社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE
IT2003-5(2003-5)

語頭条件を満たさない単語集合をもつ Word-Valued Sourceの性質について

石田 崇† 後藤 正幸†† 松嶋 敏泰† 平澤 茂一†

† 早稲田大学理工学部 〒169-8555 東京都新宿区大久保 3-4-1
†† 武蔵工業大学環境情報学部 〒224-0015 神奈川県横浜市都筑区牛久保西 3-3-1
E-mail: †{ishida,hirasawa}@hirasa.mgmt.waseda.ac.jp, †toshi@matsu.mgmt.waseda.ac.jp,
††goto@yc.musashi-tech.ac.jp

あらまし 情報源符号化における情報源モデルとして，“言語アルファベット情報源 (word-valued source)”が提案されている [1], [2]. 西新らは，i.i.d. 言語アルファベット情報源を，可算アルファベット $\mathcal{Y}$ 上の i.i.d. (定常無記憶) 情報源と，$\mathcal{Y}$ から有限アルファベット $\mathcal{X}$ の有限系列への写像 ϕ によって定義し，この情報源の漸近等分割性 (AEP) を示し，エントロピー・レートを与えた [1]. 後藤らはこれを定常エルゴード言語アルファベット情報源に対して一般化し，同様の結果を示した [2]. これらの結果は，写像 ϕ が prefix-free であるという条件もとで導かれている．一方，写像 ϕ が prefix-free でない場合については，エントロピー・レートの存在すら明らかではなく，情報源のエントロピー密度レート [6] に対して西新ら [1] がその上界と石田ら [3] が下界を与えたにとどまっている．そこで本稿では，数値計算によってエントロピー・レートや上界・下界の有効性について検証し，ϕ が prefix-free でない言語アルファベット情報源の性質について考察を行う．

キーワード 情報源符号化，言語アルファベット情報源，エントロピー・レート，漸近等分割性

Properties of a Word-valued Source with a Non-prefix-free Word Set

Takashi ISHIDA†, Masayuki GOTO††, Toshiyasu MATSUSHIMA†, and Shigeichi HIRASAWA†

† School of Science and Engineering, Waseda University Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan
†† Faculty of Environmental and Information Studies, Musashi Institute of Technology Ushikubo-nishi 3-3-1, Tuzuki-ku, Yokohama-shi, Kanagawa, 224-0015 Japan
E-mail: †{ishida,hirasawa}@hirasa.mgmt.waseda.ac.jp, †toshi@matsu.mgmt.waseda.ac.jp,
††goto@yc.musashi-tech.ac.jp

Abstract Recently, *word-valued source* is proposed as a new class of source models. A word-valued source is defined as a source which has a probability distribution over *word set*. When the word set is prefix-free, it has been shown that there exists entropy rate of the source with simple expression and the AEP holds. However, when the word set is not prefix-free, it has been shown only the upper bound and lower bound on the entropy rate of the source. In this paper, we verify the entropy rate of the source by numerical computations for some source models in order to clarify the properties of the word-valued source.

Key words source coding, word-valued source, entoropy rate, asymptotic equipartition property (AEP)

1. まえがき

情報源符号化における情報源モデルとして，“言語アルファベット情報源 (word-valued source)”が提案されている [1], [2]. この情報源は，情報源アルファベットの有限系列を単語と定義し，単語単位で確率構造を有するモデルと解釈する事ができる．

西新らは，i.i.d. 言語アルファベット情報源 $\boldsymbol{X}$ を，可算アルファベット $\mathcal{Y}$ 上の i.i.d. (定常無記憶) 情報源 $\boldsymbol{Y}$ と，$\mathcal{Y}$ から有限アルファベット $\mathcal{X}$ の有限系列 (単語) への写像 ϕ によって定義し，この情報源の漸近等分割性 (AEP) を証明し，さらにエントロピー．レートを示した [1]. 後藤らはこれを一般化し，定常エルゴード言語アルファベット情報源に対して，同様の結果を導いた [2]. また，後藤ら [2] や石田ら [4], [5] は言語アルファベット情報源に対するユニバーサル符号の漸近的な性質についても議論をしている．これらの議論では，主に写像 ϕ が prefix-free であることを仮定している．ここで，ϕ が prefix-free であるとは任意の単語が他の単語の語頭に一致していないことを意味し，このとき単語集合 $\mathcal{W}$ は語頭条件を満たしているという．一方，ϕ が prefix-free ではない言語アルファベット情報源は，エントロピー・レートが存在するのかについてもいまだ明らかにされておらず，情報源のエントロピー密度レート [6] について西新ら [1] がその上界を，石田ら [3] が下界を示しているのみである．したがって，これらの上界や下界の厳しさや，エントロピー・レートなどについてより詳細な解析が必要である．

そこで本稿では，ϕ が prefix-free ではない，すなわち単語集合 $\mathcal{W}$ が語頭条件を満たさない言語アルファベット情報源の性質をさらに明らかにするために，i.i.d. 言語アルファベット情報源に対して，様々な設定を与えて数値実験を行い，エントロ

ピー密度レートやその上界，下界などに関する理論的結果の評価を行う．

2. 言語アルファベット情報源 [1], [2]

西新ら [1] によって i.i.d. 言語アルファベット情報源が提案され，情報源のエントロピー・レートが示された．後に後藤ら [2] はそれを“定常エルゴード言語アルファベット情報源”に一般化し同様の結果を導いている．

2.1 定常エルゴード言語アルファベット情報源 [2]

確率変数 Y の無限系列 $\boldsymbol{Y}=(Y_1,Y_2,Y_3,\cdots)$ を可算アルファベット $\mathcal{Y}$ 上に値をとる定常エルゴード情報源とする．$\mathcal{X}^*$ は有限アルファベット $\mathcal{X}$ 上の有限系列すべての集合を表すものとし，写像 ϕ を $\phi:\mathcal{Y}\to\mathcal{X}^*$ で定める．

さらに w を $w=\phi(y)$ $(w\in\mathcal{X}^*, y\in\mathcal{Y})$ によって与えこれを単語とよぶ．ここで，ϕ の値域を $\mathcal{W}$ と書き，これを単語集合とよぶ．さらに確率変数 W を $W=\phi(Y)$ によって定義し，$W=\phi(Y)$ の無限系列を $\boldsymbol{W}=(W_1,W_2,\cdots)$ と表す．

ここで，系列 $\boldsymbol{Y}=(Y_1,Y_2,Y_3,\cdots)$ に対して $W_1=\phi(Y_1), W_2=\phi(Y_2), W_3=\phi(Y_3),\cdots$ の連接を $\phi(\boldsymbol{Y})=\phi(Y_1)\phi(Y_2)\phi(Y_3)\cdots=W_1W_2W_3\cdots=\boldsymbol{W}$ と書く．このとき定常エルゴード言語アルファベット情報源 $\boldsymbol{X}=(X_1,X_2,X_3,\cdots)$ は $\boldsymbol{X}\stackrel{\text{def}}{=}\phi(\boldsymbol{Y})$ で定義する．

確率変数列 $\boldsymbol{X}$ の実現値（情報源から出力される系列）を $\boldsymbol{x}=(x_1,x_2,x_3,\cdots)$ で表す．また，$n=1,2,3,\cdots$ に対して長さ n の有限系列を $X^n=(X_1,X_2,\cdots,X_n)$，$x^n=(x_1,x_2,\cdots,x_n)$ と書く．同様に $m=1,2,3,\cdots$ に対して長さ m の有限系列を $Y^m=(Y_1,Y_2,\cdots,Y_m)$，$y^m=(y_1,y_2,\cdots,y_m)$，$W^m=(W_1,W_2,\cdots,W_m)$，$w^m=(w_1,w_2,\cdots,w_m)$ とする．さらに写像 $\phi:\mathcal{Y}^m\to\mathcal{W}^m$ も同様に系列 $\phi(Y_1),\phi(Y_2),\cdots,\phi(Y_m)$ の連接 $\phi(Y_1^m)=\phi(Y_1)\phi(Y_2)\cdots\phi(Y_m)=W_1W_2\cdots W_m=W^m$ によって定義する(注1)．

言語アルファベット情報源からは，系列は単語単位で出力されるが，それぞれの単語が連接された長さ n の $\mathcal{X}$ 上の系列 x^n として観測される．すなわち $x^n=w^m$ であり，このとき任意の m に対して $n=|w_1|+|w_2|+\cdots+|w_m|$ である．ここで $|w|$ は単語の長さを表す．本稿ではこれ以降，w^n を単語系列，x^n をシンボル系列とよぶことにする．

系列 $\boldsymbol{Y},\boldsymbol{W}$ の確率分布を次のように表す．

$$P_{Y^m}(y^m)\stackrel{\text{def}}{=}\Pr\{Y^m=y^m\},\tag{1}$$

$$P_{W^m}(w^m)\stackrel{\text{def}}{=}\Pr\{W^m=w^m\}.\tag{2}$$

$m=1$ の場合は $P_{W^1}(w)=P_W(w), P_{Y^1}(y)=P_Y(y)$ と書く．

一方，系列 $\boldsymbol{X}$ の確率分布を同様に次のように表す．

$$P_{X^n}(x^n)\stackrel{\text{def}}{=}\Pr\{X^n=x^n\}.\tag{3}$$

$n=1$ の場合は $P_{X^1}(x)=P_X(x)$ と書く．

2.2 エントロピー・レート

$\boldsymbol{X}$ に対して

$$-\frac{1}{n}\log P_{X^n}(X^n)\tag{4}$$

を情報源 $\boldsymbol{X}$ のエントロピー密度レートという [6]．また，エントロピー密度レートの期待値について，

$$H(\boldsymbol{X})=\lim_{n\to\infty}\left[-\frac{1}{n}\sum_{x^n\in\mathcal{X}^n}P_{X^n}(x^n)\log P_{X^n}(x^n)\right],\tag{5}$$

が存在するとき $H(\boldsymbol{X})$ を言語アルファベット情報源 $\boldsymbol{X}$ のエントロピー・レートとよぶ [7]．

単語系列 $\boldsymbol{W}$ のエントロピー・レート $H(\boldsymbol{W})$，平均単語長レート $E[|W|]$ は以下で与えられる [2]．

$$H(\boldsymbol{W})=\lim_{m\to\infty}\left[-\frac{1}{m}\sum_{w^m\in\mathcal{W}^m}P_{W^m}(w^m)\log P_{W^m}(w^m)\right],\tag{6}$$

$$E[|W|]\stackrel{\text{def}}{=}\lim_{m\to\infty}\frac{1}{m}E_{P_{W^m}}\left[\sum_{i=1}^{m}|W_i|\right].\tag{7}$$

ここで，$E_P[\cdot]$ は確率分布 P による期待値を表す．

[定理 1] (定常エルゴード言語アルファベット情報源の AEP [2], [3]) $\boldsymbol{Y}$ を定常エルゴード情報源とする．$\boldsymbol{X}=\phi(\boldsymbol{Y})$，$H(\boldsymbol{Y})<\infty$，$E[|W|]<\infty$ のとき，次式が成り立つ．

$$\limsup_{n\to\infty}\frac{1}{n}[-\log P_{X^n}(X^n)]\leqq\frac{H(\boldsymbol{Y})}{E[|W|]},\ a.s.\tag{8}$$

$$\limsup_{n\to\infty}\frac{1}{n}E_{P_{X^n}}[-\log P_{X^n}(X^n)]\leqq\frac{H(\boldsymbol{Y})}{E[|W|]}.\tag{9}$$

が成り立つ．さらに，ϕ が prefix-free のとき，

$$\lim_{n\to\infty}\frac{1}{n}[-\log P_{X^n}(X^n)]=\frac{H(\boldsymbol{W})}{E[|W|]},\ a.s.\tag{10}$$

$$H(\boldsymbol{X})=\frac{H(\boldsymbol{W})}{E[|W|]}.\tag{11}$$

□

2.3 $\mathcal{W}$ が語頭条件を満たさない言語アルファベット情報源

単語集合 $\mathcal{W}$ が語頭条件を満たさない i.i.d. 言語アルファベット情報源を non-prefix-free i.i.d. 言語アルファベット情報源とよぶことにし，次のように定義する [3]．

[定義 1] (non-prefix-free i.i.d. 言語アルファベット情報源 [3]) non-prefix-free i.i.d. 言語アルファベット情報源 $\boldsymbol{X}=\phi(\boldsymbol{Y})$ は，有限アルファベット $\mathcal{Y}$ 上に値をとる i.i.d. 情報源 $\boldsymbol{Y}$ と，$\mathcal{Y}$ から $\mathcal{W}$ への 1 対 1 写像 $\phi:\mathcal{Y}\to\mathcal{W}=\cup_{i=1}^{K}\mathcal{X}^i$ によって与えられる．ただし，$\mathcal{X}=\{0,1\}$ とする． □

この情報源モデルでは，単語集合 $\mathcal{W}$ を深さ K の完全 2 分木で表現することができ，各ノードが単語 w に対応する (図 1 参照)．このとき，最大単語長は K であり，単語集合 $\mathcal{W}$ の要素数 $||\mathcal{W}||=2(2^K-1)$ である．単語集合 $\mathcal{W}$ に対して確率分布の与え方によって様々な構造を持つ情報源モデルとなるが，一般に $\mathcal{W}$ は語頭条件を満たなさい．石田らはこの情報源に対してエントロピー密度レートの下界式を導出した [3]．

[定理 2] (non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー密度レートの下界 [3]) non-prefix-free i.i.d. アルファベット情報源 $\boldsymbol{X}$ について，

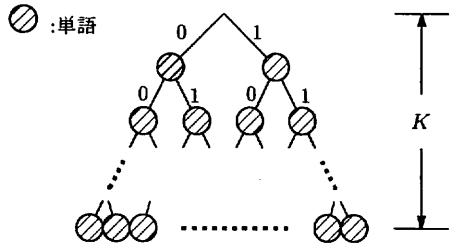
$$\liminf_{n\to\infty}\frac{1}{n}[-\log P_{X^n}(x^n)]\geqq\frac{H(\boldsymbol{W})}{E[|W|]}-\frac{H(S)}{E[|W|]},\ a.s.\tag{12}$$

$$\liminf_{n\to\infty}\frac{1}{n}E_{P_{X^n}}[-\log P_{X^n}(x^n)]\geqq\frac{H(\boldsymbol{W})}{E[|W|]}-\frac{H(S)}{E[|W|]},\tag{13}$$

が成り立つ．ここで，$H(S)$ は単語の長さの確率分布 $P_S(s)$ のエントロピーを表し，

$$H(S)=-\sum_{s=1}^{k}P_S(s)\log P_S(s),\tag{14}$$

(注1)：W^m は確率変数列 $(W_1,W_2,\cdots,W_m)$ を表しているが，特に問題がない場合には $W_1,W_2,\cdots,W_m$ の連接に対しても W^m という表記を用いることにする．w^m についても同様．



$$\mathcal{W} = \{0, 1, 00, 01, 10, 11, 000, \cdots, \underbrace{11\cdots1}_{K}\}$$

$$\|\mathcal{W}\| = \sum_{i=1}^{K} 2^i = 2(2^K - 1)$$

$\|\mathcal{W}\|$:単語集合 $\mathcal{W}$ に含まれる単語数

図 1　単語集合 $\mathcal{W}$ の 2 分木表現

$$P_S(s) = \sum_{\{w:|w|=s\}} P_W(w), \tag{15}$$

で与えられる. □

定理 1,2 より，non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー・レート $H(\boldsymbol{X})$ が存在するならば，

$$\frac{H(\boldsymbol{W})}{E[|W|]} - \frac{H(S)}{E[|W|]} \leq H(\boldsymbol{X}) \leq \frac{H(\boldsymbol{W})}{E[|W|]}, \tag{16}$$

がいえる.

この情報源におけるエントロピー密度レートの上界式，下界式は 1 本のシンボル系列 x^n に写像される単語系列 w^m の本数を評価することによって得られている．いま，十分に長い単語系列 w^m は AEP により出現確率はほぼ $2^{-mH(\boldsymbol{W})}$ である．出現する x^n に対しては少なくとも 1 つの単語系列 w^m が写像されるので，$P_{X^n}(x^n) \geq 2^{-mH(\boldsymbol{W})}$ が成り立ち，これより上界式が得られる [1], [2].

次に下界であるが，x^n を観測した場合，複数の単語系列 w^m がこのシンボル系列に写像されており，単語の切れ目がどこなのかを一意に決めることができない．しかし，単語の切れ方を確定してしまえば，同じような単語の切れ方をする異なった単語系列 w^m は必ず異なるシンボル系列 x^n に写像されることから，単語の切れ目の入れ方の総数を評価すればよいことが分かり，これによって下界を導いている [3].

以上のように，$\mathcal{W}$ が語頭条件を満たさない場合については，定常エルゴード言語アルファベット情報源に対してエントロピー密度レートの上界（式 (8),(9)），i.i.d. 言語アルファベット情報源に対してその下界 [3] が得られているのみで，エントロピー・レート $H(\boldsymbol{X})$ が存在するのか，いまだ明らかにされていない．また，上界，下界の有効性についても理論的に評価がされていない.

そこで本稿では数値実験によって，$\mathcal{W}$ が語頭条件を満たさない言語アルファベット情報源のエントロピー・レートやその性質に関する検証を行う.

3. 数値計算と考察

non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー・レートについてさらに議論を行うためには，シンボル系列 x^n の生起確率 $P_{X^n}(x^n)$ を陽な形で与える必要があると考えられるが，いまだその結果には至っていない．$P_{X^n}(x^n)$ は x^n と w^m の写像によって決定されるが，この写像は単語集合 W の構造や，単語系列 w^m における単語の並び方などに依存すると考えられ，理論的な解析を行うことは容易ではない.

そこで本稿では，数値計算によって以下の内容を検証する.

- non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー・レートは存在するのか.
- 上界，下界の有効性はどうか.
- 情報源の確率構造によってエントロピー密度レートとその上界，下界の性能はどのように変化するのか.

以上の内容を検証するために，いくつかの設定のもとで i.i.d. 言語アルファベット情報源に実際に単語の確率分布 $P_W(w)$ を数値で与え，情報源から出力されたシンボル系列 x^n に対して次式によって生起確率 $P_{X^n}(x^n)$ を計算し，エントロピー密度レートを求める.

$$P_{X^n}(x^n) = \sum_{\{w^m:x^n=w^m\}} P_{W^m}(w^m). \tag{17}$$

なお，計算ではすべて log の底を 2 としている.

3.1 情報源のエントロピー密度レートの上界と下界の有効性について

まず始めに，式 (16) について検証する．すなわち，情報源のエントロピー・レートが存在するのか，また，上界，下界の性能はどれほどなのかを調べるために以下の 4 つの単語集合モデルについて数値計算を行う.

[model 1] 全ての単語の長さが等しいモデル
[model 2] $\mathcal{W}$ が語頭条件を満たすモデル
[model 3] $\mathcal{W}$ が語頭条件を満たさないモデル
[model 4] 単語の長さが全て異なるモデル

具体的なパラメータは以下のように与える.

[model 1] 全ての単語の長さが等しいモデル

$$\mathcal{W}_1 = \{000, 001, 010, 011, 100, 101, 110, 111\}$$

$$\begin{aligned} P_W(000) &= 0.05, \quad P_W(001) = 0.05, \\ P_W(010) &= 0.10, \quad P_W(011) = 0.10, \\ P_W(100) &= 0.20, \quad P_W(101) = 0.20, \\ P_W(110) &= 0.20, \quad P_W(111) = 0.10. \end{aligned}$$

[model 2] $\mathcal{W}$ が語頭条件を満たすモデル

$$\mathcal{W}_2 = \{0, 10, 110, 111\}$$

$$\begin{aligned} P_W(0) &= 0.50, \quad P_W(10) = 0.20, \\ P_W(110) &= 0.20, \quad P_W(111) = 0.10. \end{aligned}$$

[model 3] $\mathcal{W}$ が語頭条件を満たさないモデル

$$\mathcal{W}_3 = \{0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111\}$$

$$\begin{aligned} P_W(0) &= 0.04, \quad P_W(1) = 0.05, \\ P_W(00) &= 0.12, \quad P_W(01) = 0.01, \\ P_W(10) &= 0.01, \quad P_W(11) = 0.11, \\ P_W(000) &= 0.06, \quad P_W(001) = 0.10, \\ P_W(010) &= 0.12, \quad P_W(011) = 0.01, \\ P_W(100) &= 0.11, \quad P_W(101) = 0.10, \\ P_W(110) &= 0.04, \quad P_W(111) = 0.12. \end{aligned}$$

[model 4] 単語の長さが全て異なるモデル

$$\mathcal{W}_4 = \{0, 10, 101\}$$

$$P_W(0) = 0.50, \quad P_W(10) = 0.30, \quad P_W(101) = 0.20.$$

単語集合モデル $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3, \mathcal{W}_4$ の木表現を図 2 に示す．ここで，斜線の引かれたノードが出現する単語 ($P_W(w) > 0$) を表している.

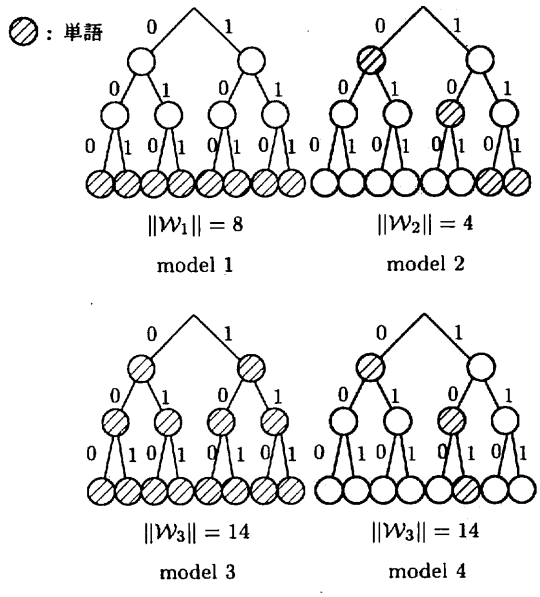


図 2　単語集合モデル $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$

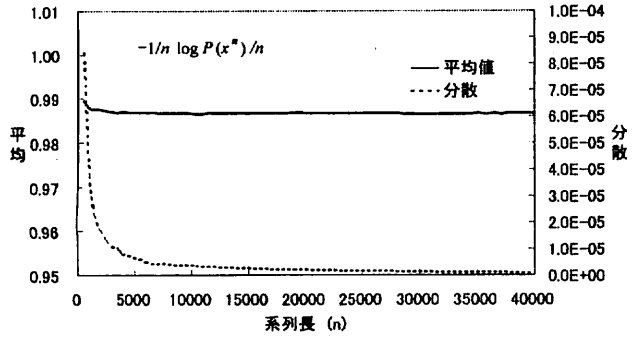


図 3　系列長 n に対するエントロピー密度レートの推移

3.1.1　エントロピー・レートに関する結果と考察

まず，系列長 n に対して，エントロピー密度レートの値がどのように推移しているかを確認する．ここでは，model 3 についてのグラフを図 3 に示す．このグラフでは，x^n の 200 系列それぞれのエントロピー密度レートの平均値（実線）と分散（破線）の推移が示されている．

この図よりエントロピー密度レートの平均値がほぼ収束していることから，non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー・レートが存在するのではないかと予想される．さらに，n が大きくなるにつれて分散の値が限りなく 0 に近づいていることから，漸近等分割性も成り立っているのではないかと考えられる．さらに単語集合 $\mathcal{W}_3$ に対して model 3 とは異なる確率分布 $P_W(w)$ を割り振って計算を何通りか行ったが，いずれについても同様の結果が得られている．

3.1.2　エントロピー・レートの上界と下界の有効性に関する結果と考察

3.1.1 の結果を踏まえて，以降の数値計算では $n = 40000$ のシンボル系列長に対して 200 系列のエントロピー密度レートの平均値を求め，これを non-prefix-free i.i.d. 言語アルファベット情報源のエントロピー・レートであると考えることにする．

model 1 から model 4 までの計算結果を表 1 に示す．

model 1，model 2 は $\mathcal{W}$ が語頭条件を満たしていることから，これらの情報源のエントロピー・レート $H(\boldsymbol{X})$ は存在し，式 (11) から求められる．

（1）model 1 では上界と下界がエントロピー・レートと一致し，エントロピー密度レートの平均値もエントロピー・レートに一致していることが計算結果より分かる．式 (16) より，上界と下界との差分は $H(S)/E[|\boldsymbol{W}|]$ で与えられるが，単語の長さが全て等しいモデルでは $H(S)$ が 0 となることから確かに下界が上界と一致するモデルであること分かる．単語の長さが全て等しいということはシンボル系列 x^n に対して単語の切れ目が一意に決定できることから，シンボル系列 x^n と単語系列 w^m が 1 対 1 に対応していることを意味している．

表 1　エントロピー密度レートの平均値と分散 (n=40000)

model	上界	平均	entropy rate	下界	分散
1	0.941	0.941	0.941	0.941	$3.92 \cdot 10^{-6}$
2	0.978	0.978	0.978	0.153	$1.60 \cdot 10^{-6}$
3	1.361	0.987	—	0.890	$8.50 \cdot 10^{-7}$
4	0.874	0.747	—	0.000	$3.21 \cdot 10^{-6}$

entropy rate：式 (11) より計算．
上界：式 (16) の右辺より計算．
下界：式 (16) の左辺より計算．

（2）model 2 は各単語の長さは一定値ではないものの $\mathcal{W}$ が語頭条件を満たしていることから，上界とエントロピー・レートが一致するモデルである．しかし，このモデルでは $H(S)$ が必ずしも 0 とはならないことから下界は上界と一致しない．これは，下界を導出する過程で単語の長さのみに着目して x^n へ写像される w^m の本数を評価したために，本来 1 対 1 対応してしているものを過大に評価しているためだといえる．

（3）model 3 は $\mathcal{W}$ が語頭条件を満たさないため，陽な形でエントロピー・レートを求めることはできない．このモデルではエントロピー密度レートの平均値は上界にも下界にも一致しないが，計算結果より上界とは大きく離れて下界に近い値をとっていることが分かる．これは，下界の有効性を示しており，x^n へ写像される w^m の本数は単語の長さの確率分布 $P_S(s)$ によってよい精度で評価できているのではないかと考えられる．単語集合 $\mathcal{W}_3$ に対して，異なる確率分布 $P_W(w)$ を与えたモデルについても，ほとんどの場合で同様の結果を得た．

（4）model 4 では下界値が 0 になっている．これは各単語の長さが全て異なり $H(S) = H(\boldsymbol{W})$ となるためで，式 (16) からも明らかである．このとき下界は意味をなさない．これは語頭条件を満たしている場合にも起こる．

結果をまとめると以下のことが言える．

- non-prefix-free i.i.d. 言語アルファベット情報源のエントロピーレートは存在すると考えられる．
- 一般的な non-prefix-free i.i.d. 言語アルファベット情報源に対して，上界はあまくなるが，下界はタイトであると考えられる．

3.2　情報源の確率構造とエントロピー・レートの振る舞いについて

前項の結果からも明らかなように，$\mathcal{W}$ が語頭条件を満たさないモデルでは確率分布 $P_W(w)$ の与え方によって様々な性質を有することがわかる．

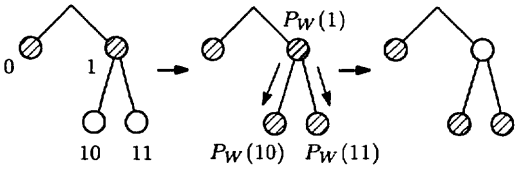
そこで本項では $\mathcal{W}$ が語頭条件を満たさないモデルの確率構造を明らかにすることを目的として，いくつかの場合を想定して単語の確率分布を与え，エントロピー密度レートの平均と上界，下界を計算し，それらがどのような性質をもつのか数値計算によって検証する．

単語集合 $\mathcal{W}$ が語頭条件を満たさない情報源モデルとして以下のような場合を考える．

[case A]　$\mathcal{W}$ の語頭条件が徐々に崩れる場合
[case B]　$H(S)$ 一定のもとで $E[|W|]$ が変化する場合
[case C]　同じ長さの単語同士が等確率で出現する場合
[case D]　上界と下界が一定の場合

表 2　model 5-1 – 5-4 の $P_W(w)$

	$P_W(0)$	$P_W(1)$	$P_W(10)$	$P_W(11)$
model 5-1	0.10	0.90	0.00	0.00
model 5-2	0.10	0.80	0.02	0.08
model 5-3	0.10	0.10	0.20	0.60
model 5-4	0.10	0.00	0.25	0.65



model 5-1　　model 5-2, 5-3　　model 5-4

図 4　case A の単語集合

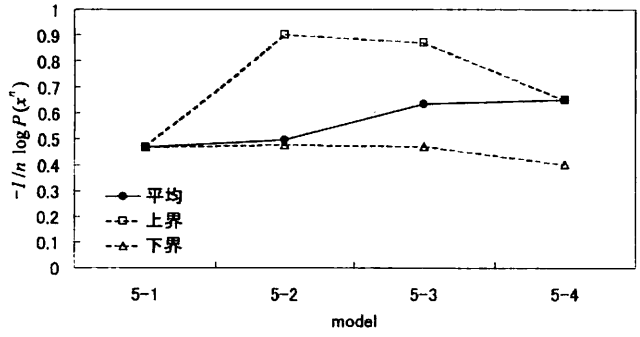


図 5　model 5-1 – 5-4 の数値計算結果

(1) case A : $\mathcal{W}$ の語頭条件が徐々に崩れる場合

任意の s について単語長の確率分布 $P_S(s)$ が 0 か 1 の値をとるとき，このモデルは語頭条件を満たすモデルとなるが，わずかでも違う長さの単語に出現確率が割り当てられた瞬間，このモデルは語頭条件を満たさなくなる．

そこで，単純なモデルとして単語集合 $\mathcal{W}_5 = \{0, 1, 10, 11\}$ を考え，表 2 のように単語の出現確率 $P_W(w)$ を割り当てた model 5-1 – 5-4 を設定する．

これらのモデルは次のような構造になっている．model 5-1 は長さ 1 の単語しか出現しない，語頭条件を満たすモデルであるが，model 5-2,5-3 になるにつれて次第に $P_W(1)$ を $P_W(10), P_W(11)$ に割り振っていき，model 5-4 で完全に $P_W(1) = 0$ となって再び語頭条件を満たすモデルとなる (図 4)．数値計算の結果を図 5 に示す．

model 5-1 は全ての単語長が等しいモデルなので上界，下界，エントロピー密度レートが情報源のエントロピー・レートと一致する特殊な場合である．model 5-2 は $P_W(1)$ を $P_W(10), P_W(11)$ に割り振ることによって model 5-1 の語頭条件がわずかに崩れたモデルとなっているが，上界が一気にエントロピー密度レートの平均値よりも大きくかけ離れてしまうことが分かる．一方，下界とエントロピー密度レートとの差はほとんどない．しかし，さらに $P_W(1)$ の割り振りを続けると徐々にエントロピー・密度レートの平均値と上界の値が近づいてゆき (model 5-3)，$P_W(1) = 0$ となったとき (完全に語頭条件を満たすとき) 両者は完全に一致する．

(2) case B : $H(S)$ 一定のもとで $E[|W|]$ が変化する場合

case B では $H(S)$ が一定のもとで平均単語長 $E[|W|]$ が増加したときにエントロピー密度レートの値と上界，下界がどのような性能を示すのかを調べる．

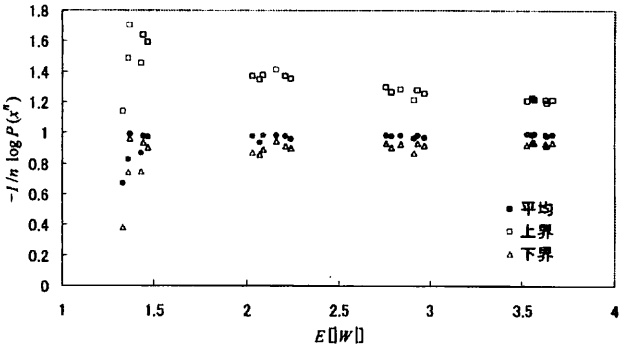


図 6　case B の数値計算結果

表 3　model 7-1 – 7-4 の $P_S(s)$

	$P_S(1)$	$P_S(2)$	$P_S(3)$	$P_S(4)$
model 7-1	1.00	0.00	0.00	0.00
model 7-2	0.40	0.60	0.00	0.00
model 7-3	0.20	0.30	0.40	0.00
model 7-4	0.10	0.20	0.30	0.40

このモデルでは，最大深さ $K = 4$ の完全 2 分木全てのノードに対応する単語集合 $\mathcal{W}_6$ を考える ($||\mathcal{W}_6|| = 30$)．長さ s の出現確率 $P_S(s)$ $(s = 1, 2, 3, 4)$ は 4 つの定数の組 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ (ただし，$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$) を割り当てることによって定めるものとする．このとき，α の割り当て方の組み合わせは $4! = 24$ 通りあるので，それぞれを 1 つの情報源モデルとする (model 6-1–model 6-24)．このとき，24 個全てのモデルの $H(S)$ は等しい．また，24 個の情報源モデルにおいて各単語の出現確率 $P_W(w)$ は割り当てられた $P_S(s)$ を満たすように，ランダムに割り振ることにする．

$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0, 80, 0.10, 0.07, 0.03)$ として，24 個のモデルに対する数値実験の結果を図 6 に示す．このグラフは 24 個それぞれのモデルについて求めたエントロピー密度レートの平均値，上界，下界の値を，横軸に平均単語長 $E[|W|]$ をとってプロットしたものである．

グラフより，$H(S)$ 一定のもとで平均単語長 $E[|W|]$ の増加にともなって上界と下界の差が小さくなっていく様子が分かる．これは，上界と下界のの差分が $H(P_S)/E[||W]$ で与えられることから，明らかな結果である．

(3) case C : 同じ長さの単語同士が等確率で出現する場合

最大深さ $K = 4$ の完全 2 分木によって定まる単語集合 $\mathcal{W}_7$ について，長さが等しい単語同士が等確率で出現するような情報源モデルを考える．すなわち，単語長の確率分布 $P_S(s)$ が与えられたもとで，長さが s の単語の出現確率が $P_S(s)/2^s$ で与えられるモデルである．単語長の確率分布 $P_S(s)$ $(s = 1, 2, 3, 4)$ を表 3 のように与える．

このモデルに対する数値計算結果を図 7 に示す．グラフから，エントロピー密度レート平均値と下界は常に一致し，1.0 となっていることが分かる．このような確率構造を持つモデルについては，長さ s の単語の個数が 2^s 個で，その出現確率が全て $P_S(s)/2^s$ であることから，式 (16) の左辺 (下界式) より

$$\frac{H(W)}{E[|W|]} - \frac{H(S)}{E[|W|]}$$
$$= \frac{-\sum_w P_W(w) \log P_W(w) + \sum_s P_S(s) \log P_S(s)}{E[|W|]}$$
$$= \frac{-\sum_s P_S(s) \log(P_S(s)/2^s) + \sum_s P_S(s) \log P_S(s)}{E[|W|]}$$

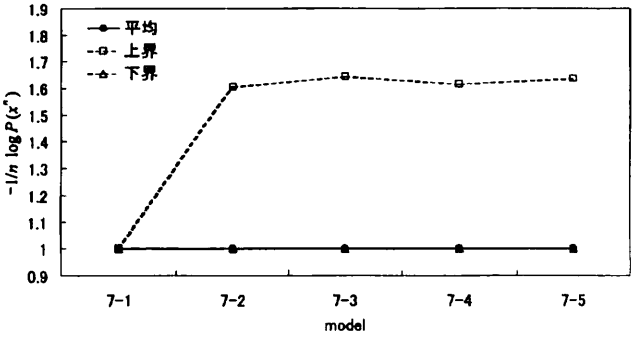


図 7　case C の数値計算結果

表 4　model 8-1 - 8-12 の $P_W(w)$

	$P_W(00)$	$P_W(01)$	$P_W(10)$	$P_W(11)$
model 8-1	0.03	0.07	0.00	0.00
model 8-2	0.07	0.03	0.00	0.00
model 8-3	0.00	0.00	0.03	0.07
model 8-4	0.00	0.00	0.07	0.03
model 8-5	0.03	0.00	0.07	0.00
model 8-6	0.07	0.00	0.03	0.00
model 8-7	0.03	0.00	0.00	0.07
model 8-8	0.07	0.00	0.00	0.03
model 8-9	0.00	0.03	0.07	0.00
model 8-10	0.00	0.07	0.03	0.00
model 8-11	0.00	0.03	0.00	0.07
model 8-12	0.00	0.07	0.00	0.03

全ての model について $P_W(0) = 0.2$, $P_W(1) = 0.7$

$$= \frac{\sum_s s \cdot P_S(s)}{E[|W|]} = 1, \qquad (18)$$

より，下界値が 1 となることは容易に導かれる．また，エントロピー密度レートの値が下界と一致するという事は，[3] で下界を導出する際に評価した x^n へ写像される w^m の本数が等号で成り立っていることが分かる．

（4）case D：上界と下界が一定の場合

最後に，上界と下界は一定であるが確率構造の異なる情報源モデルを考える．$\mathcal{W}_8 = \{0, 1, 00, 01, 10, 11\}$ とし，単語 0 と 1 の生起確率を $P_W(0) = 0.2$, $P_W(1) = 0.7$ とし，残りの 0.1 を表 4 に示すように割り当てる．すなわち，case D では長さ 1 の単語の生起確率を固定し，長さ 2 の単語について生起確率の値を入れ替える事によってできるモデルの集合を考えている．

各モデルで単語の生起確率 $P_W(w)$ がとる値の組み合わせはどのモデルについても共通であり，また，$P_S(s)$ も共通であることから，どのモデルの $H(\boldsymbol{W})$ も $H(S)$ も等しい．さらに，平均単語長 $E[|W|]$ も共通なので，すべての情報源について式 (16) によって計算される上界，下界は等しい．

case D の数値計算結果を図 8 に示す．この結果から，上界，下界の値が同じでもエントロピー密度レートの値はモデルによって異なる事がわかる．すなわち，$\mathcal{W}$ が語頭条件を満たさない言語アルファベット情報源のエントロピー・レートの性質を明らかにするにはこれまでの解析でなされてきた議論やパラメータだけでは不十分である可能性を示唆している．

しかし，model 8-1 と model 8-5 や model 8-2 と model 8-6 などはエントロピー密度レートが同じ値になっており，規則性が見られる．これらは $P_W(01)$ と $P_W(10)$ の確率が入れ替わっただけのモデル同士であるが，このような情報源は同じエントロピー・レートの値を持つと考えられる．

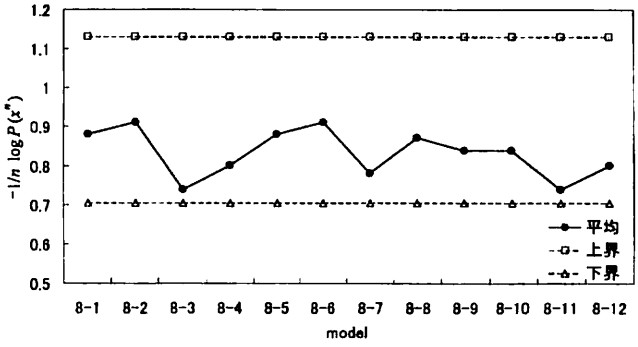


図 8　model 8 の数値計算結果

4. むすび

本稿では，non-prefix-free i.i.d. 言語アルファベット情報源について，エントロピー・レートや情報源の構造に関する性質を明らかにすることを目的として，様々な設定のもとで情報源モデルに対して数値計算を行い，エントロピー密度レートやその上界，下界のとる値について検証を行った．検証の結果として，以下のようなことが挙げられる．

non-prefix-free i.i.d. 言語アルファベット情報源について，

- エントロピー・レートが存在すると考えられる．
- AEP が成り立つと考えられる．
- 一般に式 (16) の上界式はあまく，下界式タイトである．
- 単語の確率分布 $P_W(w)$ の与え方によってエントロピー密度レートとその上界，下界は大きく異なる性質を有する．
- 情報源の確率構造をさらに解析するためには新たな視点が必要であると考えられる．

しかし，non-prefix-free i.i.d. 言語アルファベット情報源に関する理論的な解析はいまだ十分になされていない．本稿における数値計算から得られた結果や，予想をもとにさらに語頭条件を満たさない情報源モデルの構造を明らかにし，エントロピー・レート $H(\boldsymbol{X})$ の収束性や，漸近等分割性についての理論的な解析を進めていく必要がある．また，$\mathcal{W}$ が語頭条件を満さないが，モデルに制約を加えることで情報源のエントロピー・レートが定義できるような情報源モデルがないか考えたい．

謝辞　著者の 1 人石田は，本研究を進めるにあたり熱心に議論をしてくださる湘南工科大学 小林学先生に感謝いたします．

文　献

[1] M.Nishiara and H.Morita, "On the AEP of word-valued sources," *IEEE Trans. Inform. Theory*, vol.IT-46, no.3, pp.1116-1120, May 2000.

[2] Masayuki GOTO, Toshiyasu MATSUSHIMA, and Shigeichi HIRASAWA, "A source model with probability distribution over word set and recurrence time theorem," submitted to IEICE Trans. Fundamentals. Special Section on Information Theory and Its Applications, 2003.

[3] 石田 崇，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の性質について，" 第 25 回情報理論とその応用シンポジウム予稿集 (SITA2002), pp.695-698, 2002.

[4] 石田 崇，後藤 正幸，平澤 茂一，"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，" 電子情報通信学会論文誌, vol.J84-A, no.9,pp.1167-1178, 2001.

[5] 石田 崇，後藤 正幸，平澤 茂一，"単語単位で系列を出力する情報源に対する LZ78 符号のユニバーサル性について，" 第 24 回情報理論とその応用シンポジウム予稿集 (SITA2001), pp.243-246, 2001.

[6] 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.

[7] 韓 太舜，小林 欣吾，情報と符号化の数理，培風館,1999.