Vol.2010-CE-104 No.6
2010/5/22

情報エントロピーの多項式表現

高木和久†

情報エントロピー（平均情報量）の概念は C.E.シャノンによって 1948 年に導入された．この定義には対数が用いられているため，必ずしも全ての学生にとって理解が容易であるとは言えない．情報エントロピーは全ての事象の発生確率が同じとき最大となる．この証明には通常は偏微分が用いられるが，微積分を用いずに証明することもできる．この証明の過程で情報エントロピーを多項式を用いて表現する方法を発見した．この定義には対数は用いられず，また理解が容易である．本論文ではこの多項式表現の応用例についても述べる．

Polynomial Expression of Shannon Entropy

Kazuhisa Takagi†

The concept of entropy is introduced by C. E. Shannon in 1948. As the definition needs knowledge of logarithm, some students have difficulties to calculate Shannon entropy. As uncertainty is highest when all possible events are equi-probable, entropy function is maximal if all the outcomes are equally likely. It can be provable without using calculus. When analyzing this proof, I found the polynomial expression of Shannon entropy. It doesn't use logarithm, and is very easy to understand. Some examples of this expression will be shown in this paper.

1. はじめに

C. E. シャノンは確率 $p_1, p_2, \cdots, p_n$ で起こる事象のエントロピー H を

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

と定義した[1]．$n = 2$ のときのHは 2 値エントロピー関数と呼ばれ，式は

$$H = -p \log_2 p - (1-p) \log_2 (1-p)$$

で与えられる．このグラフを図 1 に示す．2 値エントロピー関数は $p = \frac{1}{2}$ のとき最大となる．偏微分のラグランジュの乗数法を用いることにより， $p_1 = p_2 = \cdots = p_n$ のときにエントロピー H が最大となることを証明することができる．

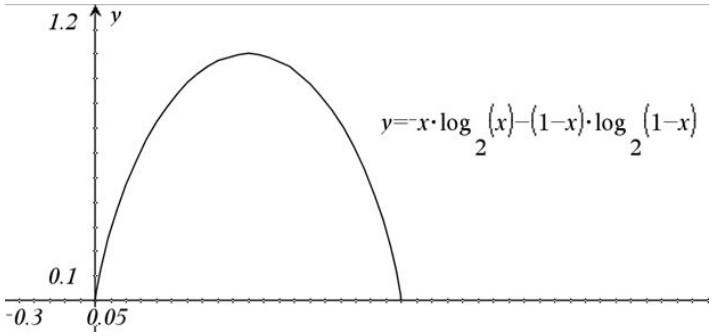

図 1　2 値エントロピー関数

以前，大学の非常勤講師として情報処理概論の講義を担当した際に，エントロピーを題材として取り扱ったことがある．対象は高等学校で対数を履修している理科系の学生である．このときは 2 値エントロピー関数などには触れずに簡単な例のみに限定して説明した．

例えば，赤玉 2 個，白玉 1 個の入った袋から 1 個の玉を無作為に取り出すときの確率は $\frac{2}{3}, \frac{1}{3}$ であるから，この事象のエントロピー H は次のようになる．

$$H = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = \frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3$$
$$= \left(\frac{2}{3} + \frac{1}{3}\right)\log_2 3 + \frac{2}{3}\log_2\frac{1}{2} = \log_2 3 - \frac{2}{3} = 0.918$$

しかし，この程度の簡単な計算でも式変形に困難を感じる学生が一定数いた．本研究では，対数を全く用いずにエントロピーを定義する．このことにより，一般の学部の大学生，高校生，および数学に興味のある中学生にエントロピーの概念を説明することができる．

† 高知工業高等専門学校
Kochi National College of Technology

2. 本研究に至る経緯

大学入試や公務員試験によく出題される問題の 1 つに「最短経路の問題」がある．これは，例えば図 1 のような碁盤目状の街路があり，左下の点Oから右上の点Pまで後戻りせずに（つまり最短経路で）行く道順の数を問うものである．横方向を x 軸，縦方向を y 軸とするとPの座標は $(6,6)$ となる．OからPまでは 12 ステップかかるのでPまでの最短経路の数は $(a+b)^{12}$ の展開式中の a^6b^6 の係数と一致する．

最短経路の問題は 2 次元のランダムウォークと捉えることもできる．酔客が原点Oから出発し，分岐点では右または上に等確率で進む．k ステップまで来たときに，彼は途中で財布を落としたことに気がつく．途中の経路の記憶が全くないとすると，可能な経路の数が多いほど財布を見つけることが困難である．酔客が x 軸，または y 軸上にいるときはまっすぐ原点に戻ればよい．これはエントロピーが 0 の状態である．一般に，酔客の位置が直線 $y=x$ に近いときほど捜索は困難である．

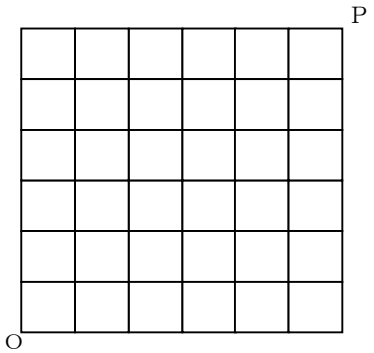

図 2　最短経路の問題

$x+y$ 個の玉の入った袋があり，そのうち x個および y 個がそれぞれ同じ色であったときにこの袋の持つエントロピー(平均情報量)を $H(x,y)$ で表わす．

数式で記述すると

$$H(x,y)=-\frac{x}{x+y}\log_2\frac{x}{x+y}-\frac{y}{x+y}\log_2\frac{y}{x+y}$$

となる．また $H(x,0)=0,\ H(0,y)=0$ と定める．以下，対数の底は 2 を用いるものとし，底は省略する．

$H(x,y)$ の値と点 (x,y) に至る最短経路の数の関係を調べてみよう．

表 1 は $x+y=6$ を満たす点 (x,y) までの最短経路の数と $H(x,y)$ の値の表である．

表 1　$H(x,y)$ の値と最短経路の数

x	y	最短経路の数	$H(x,y)$
6	0	1	0
5	1	6	0.45
4	2	15	0.64
3	3	20	0.69
2	4	15	0.64
1	5	6	0.45
0	6	1	0

点 (x,y) までの最短経路の数と $H(x,y)$ の値には強い正の相関関係があり，しかも点の位置が直線 $y=x$ に近いときほど $H(x,y)$ の値が大きいことがわかる．

以下，任意の 2 点 $(x_1,y_1),\ (x_2,y_2)$ について $H(x_1,y_1)$ と $H(x_2,y_2)$ の値の大小を調べることにする．

まず，任意の正の実数 k に対して

$$H(kx,ky)=H(x,y)$$

が成り立つので，原点を通る同一直線上にある 2 点のエントロピーは一致する．

次に，任意の自然数 x_1,y_1,x_2,y_2 に対し

$$H(x_1,y_1)=H(\,x_1(x_2+y_2),\ y_1(x_2+y_2)\,)$$
$$H(x_2,y_2)=H(\,x_2(x_1+y_1),\ y_2(x_1+y_1)\,)$$

が成り立つから $H(x_1,y_1)$ と $H(x_2,y_2)$ の値の大小は直線 $x+y=(x_1+y_1)(x_2+y_2)$ 上の 2 点 $(\,x_1(x_2+y_2),\ y_1(x_2+y_2)\,),(\,x_2(x_1+y_1),\ y_2(x_1+y_1)\,)$ における Hの値の大小と同じである．

次の定理は任意の正の数 k に対し，直線 $x+y=k$ 上の点 (x,y) に対する $H(x,y)$ の値の大小関係を示すものである．

<u>定理 1</u>　任意の自然数 x_1,y_1,x_2,y_2 に対し次が成り立つ．

(1) $|x_1-y_1|=|x_2-y_2|$ ならば $H(x_1,y_1)=H(x_2,y_2)$

(2) $|x_1-y_1|<|x_2-y_2|$ ならば $H(x_1,y_1)>H(x_2,y_2)$

（1）が成り立つことは明らかである．（2）の証明には数列 $\left\{\left(1+\frac{1}{n}\right)^n\right\}$ の単調性を用いる．証明は最後の補遺で述べる．

Vol.2010-CE-104 No.6
2010/5/22

3. エントロピーを表現する多項式

$N = a_1 + a_2 + \cdots + a_n$ 個の玉の入った袋があり，各 a_i 個が同じ色（色は全部で n 色）であったときにこの袋の持つエントロピーを $H(a_1, a_2, \cdots, a_n)$ で表わす．

このとき，

$$H(a_1, a_2, \cdots, a_n) = -\sum_{i=1}^{n} \frac{a_i}{N} \log \frac{a_i}{N} = -\sum_{i=1}^{n} \frac{a_i}{N} (\log a_i - \log N)$$

$$= \sum_{i=1}^{n} \frac{a_i}{N} \log N - \sum_{i=1}^{n} \frac{a_i}{N} \log a_i = \log N - \frac{1}{N} \sum_{i=1}^{n} a_i \log a_i$$

$x = \frac{1}{N}$ とおくと

$$H(a_1, a_2, \cdots, a_n) = N(\log N)x - \left(\sum_{i=1}^{n} a_i \log a_i \right) x$$

となる．具体例をいくつか見てみよう．

1. $H(1,1) = 2(\log 2)x$
2. $H(1,1,1) = 3(\log 3)x = 3\left(\log 2 + \log \frac{3}{2}\right)x$
3. $H(2,1) = 3(\log 3)x - 2(\log 2)x$
4. $H(1,1,1,1) = 4(\log 4)x = 4\left(\log 2 + \log \frac{3}{2} + \log \frac{4}{3}\right)x$
5. $H(3,1) = 4(\log 4)x - 3(\log 3)x = 4\left(\log 2 + \log \frac{3}{2} + \log \frac{4}{3}\right)x - 3\left(\log 2 + \log \frac{3}{2}\right)x$
6. $H(2,2) = 4(\log 4)x - 2(\log 2)x - 2(\log 2)x = 4\left(\log 2 + \log \frac{3}{2} + \log \frac{4}{3}\right)x - 2(\log 2)x - 2(\log 2)x$

エントロピーには一定のパターンがあることがわかる．そこで，この値を定数 $a, b, c, d, \cdots$ および変数 x を用いた多項式で表現する．

まず，定数の定義は次の通りである．

定義　$a = \log 2 = 1$，$b = \log \frac{3}{2}$，$c = \log \frac{4}{3}$，$d = \log \frac{5}{4}$，$\cdots$

定義　$H(a_1, a_2, \cdots, a_n)$の多項式表現を表 2 の通りに定める．

表 2　$H(a_1, a_2, \cdots, a_n)$ の多項式表現

総数	色のパターン	x	多項式表現
2	1+1	1/2	$2ax$
3	1+1+1	1/3	$3(a+b)x$
3	2+1	1/3	$3(a+b)x - 2ax$
4	1+1+1+1	1/4	$4(a+b+c)x$
4	3+1	1/4	$4(a+b+c)x - 3(a+b)x$
4	2+2	1/4	$4(a+b+c)x - 2ax - 2ax$
4	2+1+1	1/4	$4(a+b+c)x - 2ax$
5	1+1+1+1+1	1/5	$5(a+b+c+d)x$
5	4+1	1/5	$5(a+b+c+d)x - 4(a+b+c)x$
5	3+1+1	1/5	$5(a+b+c+d)x - 3(a+b)x$
5	3+2	1/5	$5(a+b+c+d)x - 3(a+b)x - 2ax$
5	2+2+1	1/5	$5(a+b+c+d)x - 2ax - 2ax$
5	2+1+1+1	1/5	$5(a+b+c+d)x - 2ax$

玉の総数が 6 個以上の場合も同じ規則に従って定めればよい．このとき次の定理が成り立つ．

定理 2　$a > b > c > d > \cdots$

定理 3　$a = b + c,\ b = d + e,\ c = f + g,\ \cdots$

証明

$$b + c = \log \frac{3}{2} + \log \frac{4}{3} = \log \frac{4}{2} = \log 2 = a$$

$$d + e = \log \frac{5}{4} + \log \frac{6}{5} = \log \frac{6}{4} = \log \frac{3}{2} = b$$

$$f + g = \log \frac{7}{6} + \log \frac{8}{7} = \log \frac{8}{6} = \log \frac{4}{3} = c$$

(Q.E.D.)

また，数列 $\left\{\left(1 + \frac{1}{n}\right)^n\right\}$ が単調に増加することから次の定理 4 が得られる．

定理 4　$a < 2b < 3c < 4d < 5e < \cdots$

以下にエントロピー計算の具体例を挙げる．

例 1

$$H(3,3) = 6(a+b+c+d+e)x - 3(a+b)x - 3(a+b)x = 6(c+d+e)x$$

定理 2 より $d+e=b, c+b=a$ であるから，

$$H(3,3) = 6ax = 6 \cdot 1 \cdot \frac{1}{6} = 1$$

即ち，赤玉 3 個と白玉 3 個の入った袋の持つエントロピーは 1 ビットである．

例 2

$$H(4,2) = 6(a+b+c+d+e)x - 4(a+b+c)x - 2ax = 2(b+c)x + 6(d+e)x$$

$$= 2ax + 6bx = \frac{1}{3} + b = \frac{1}{3} + \log\frac{3}{2} = 0.6365$$

4. 応用

次の表 3 は 2010 年 4 月現在の国内線の一部の路線の就航航空会社の表である．例えば，新潟と大阪を結ぶ路線は JAL，ANA，JAC の 3 社が参入していて，その便数の比は

JAL : ANA : JAC = 4 : 4 : 2 = 2 : 2 : 1

である．エントロピーを多項式で表わすと

$$5(a+b+c+d)x - 2ax - 2ax$$

となる．2+2+1=5 であるから変数 x の値は $\frac{1}{5}$ である．

表 3 国内線就航航空会社の表

発空港	着空港	JAL	ANA	JAC	SNA	JEX	IBX
新潟	大阪	4	4	2			
羽田	長崎	4	4		4		
福岡	宮崎	2		1		4	
福岡	五島福江		2				3
鹿児島	奄美大島	4		1			
鹿児島	那覇		1		2		

表 3 の他の路線についても計算すると表 4 のようになる．

表 4 航空路線とエントロピー

発空港	着空港	エントロピー	計算結果
新潟	大阪	$5(a+b+c+d)x - 2ax - 2ax$	$6ax + 5dx$
羽田	長崎	$3(a+b)x$	$3ax + 3bx$
福岡	宮崎	$7(a+b+c+d+e+f)x - 4(a+b+c)x - 2ax$	$4ax + 7bx + 7fx$
福岡	五島福江	$5(a+b+c+d)x - 2ax - 3(a+b)x$	$5ax - 3bx + 5dx$
鹿児島	奄美大島	$5(a+b+c+d)x - 4(a+b+c)x$	$2ax + 5dx$
鹿児島	那覇	$3(a+b)x - 2ax$	$ax + 3bx$

$a, b, c, \cdots$ の値を代入しなくても，定理 4 の不等式を用いるとエントロピーの大小を計算することができる．例えば，鹿児島と那覇を結ぶ路線のエントロピーは

$$ax + 3bx = \frac{1}{3}a + b = \frac{1}{3}a + c + d = \frac{5}{15}a + c + d$$

であり，鹿児島と奄美大島を結ぶ路線のエントロピーは

$$2ax + 5dx = \frac{2}{5}a + d = \frac{6}{15}a + d = \frac{5}{15}a + \frac{1}{15}a + d$$

である．ここで，定理 4 より $a < 3c$ だから $\frac{1}{15}a < c$ となる．このとき

$$\frac{2}{5}a + d < \frac{1}{3}a + b$$

であるので，鹿児島と那覇を結ぶ路線のエントロピーは，鹿児島と奄美大島を結ぶ路線のエントロピーより大きい．

5. 補遺

この章ではエントロピー関数 $H(a_1, a_2, \cdots, a_n)$

$$H(a_1, a_2, \cdots, a_n) = -\sum_{i=1}^{n} \frac{a_i}{N} \log\frac{a_i}{N} = \log N - \frac{1}{N}\sum_{i=1}^{n} a_i \log a_i$$

が $a_1 = a_2 = \cdots = a_n$ のとき最大になることを証明する．

補題　m, n を自然数とするとき，$n < m$ ならば $n\log\frac{n+1}{n} < m\log\frac{m+1}{m}$

証明　$n<m$ のとき $\left(1+\frac{1}{n}\right)^n<\left(1+\frac{1}{m}\right)^m$ だから両辺の対数を取ると

$$\log\left(1+\frac{1}{n}\right)^n<\log\left(1+\frac{1}{m}\right)^m$$
$$\therefore n\log\frac{n+1}{n}<m\log\frac{m+1}{m}$$

(Q.E.D.)

$a_1,a_2,\cdots,a_n$ のうち任意の 2 つを $x,\ y$ とする．$a_1,a_2,\cdots,a_n$ の順番を入れ替えても H の値は同じだから $a_1=x,\ a_2=y$ としても一般性を失わない．このとき，

$$H(x,y,a_3,\cdots,a_n)=\log N-\frac{1}{N}\sum_{i=3}^{n}a_i\log a_i-\frac{1}{N}(\,x\log x+y\ \log y\,)$$

であるから，$f(x,y)$ を

$$f(x,y)=x\log x+y\ \log y$$

とおく．

$f(x,y)$ に対して次の定理が成り立つ．

定理 5
(1) $f(x+1,x-1)>f(x,x)$
(2) $f(x+1,x)=\ f(x,\ x+1)$
(3) $y>x+1$ のとき

$$f(x-1,y+1)>\ f(x,y)>\ f(x+1,y-1)$$

証明　(1) $f(x+1,x-1)=(x+1)\log(x+1)+(x-1)\log(x-1)$

$$f(x+1,x-1)=(x+1)\sum_{i=1}^{x}\log\frac{i+1}{i}+(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}$$
$$=x\sum_{i=1}^{x}\log\frac{i+1}{i}+\sum_{i=1}^{x}\log\frac{i+1}{i}+(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}$$
$$=x\sum_{i=1}^{x-1}\log\frac{i+1}{i}+x\log\frac{x+1}{x}+\sum_{i=x-1}^{x}\log\frac{i+1}{i}+x\sum_{i=1}^{x-2}\log\frac{i+1}{i}$$
$$>x\sum_{i=1}^{x-1}\log\frac{i+1}{i}+x\log\frac{x}{x-1}+x\sum_{i=1}^{x-2}\log\frac{i+1}{i}$$
$$=x\sum_{i=1}^{x-1}\log\frac{i+1}{i}+x\sum_{i=1}^{x-1}\log\frac{i+1}{i}=f(x,x)$$

(3)

$$f(x,y)=x\sum_{i=1}^{x-1}\log\frac{i+1}{i}+y\sum_{i=1}^{y-1}\log\frac{i+1}{i}$$
$$=x\left(\sum_{i=1}^{x-2}\log\frac{i+1}{i}+\log\frac{x}{x-1}\right)+y\sum_{i=1}^{y-1}\log\frac{i+1}{i}$$
$$=(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}+\sum_{i=1}^{x-2}\log\frac{i+1}{i}+x\log\frac{x}{x-1}+y\sum_{i=1}^{y-1}\log\frac{i+1}{i}$$
$$=(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}+\sum_{i=1}^{x-1}\log\frac{i+1}{i}+(x-1)\log\frac{x}{x-1}+y\sum_{i=1}^{y-1}\log\frac{i+1}{i}$$

補題より

$$<(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}+\sum_{i=1}^{y}\log\frac{i+1}{i}+y\log\frac{y+1}{y}+y\sum_{i=1}^{y-1}\log\frac{i+1}{i}$$
$$=(x-1)\sum_{i=1}^{x-2}\log\frac{i+1}{i}+\ (y+1)\sum_{i=1}^{y}\log\frac{i+1}{i}=f(x-1,y)$$

一方，

$$f(x,y)=y\sum_{i=1}^{y-1}\log\frac{i+1}{i}+x\sum_{i=1}^{x-1}\log\frac{i+1}{i}$$

Vol.2010-CE-104 No.6
2010/5/22

$$= y\left(\sum_{i=1}^{y-2}\log\frac{i+1}{i}+\log\frac{y}{y-1}\right)+x\sum_{i=1}^{x-1}\log\frac{i+1}{i}$$

$$= (\mathrm{y}-1)\sum_{i=1}^{y-2}\log\frac{i+1}{i}+\sum_{i=1}^{y-2}\log\frac{i+1}{i}+y\log\frac{y}{y-1}+x\sum_{i=1}^{x-1}\log\frac{i+1}{i}$$

$$> (\mathrm{y}-1)\sum_{i=1}^{y-2}\log\frac{i+1}{i}+\sum_{i=1}^{x}\log\frac{i+1}{i}+x\log\frac{x+1}{x}+x\sum_{i=1}^{x-1}\log\frac{i+1}{i}$$

$$= (\mathrm{y}-1)\sum_{i=1}^{y-2}\log\frac{i+1}{i}+\sum_{i=1}^{x}\log\frac{i+1}{i}+x\sum_{i=1}^{x}\log\frac{i+1}{i}$$

$$= (\mathrm{y}-1)\sum_{i=1}^{y-2}\log\frac{i+1}{i}+(x+1)\sum_{i=1}^{x}\log\frac{i+1}{i}=f(x+1,y-1)$$

(Q.E.D.)

この定理 5 から，ただちに定理 6 が得られる．

定理 6　自然数 x,y に対し次が成り立つ
(1) $H(x+1,x-1,a_3,\cdots,a_n)<H(x,x,a_3,\cdots,a_n)$
(2) $H(x+1,x,a_3,\cdots,a_n)=H(x,\ x+1,a_3,\cdots,a_n)$
(3) $y>x+1$ のとき
$$H(x-1,y+1,a_3,\cdots,a_n)<H(x,y,a_3,\cdots,a_n)<H(x+1,y-1,a_3,\cdots,a_n)$$

$H(ka_1,ka_2,\cdots,ka_n)=H(a_1,a_2,\cdots,a_n)$ であるから，$N=a_1+a_2+\cdots+a_n$ をnの倍数になるようにとることにする．このとき，定理 6 より $H(a_1,a_2,\cdots,a_n)$ は$a_1=a_2=\cdots=a_n=\frac{N}{n}$のとき最大となる．

H は連続関数であるから，シャノンのエントロピー H

$$H=-\sum_{i=1}^{n}p_i\log_2 p_i$$

は $p_1=p_2=\cdots=p_n=\frac{1}{n}$ のとき最大になることがわかる．

参考文献

1) Claude E. Shannon: A Mathematical Theory of Communication, Bell System Technical Jounal, 27:379-423, 623-656, 1948