $CE(k)$ scheme, becomes

$$B_{CE(k)}(n)=B_{C(2^k)}(p)[q-1]_{p-1}$$
$$=B_{C(2^k)}(p)[n]_-. \qquad (23)$$

From $T(p)=2^p-1$ and Eq. (10), p is given as

$$p=\lfloor \log_2 n\rfloor+1. \qquad (24)$$

Furthermore, from Eqs. (16), (17), $B_{C(2^k)}(p)$ can be represented by

$$B_{C(2^k)}(p)=B_U(p')[q'-1]_k, \qquad (25)$$

where

$$p'=\left\lfloor \frac{p-1}{2^k}\right\rfloor+1, \qquad (26)$$
$$q'=p-2^k(p'-1). \qquad (27)$$

From Eqs. (24), (26), (27), we have

$$p'=\left\lfloor \frac{\lfloor \log_2 n\rfloor}{2^k}\right\rfloor+1, \qquad (28)$$
$$q'=\lfloor \log_2 n\rfloor+1-\left\lfloor \frac{\lfloor \log_2 n\rfloor}{2^k}\right\rfloor 2^k$$
$$=[\lfloor \log_2 n\rfloor \bmod 2^k]_k+1. \qquad (29)$$

Consequently, $CE(k)$ scheme is represented as

$$B_{CE(k)}(n)=B_U\left(\left\lfloor \frac{\lfloor \log_2 n\rfloor}{2^k}\right\rfloor+1\right)$$
$$\cdot[\lfloor \log_2 n\rfloor \bmod 2^k]_k[n]_-, \qquad (30)$$

and the codeword length is given by

$$L_{CE(k)}(n)=\left\lfloor \frac{\lfloor \log_2 n\rfloor}{2^k}\right\rfloor+1+k+\lfloor \log_2 n\rfloor$$
$$\leqq\left(1+\frac{1}{2^k}\right)\log_2 n+k+1. \qquad (31)$$

From Eq. (31) and Lemma 2, the family of $CE(k)$ schemes is *almost asymptotically optimal*. Furthermore, from the properties 1~3 shown in Sect. 2, we can easily show that $CE(k)$ scheme preserves *the lexicographic, numerical,* and *length orders.* We note that although the $YO(f)$ and $Fib(f)$ schemes are efficient, they preserve neither the numerical order nor the lexicographic order. Furthermore, the $YO(f)$ scheme does not preserve the length order, either, and the

Table 1 Codeword lengths for $n=2^M$.

M	L_{SP}	L_{DP}	$L_{Fib(3)}$	$L_{Fib(4)}$	$\overline{L_{YO(3)}}$	$\overline{L_{YO(4)}}$	$L_{CE(1)}$	$L_{CE(2)}$	$L_{CE(3)}$
0	1	1	3	4	3.00	4.000	2	3	4
1	3	4	4	5	4.00	5.000	3	4	5
2	5	5	5	6	5.25	6.000	5	5	6
3	7	8	6	7	6.50	7.125	6	6	7
4	9	9	7	8	7.75	8.250	8	8	8
5	11	10	8	9	9.00	9.375	9	9	9
6	13	11	9	10	10.25	10.500	11	10	10
7	15	14	10	11	11.50	11.625	12	11	11
8	17	15	12	12	12.75	12.750	14	13	13
9	19	16	13	13	14.00	13.875	15	14	14
10	21	17	15	14	15.25	15.000	17	15	15
12	25	19	17	16	17.75	17.250	20	18	17
14	29	21	20	19	20.25	19.500	23	20	19
16	33	25	23	21	22.75	21.750	26	23	22
18	37	27	26	23	25.25	24.000	29	25	24
20	41	29	29	25	27.75	26.250	32	28	26
22	45	31	32	28	30.25	28.500	35	30	28
24	49	33	35	30	32.75	30.750	38	33	31
26	53	35	38	32	35.25	33.000	41	35	33
28	57	37	40	35	37.75	35.250	44	38	35
30	61	39	43	37	40.25	37.500	47	40	37

encoding and decoding algorithms of the *Fib*(f) scheme are complex because Fibonacci numbers must be calculated. On the other hand, the encoding and decoding algorithms of the $CE(k)$ scheme is very simple because the division and the modulo operations in Eq. (30) is easily realized by the k bit-shift operation.

We next compare the $CE(k)$ scheme with other schemes about the codeword length. The codeword lengths necessary to represent $n=2^M$ $(0 \leqq M \leqq 30)$ are shown in Table 1, where the flag pattern of the $YO(f)$ scheme is 10^{f-1} and $\overline{L_{YO(f)}(n)}$ is the average codeword length given by

$$\overline{L_{YO(f)}(n)}=\frac{1}{2^M}\sum_{n=2^M}^{2^{M+1}-1} L_{YO(f)}(n) \tag{32}$$

while the codeword lengths of the other schemes are the actual lengths.

We note from Table 1 that $L_{CE(2)}(n)$ is not larger than $L_{SP}(n)$ and $L_{DP}(n)$ for $2^2 \lessapprox n \lessapprox 2^{27}$ and is almost as short as the *Fib*(3) scheme or the $YO(3)$ scheme.

As an example of the probability distribution that satisfies Eq. (1), we consider the geometric probability distribution

$$P_G(n)=(1-\theta)\theta^{n-1}, \quad n \in \mathcal{N}, 0<\theta<1, \tag{33}$$

where the average of n, say $\bar{n}$, is given by

$$\bar{n}=\frac{1}{1-\theta}. \tag{34}$$

The average codeword lengths are compared in Fig. 1.† The $YO(f)$ scheme is omitted in the figure since its performance for the geometric probability distribution is almost the same as the Capocelli's *Fib*(f) scheme.(18) We note from Fig. 1 that the $CE(2)$ scheme is more efficient than the SP and DP schemes for $\bar{n}>8$ and is almost as efficient as the *Fib*(3) scheme.

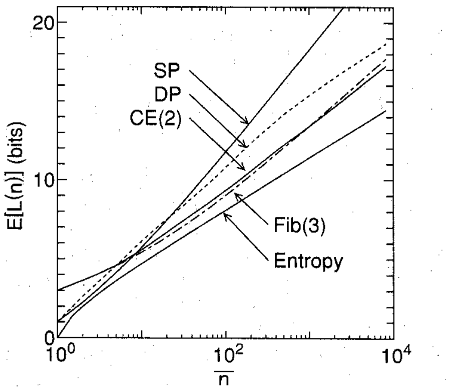


Fig. 1 Expected codeword lengths for the geometrically distributed integers.

5. Conclusion

In this paper, a grouping strategy was proposed to represent the positive integers, which includes the message length strategy as a special case. We showed that the $EX(c)$ and $SE(c, m)$ schemes, which are derived from the grouping strategy, can attain the almost asymptotic optimality and the asymptotic optimality, respectively. We also derived the $CE(k)$ scheme from the grouping strategy, which has a good practical performance in addition to the almost asymptotic optimality. The $CE(k)$ scheme is almost as efficient as the best of known representation schemes. Furthermore, it preserves the numerical, lexicographic, and length orders, and its encoding and decoding algorithms are simple. Therefore, the $CE(k)$ scheme is one of the best representations for the positive integers to use in practical applications.

Various representation schemes can be derived from the grouping strategy by devising $B_\alpha(p)$, $B_\beta(q|S(p))$, and $S(p)$ in Eq. (12). Hence, it may be possible to obtain a better representation scheme from the grouping strategy than the ones we treated in this paper.

References

(1) Storer, J., "Data Compression," Computer Science Press, 1988.

(2) Bell, T., Cleary, J. G. and Witten, I. H., "Text Compression," Prentice Hall, Inc, 1989.

(3) Elias, P., "Universal Codeword Sets and Representations of the Integers," *IEEE Trans. Inform. Theory,* vol. IT-21, no. 2, pp. 194-203, Mar. 1975.

(4) Even, S. and Rodeh, M., "Economical Encoding Commas between Strings," *Commun. ACM,* vol. 21, no. 4, pp. 315

† Although the optimal (but not universal) code is known for the geometric source,(22) its graph is omitted in the figure because it is very close to the entropy.

-317, Apr. 1978.

(5) Knuth, D. E., "Supernatural Numbers," in D. A. klarner (Ed.), *The Mathematical Gardner,* pp. 310-325, Prindle Weber and Schmidt, Boston, 1980.

(6) Stout, Q. F., "Improved Prefix Encodings of the Natural Numbers," *IEEE Trans. Inform. Theory,* vol. IT-26, no. 5, pp. 607-609, Sep. 1980.

(7) Park, J. H., Takashima, Y. and Imai, H., "Adaptive Data Compression Using Self-Organizing Rule," *Trans. IEICE,* vol. J72-A, no. 8, pp. 1353-1359, Aug. 1989.

(8) Yokoo, H., "An Efficient Representation of the Integers for the Distribution of Partial Quotients over the Continued Fractions," *J. of Inform. Processing,* vol. 11, no. 4, pp. 288-293, 1988.

(9) Shannon, C. E., "The Mathematical Theory of Communication," *Bell Syst. Techn. J.,* vol. 27, no. 3, p. 405, Jul. 1948.

(10) Guibas, L. J. and Odlyzko, A. M., "Maximum Prefix-synchronized Codes," *SIAM J. Appl. Math.,* vol. 35, no. 2, pp. 401-408, Sep. 1978.

(11) Stone, R. G., "On Encoding of Commas between Strings," *Commun. ACM,* vol. 22, no. 5, pp. 310-311, May 1979.

(12) Lakshmanan, K. B., "On Universal Codeword Sets," *IEEE Trans. Inform. Theory,* vol. IT-27, no. 5, pp. 659-662, Sep. 1981.

(13) Capocelli, R. M. and De Santis, A., "Regular Universal Codeword Sets," *IEEE Trans. Inform. Theory,* vol. IT-32, no. 1, pp. 129-133, Jan. 1986.

(14) Apostolico, A. and Fraenkel, A. S., "Robust Transmission of Unbounded Strings Using Fibonacci Representations," *IEEE Trans. Inform. Theory,* vol. IT-33, no. 2, pp. 238-245, Mar. 1987.

(15) Wang, M., "Almost Asymptotically Optimal Flag Encoding of the Integers," *IEEE Trans. Inform. Theory,* vol. IT -34, no. 2, pp. 324-326, Mar. 1988.

(16) Capocelli, R. M., "Comments and Additions to 'Robust Transmission of Unbounded Strings Using Fibonacci Representations," *IEEE Trans. Inform. Theory,* vol. 35, no. 1, pp. 191-193, Jan. 1989.

(17) Capocelli, R. M., "Flag Encoding Related to the Zeckendorf Representation of Integers," in R. M. Capocelli (Ed.), *Sequences: Combinatorics, Compression, Security, and Transmission,* pp. 449-466, Springer-Verlag, 1990.

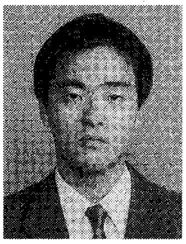
(18) Yamamoto, H. and Ochi, H., "A New Asymptotically Optimal Code of the Positive Integers," *IEEE Trans. Inform. Theory,* vol. 37, no. 5, pp. 1420-1429, Sep. 1991.

(19) Stout, Q. F., "Searching and Encoding for Infinite Order Sets," *Int. J. of Comp. and Inf. Sci.,* vol. 11, no. 1, pp. 55 -72, Feb. 1982.

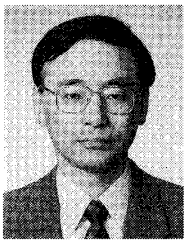
(20) Willems, F. M. J., "Universal Data Compression and Repetition Times," *IEEE Trans. Inform. Theory,* vol. 35, no. 1, pp. 54-58, Jan. 1989.

(21) Wyner, A. D., "An Upper bound on the Entropy Series," *Inform. Contr.,* vol. 20, no. 20, pp. 176-181, 1972.

(22) Gallager, R. G. and Van Voorhis, D. C., "Optimal Source Codes for Geometrically Distributed Integers Alphabets," *IEEE Trans. Inform. Theory,* vol. IT-21, no. 3, pp. 228-230, Mar. 1975.



Takashi Amemiya was born in Kanagawa, Japan, on July 6, 1966. He received the B.E. and M.E. degrees from University of Electro-Communications, Japan, in 1989 and 1991, respectively. He studied universal source coding theories and algorithms in the graduate school. He has been with the software department, NEC Corporation, since April 1991.



Hirosuke Yamamoto was born in Wakayama, Japan, on November 15, 1952. He received the B.E. degree from Shizuoka University, Shizuoka, Japan, in 1975 and the M.E. and Dr.E. degrees from the University of Tokyo, Tokyo, Japan, in 1977 and 1980, respectively, all in electrical engineering. In 1980 he joined the Department of Electronic Engineering at Tokushima University. From 1983 to 1987 he was an Associate Professor at that university. Since 1987 he has been an Associate Professor in the Department of Communications and Systems at University of Electro-Communications, Tokyo, Japan. In 1989-90, he was a Visiting Scholar at the Information Systems Laboratory, Stanford University. His research interests are in Shannon theory, coding theory, and cryptography. Dr. Yamamoto is a member of the IEEE.