

1 平均符号長とエントロピー

一意復号化可能な符号においてエントロピーは平均符号長の下界であることを示したが、実はある意味で下限でもあることが今回の主題である。

エントロピーが下界であることは本質的に

Proposition 1 (クラフトの不等式). 任意の一意復号化可能な r 元符号 $\varphi: \Sigma \rightarrow A^*$ について

$$\sum_{\sigma \in \Sigma} r^{-|\varphi(\sigma)|} \leq 1$$

が成り立つ。

から導かれた事に注意する。クラフトの不等式のほぼ逆に相当する次の事実が平均符号長のエントロピーに関する逆の評価を与える。

Proposition 2. 有限集合 Σ で添え字付けられた自然数の組 $(l_\sigma)_{\sigma \in \Sigma}$ が

$$\sum_{\sigma \in \Sigma} r^{-l_\sigma} \leq 1$$

を満たすとき、瞬時復号化可能な r 元符号 $\varphi: \Sigma \rightarrow A^*$ で

$$|\varphi(\sigma)| = l_\sigma$$

を満たすものが存在する。

この事実は実際に構成手段を与えることによって証明できる。証明は後に回してこれを使ったエントロピーの評価を先に見よう。

Theorem 1. 任意の情報源 $(\Sigma, (p_\sigma))$ について

$$\sum |\varphi(\sigma)| p_\sigma < H_r((\Sigma, (p_\sigma))) + 1$$

を満たす瞬時復号化可能な符号が存在する。

「符号長」は整数である事を考えれば $+1$ の余裕で上から押さえるこの評価は最善の評価であると言える。

Proof. 各 $\sigma \in \Sigma$ について l_σ を

$$r^{-(l_\sigma-1)} > p_\sigma \geq r^{-l_\sigma}$$

を満たすものとする。¹ 定義より

$$\sum_{\sigma \in \Sigma} r^{-l_\sigma} \leq \sum_{\sigma \in \Sigma} p_\sigma = 1$$

を得るので Proposition 2 から $|\varphi(\sigma)| = l_\sigma$ を満たす瞬時復号化可能な符号が存在する。

一方 l_σ の定義から両辺に r を底とした対数を取って -1 をかけたもの

$$l_\sigma - 1 < -p_\sigma \leq l_\sigma$$

を用いれば平均符号長について求める評価が与えられる。即ち、

$$\sum |\varphi(\sigma)| p_\sigma = \sum l_\sigma p_\sigma < \sum (-\log_r p_\sigma + 1) p_\sigma = \sum (-\log_r p_\sigma) p_\sigma + \sum p_\sigma = H_r((\Sigma, (p_\sigma))) + 1$$

□

¹ l_σ は $-\log_r p_\sigma$ を整数に切り上げた値となる。

2 シャノン-ファノ符号

2.1 構成例

情報源文字		確率	を考える. この情報源の 2 を底としたエントロピーは
情報源	a	0.45	
	b	0.20	
	c	0.05	
	d	0.05	
	e	0.25	

$$-0.45 \log_2 0.45 - 0.20 \log_2 0.20 - 0.05 \log_2 0.05 - 0.05 \log_2 0.05 - 0.25 \log_2 0.25 = 1.9149 \dots \text{ [bit]}$$

であるから, この情報源の一意復号化可能な 2 元符号の平均符号長は必ず $1.9149 \dots$ 以上である. また平均符号長が $1.9149 \dots + 1 = 2.9149 \dots$ 以下の瞬時復号化可能な 2 元符号が次の手順で構成できる.

l	1	2	3	4	5	6	...
2^{-l}	0.5	0.25	0.125	0.0625	0.03125	0.015625	...

1. $2^{-(l_\sigma-1)} > p_\sigma \geq 2^{-l_\sigma}$ なる l_σ を求める.	情報源文字 σ	確率 p_σ	l_σ
	a	0.45	2
	b	0.20	3
	c	0.05	5
	d	0.05	5
	e	0.25	2

2. l_σ によって昇順に並べる	i	情報源文字 σ_i	l_{σ_i}
	1	a	2
	2	e	2
	3	b	3
	4	c	5
	5	d	5

3. $q_i := \sum_{j=1}^{i-1} 2^{-l_{\sigma_j}}$ を計算する.	i	情報源文字 σ_i	l_{σ_i}	$2^{l_{\sigma_i}}$	q_i
	1	a	2	0.25	0
	2	e	2	0.25	0.25
	3	b	3	0.125	0.5
	4	c	5	0.03125	0.625
	5	d	5	0.03125	0.628125

4. q_i の 2 進小数表示の l_{σ_i} 桁を符号にする	i	情報源文字 σ_i	l_{σ_i}	$2^{l_{\sigma_i}}$	q_i	q_i の 2 進小数表示	$\varphi(\sigma_i)$
	1	a	2	0.25	0	0.000000...	00
	2	e	2	0.25	0.25	0.010000...	01
	3	b	3	0.125	0.5	0.100000...	100
	4	c	5	0.03125	0.625	0.101000...	10100
	5	d	5	0.03125	0.628125	0.101010...	10101

結果として符合	情報源文字 σ	確率 p_σ	符号長 l_σ	符号 $\varphi(\sigma_i)$	ができた. この符号の平均符号長は
	a	0.45	2	00	
	b	0.20	3	100	
	c	0.05	5	10100	
	d	0.05	5	10101	
	e	0.25	2	01	

2.5 である. また任意の符号語が他の符号語の接頭語で無いから瞬時復号化可能である.

2.2 準備: r 進小数

r を 2 以上の整数とする.

Definition 1. $x \in [0, 1]$ について

$$y = \sum_{k=1}^l \frac{a_k}{r^k}$$

を満たす $(a_1, \dots, a_l) \in \{0, 1, \dots, r-1\}^+$ が存在する時, $(a_1, \dots, a_l)$ を x の r 進小数と呼び, x を $0.a_1 \dots a_l$ で表現する.²

Proposition 3. r 進小数の最後に 0 をつけても値は変わらない, 即ち,

$$0.a_1 \dots a_l = 0.a_1 \dots a_l 0.$$

Proposition 4 (切捨て/切り上げ). r 進小数 $x = \sum_{k=1}^l \frac{a_k}{r^k}$ について

$$\sum_{k=1}^{l'} \frac{a_k}{r^k} \leq x < \sum_{k=1}^{l'} \frac{a_k}{r^k} + \frac{1}{r^{l'}}$$

が成り立つ. ($l' \leq l$)

2.3 Proof Proposition 2

構成的に証明を与える.

Proof. 簡単の為に符号アルファベットは $A := \{0, 1, 2, \dots, r-1\}$ なる r 元集合とする.

r 元シャノンファノ符号の構成法

1. $r^{-l_\sigma} \leq p_\sigma < r^{-l_\sigma+1}$ なる l_σ を求める. ($-\log_r p_\sigma$ の端数を切り上げた値を計算しても同じ.)
2. l_σ によって昇順に並べる
3. $q_i := \sum_{j=1}^{i-1} r^{-l_{\sigma_j}}$ を計算する.
4. q_i の r 進小数表示の l_{σ_i} 桁を符号にする

以下でこの φ は瞬時復号化であること, 具体的には Σ の任意の異なる 2 元 σ_i と σ_j について一方が他方の接頭語でない事を示す. 必要ならば順序を入れ替えれば良いので $i < j$ と仮定して一般性を失わない. 昇順に並べたので $l_{\sigma_i} \leq l_{\sigma_j}$ である.

q_i, q_j について, q_j の定義と $i \leq j-1$ から,

$$q_j = r^{-l_{\sigma_1}} + \dots + r^{-l_{\sigma_i}} + \dots + r^{-l_{\sigma_{j-1}}} \geq r^{-l_{\sigma_1}} + \dots + r^{-l_{\sigma_i}} = q_i + r^{-l_{\sigma_i}}$$

であるから, q_j と q_i の r 進小数展開 $.a_1^{(j)} a_2^{(j)} \dots a_{l_j}^{(j)}$ と $.a_1^{(i)} a_2^{(i)} \dots a_{l_i}^{(i)}$ に関して

$$\sum_{k=1}^{l_{\sigma_i}} \frac{a_k^{(j)}}{r^k} + \frac{1}{r^{l_{\sigma_i}}} > q_j \geq q_i + r^{-l_{\sigma_i}} \geq \sum_{k=1}^{l_{\sigma_i}} \frac{a_k^{(i)}}{r^k} + \frac{1}{r^{l_{\sigma_i}}}$$

²厳密に言えば, $\{0, 1, \dots, r-1\}$ の元 a_i ではなく, $\{0, 1, \dots, r-1\}$ で添え字付けられた文字の集合 $A = \{\alpha^{(0)}, \dots, \alpha^{(r-1)}\}$ の $a^{(i)}$ 番目の文字 $\alpha^{(a_i)}$ を用いて「 $0.a^{(a_1)} \dots a^{(a_l)}$ 」で表現する」とすべきである.

具体例をあげてみると, $x = \frac{1}{16} + \frac{11}{(16)^2} + \frac{15}{(16)^3}$ の 16 進小数表現は 0.11115 ではなく 16 種類の文字の集合 $A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$ の 1 番目, 11 番目, 15 番目の文字を用いて 0.1BF で表す.

然しながら, 特に r が 10 以下の場合には通常の「数字」自体を文字集合として扱うことができるので, 記号の煩雑さを避ける為に以下の議論では数字 $a \in \{0, 1, \dots, r-1\}$ と数字を表す文字 $\alpha^{(a)}$ を区別しない.

よって, 左の不等号が等号を含まないことから,

$$\sum_{k=1}^{l_{\sigma_i}} \frac{a_k^{(j)}}{r^k} \neq \sum_{k=1}^{l_{\sigma_i}} \frac{a_k^{(i)}}{r^k}$$

即ち

$$a_1^{(j)} a_2^{(j)} \cdots a_{l_{\sigma_i}}^{(j)} \neq a_1^{(i)} a_2^{(i)} \cdots a_{l_{\sigma_i}}^{(i)}$$

つまり $\varphi(\sigma_j)$ (長い方の符号) の l_{σ_i} 文字の接頭語が $\varphi(\sigma_i)$ (短い方の符号) と一致しない事を得た. $\square$