情報理論

第14回　講義

エントロピーと情報量

2015. 7. 22

植松　芳彦

本日の講義内容

- 情報量の定義
  - 平均符号長の下限としての情報量
  - 直感的な立場から見た情報量
- エントロピーと情報量
  - 「あいまいさ」の尺度としてのエントロピー
  - エントロピーの最小値／最大値

「情報源の符号化」とは何だったか？

- 元々の情報源系列が壊れない（受信側で一意復号可能）な前提で，最小限の量の符号（0, 1の並び）にして送る.
- 符号長（0，1の並びの数）は，元々の情報源系列に関わる何か重要な量を意味してないか？
- さらに平均符号長の下限値を与える「エントロピー」とは何を意味するか？

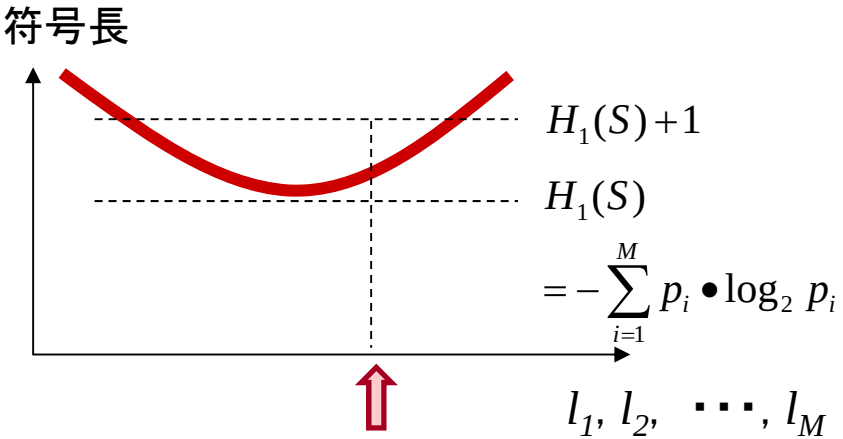
エントロピーについて知ってること(1)

- 情報源 S が発生する記号列を符号化する時の平均符号長の下限値を与える.
  - クラフトの不等式を満たす条件で符号長を最小化
- 情報源 S の統計的性質に依存した量

情報源記号　$\{a_1, a_1, \cdots, a_M\}$

発生確率　$P(a_i) = p_i \quad (i = 1, 2, \cdots, M)$

1次エントロピー　$H_1(S) = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i$



$-\log_2 p_1 \le l_1 < -\log_2 p_1 + 1$

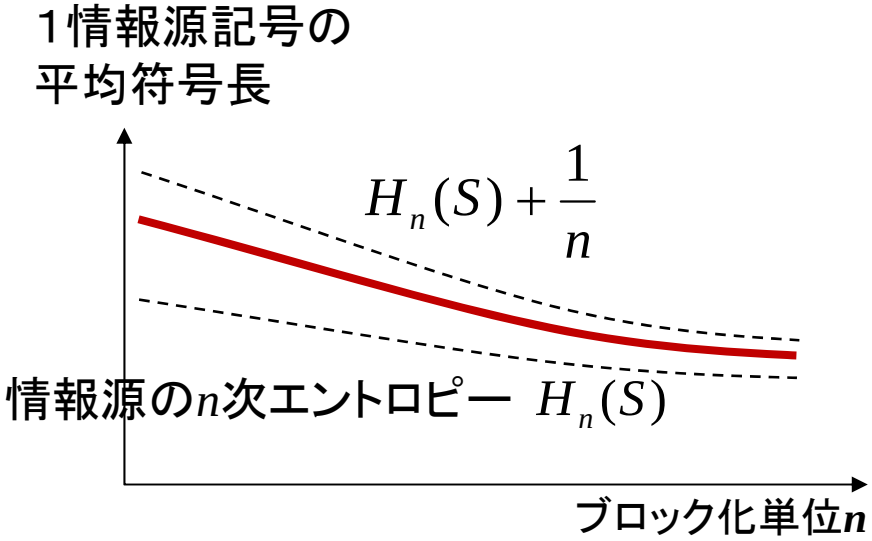
$\cdots$

$-\log_2 p_M \le l_M < -\log_2 p_M + 1$

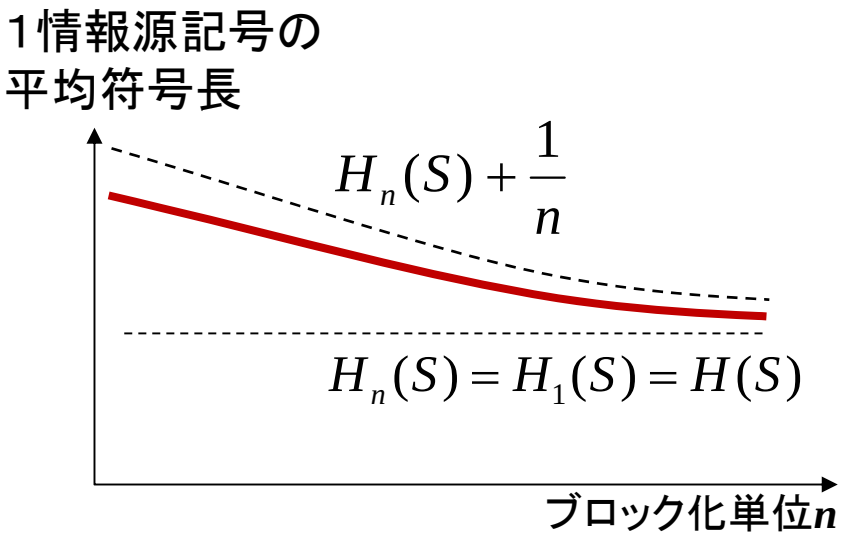
教科書p57-60
第8回講義資料

エントロピーについて知ってること(2)

- n 個の情報源記号を纏めて符号化することで，平均符号長は短くなる傾向
- 1情報源記号あたりの平均符号長の下限となる n 次エントロピーも小さくなる傾向



一般的な傾向



記憶のない情報源の場合

エントロピーの意味に関する仮説

- 情報源記号列が持っている情報の価値，またはその裏返しとして情報を知らない場合の曖昧さのようなものを表してないか？
  - 情報の意味が壊れない範囲で相手に伝えられる最小のビット数

直感的な立場からの情報量

- ある情報源から確率的に(統計的性質をもって)情報が発生する場合の情報量を考える.
- 情報量はその情報が発生する確率に依存すべき.
  - 確率1で発生する情報から得られる情報量はゼロ
    - 太陽が西に沈む
    - 犬が人を噛む
  - 非常に発生確率の低い情報から得られる情報量は非常に大きい
    - 新たに油田が発見される
    - 人が犬を噛む

直感的な立場からの情報量

- 情報源Sから情報 a_i が発生したことを知ることにより　得る情報量　$I(p_i)$　とは？

情報源 S

・各情報源記号の発生確率

情報源記号	発生確率
a_1	p_1
a_2	p_2
・・・	
a_M	p_M

情報量 $I(p)$ に対する条件

1．$I(p)$ は p の単調減少関数

2．$I(p_1 \cdot p_2) = I(p_1) + I(p_2)$

3．$I(p)$ は p の連続関数

上記の条件を満たす関数$I(p)$は以下の形のみ

$$I(p) = -a \cdot \log_2 p \qquad (式5.3)$$

$$I(p) = -\log_2 p \qquad (a = 1) \qquad (式5.4)$$

直感的な立場からの情報量

- 情報量 $I(p_i)$ の平均値とは？

$$\bar{I} = \sum_{i=1}^{M} p_i \cdot I(p_i) = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i \qquad \text{（式5.5）}$$

- 符号長と情報量の対応付け

情報源記号a_iに割当てる符号長の目安

$$-\log_2 p_i \le l_i < -\log_2 p_i + 1$$

平均符号長の下限（1次エントロピー）

$$H_1(S) = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i$$

⇔

情報 a_i の情報量

$$I(p_i) = -\log_2 p_i$$

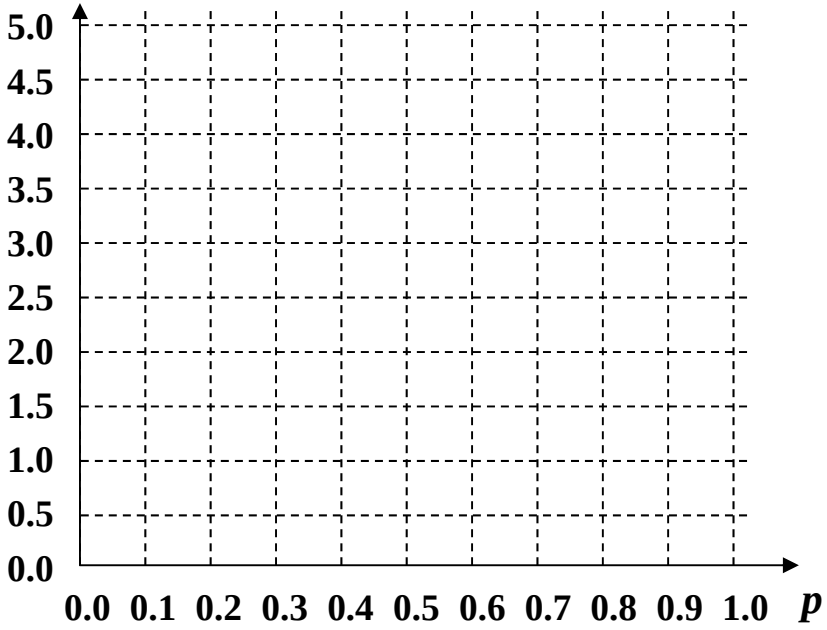
情報量の平均値

$$\bar{I} = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i$$

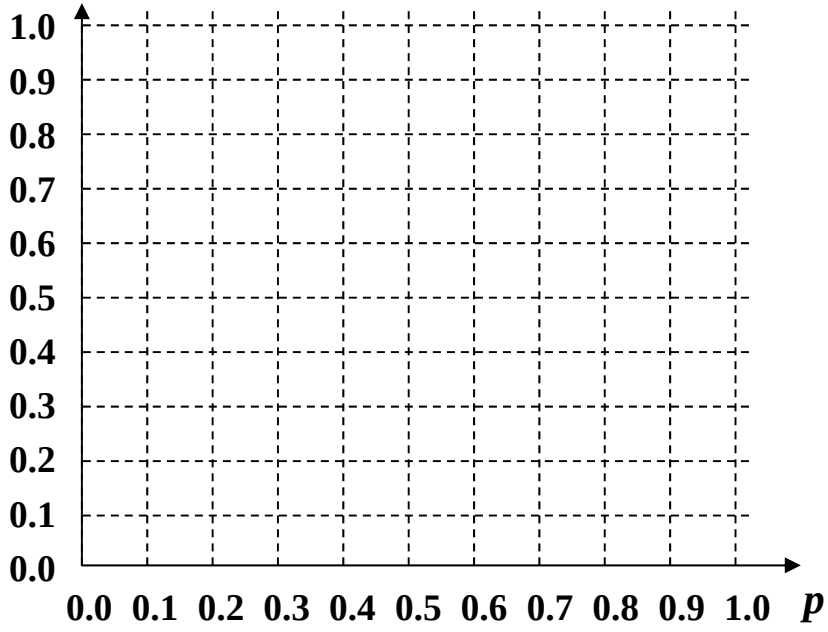
【演習1】 情報量の関数のかたち

- 情報量 $I(p)$, 情報量の平均値(期待値)の要素関数 $p \cdot I(p)$ のかたちを求めておこう

$$I(p) = -\log_2 p$$



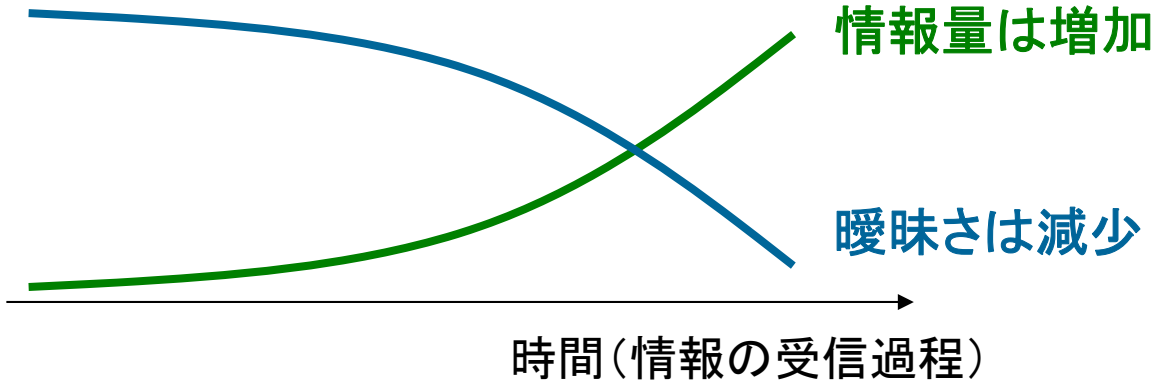
$$p \cdot I(p) = -p \cdot \log_2 p$$



p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−log2p	∞	3.3219	2.3219	1.7370	1.3219	1.0000	0.7370	0.5146	0.3219	0.1520	0.0000
ただし p=0 のとき p･log2p = 0 とする											

「あいまいさ」の尺度としてのエントロピー

- 元々熱力学における系の「無秩序さ」を表す尺度.
- 情報理論においても,「無秩序さ」を表す尺度.
  - ある時点の出力記号を知る以前における，情報の受け手の知識の「あいまいさ」
- 情報を得る過程の2つの捉え方
  - 受け手の知識の「あいまいさ」が減っていく過程.
  - 受けての情報量が増える過程.



エントロピーの最小値

- エントロピーが各情報源記号 a_i の発生確率 p_i の変化に伴いどのような値を取り得るか考察

情報源記号　$\{a_1, a_2, \cdots, a_M\}$

発生確率　$P(a_i) = p_i \quad (i = 1, 2, \cdots, M)$

1次エントロピー　$H_1(S) = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i = H(S)$　（記憶ない情報源）

- p_i は確率であることから　$0 \le H(S)$　（式5.8）
- 等号の成立は特定の p_j につき　$p_j = 1$

$p_k = 0 \quad (k \neq j)$

- どの記号が発生するか予め明らかなので，「あいまいさ」が全くない．

エントロピーの最大値

- 教科書p58の補助定理により

$$H(S) = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i$$

$$\leq -\sum_{i=1}^{M} p_i \cdot \log_2 \frac{1}{M} = \log_2 M \qquad \text{（式5.9）}$$

- 等号成立は　$p_1 = p_2 = \cdots = p_M = \frac{1}{M}$

- どの記号が発生する確率も等しく，どれが発生するか全く見当がつかない時「あいまいさ最大」（エントロピー最大）

【参考】補助定理

教科書p58-59
第 7 回講義資料

- $p_1, \cdots, p_M$（p_iは非負）に

$$p_1 + p_2 + \cdots + p_M = 1$$

- $q_1, \cdots, q_M$（q_iも非負）に

$$q_1 + q_2 + \cdots + q_M \leq 1 \qquad \text{（式4.9）}$$

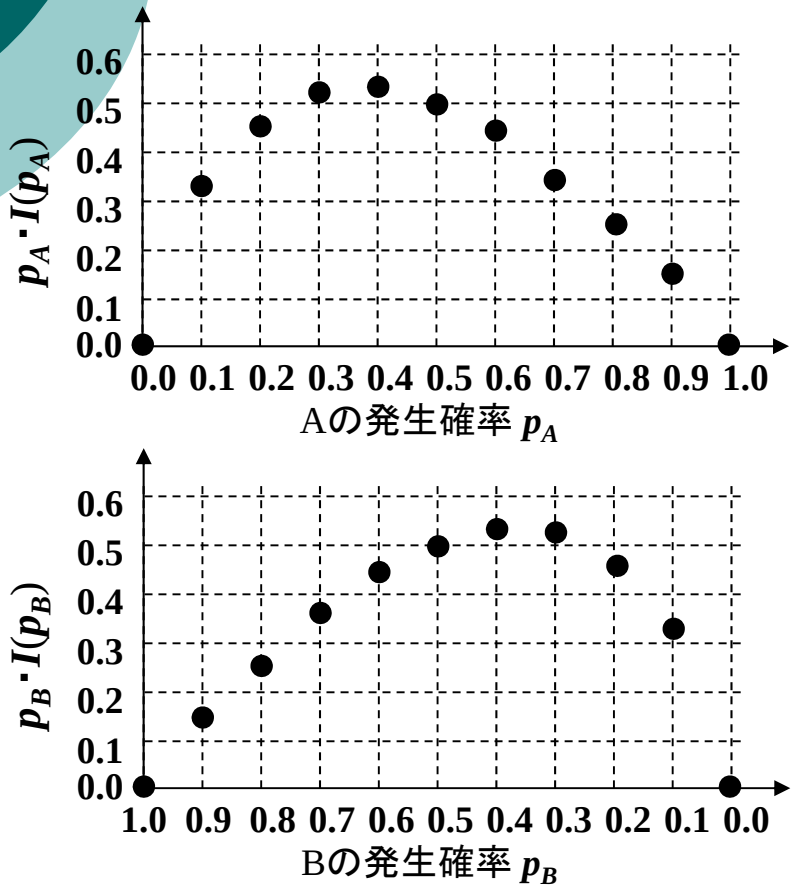
- が成り立つとき，以下の関係が成立．

$$H_1(S) = -\sum_{i=1}^{M} p_i \bullet \log_2 p_i \leq -\sum_{i=1}^{M} p_i \bullet \log_2 q_i \qquad \text{（式4.10）}$$

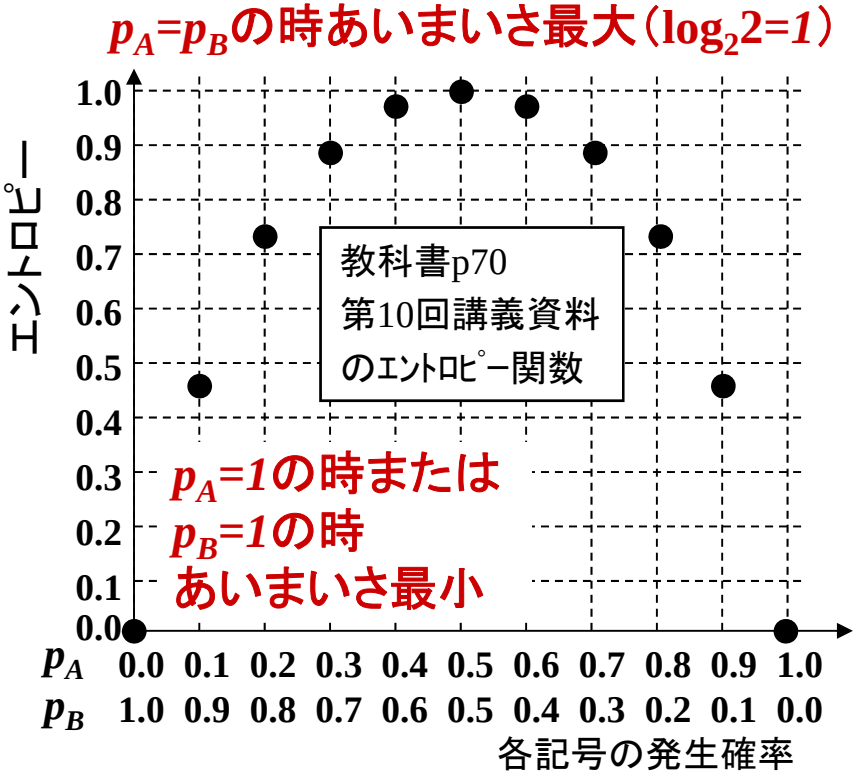
等号条件　$p_i = q_i \quad (i = 1, 2, \cdots, M)$

各記号の発生確率とエントロピー

- 情報源記号数M =2の場合（発生記号はA, B のみ）の，各記号の発生確率とエントロピーの関係.



$p_A+p_B=1$



記憶ない2元情報源のエントロピー

【演習2】各記号の発生確率とエントロピー

- 情報源記号数M =10（発生記号$a_1, \cdots, a_{10}$）場合の，各記号の発生確率とエントロピーの関係．

特定記号の発生確率が高い場合

$$p_1 = p_2 = \cdots = p_8 = 0.001$$

$$p_9 = 0.002, \quad p_{10} = 0.99$$

$$H(S) = -\sum_{i=1}^{10} p_i \cdot \log_2 p_i$$

$$=$$

すべての発生確率が同じ場合

$$p_1 = p_2 = \cdots = p_{10} = 0.1$$

$$H(S) = -\sum_{i=1}^{10} p_i \cdot \log_2 p_i$$

$$= -\sum_{i=1}^{10} 0.1 \cdot \log_2 0.1$$

$$=$$

p	0.001	0.002	0.01	0.02	0.1	0.9	0.99		M	10
−log2p	9.9658	8.9658	6.6439	5.6439	3.3219	0.1520	0.0145		log2M	3.3219

各記号の発生確率とエントロピー

- 情報源が「英文」の場合のエントロピーを試算
- 英文は記憶ある情報源のため，エントロピーは更に低い

各アルファベットの発生確率が同じ場合

$$p_A = p_B = \cdots = p_Z = 1/26$$

$$H(S) = \log_2 26 = 4.70$$

各アルファベットの発生確率に表5.1の偏りがある場合

$$H(S) = -\sum_{i=A}^{Z} p_i \cdot \log_2 p_i$$

$$= 4.17$$

実際の英文は記憶のある情報源でありエントロピー1.2程度と推定されてる

表 5.1 英文における文字の出現確率

文字	確率	文字	確率	文字	確率
A	8.29%	J	0.21%	S	6.33%
B	1.43	K	0.48	T	9.27
C	3.68	L	3.68	U	2.53
D	4.29	M	3.23	V	1.03
E	12.08	N	7.16	W	1.62
F	2.20	O	7.28	X	0.20
G	1.71	P	2.93	Y	1.57
H	4.54	Q	0.11	Z	0.09
I	7.16	R	6.90		

本日のまとめ

- これまで学んできた平均符号長，直感的な視点等を踏まえ，「情報量」を定義した．
  - 各情報源記号の情報量 $I(p_i) = -\log_2 p_i$
  - 情報量の平均値 $\bar{I} = -\sum_{i=1}^{M} p_i \cdot \log_2 p_i$
- エントロピーをその情報を受け取る前の知識のあいまいさと捉え直し，各記号の発生確率への依存性を考察．