第3章 エントロピー

3.1 ～ 3.3

動機

- 情報源Sの各シンボルで伝達される“情報の量”を定義したい
- “情報の量”として次の条件が自然に課される
  - シンボルが出易くなると情報として面白くないので、”情報の量”は減少する。特に確率1のときに”情報の量”は0
  - シンボルs_iと$s_j (i \neq j)$が独立ならば”情報の量”は足し合わさった結果になっている

情報の量$I(s)$

- 情報源Sのシンボルsの情報量を$I(s)$とすると
  - シンボルが出易くなると情報として
    面白くないので、”情報の量”は減少する。
    特に確率1のときに”情報の量”は0

    $$\Rightarrow p_i = 1 \Rightarrow I(s_i) = 0$$

  - シンボルs_iと$s_j$$(i \neq j)$が独立ならば”情報の量”は
    足し合わさった結果になっている(独立でなれば
    これよりも”情報の量”は小さくなる)

    $$\Rightarrow I(s_i s_j) = I(s_i) + I(s_j)$$

情報の量$I(s)$

- 実はこのような条件を見たす関数は、シンボルs_iの確率をp_iとすると
$$I(s_i) = -\log p_i$$
となる(演習問題3.7)
- 実際、
  - $p_i = 1 \Rightarrow -\log p_i = -\log 1 = 0$
  - シンボルの独立性から $\Pr(s_i s_j) = \Pr(s_i)\Pr(s_j)$ ゆえ、
$I(s_i s_j) = -\log p_i \cdot p_j = -\log p_i - \log p_i = I(s_i) + I(s_j)$

補足

- 底の取り方は重要ではない
  - $x = r^{\log_r x}$なので、$\log_s x = \log_s r \log_r x$
- 底が2のとき、情報の単位は**ビット(binary digit)**と呼ばれる
- 底rを強調するときは$I_r(s)$と表記

偏りのない硬貨による例

- 偏りのない硬貨の出目の確率は$\frac{1}{2}$なので、

$$I_2(s_i) = -\log_2 \frac{1}{2} = \log_2 2 = 1$$

- つまり、情報量の単位は偏りのない硬貨の出目における情報量と同じ

エントロピー

- 情報源Sが出す平均的な情報量は

$$H_r(S) \coloneqq \sum_{i=1}^{q} p_i I_r(s_i) = -\sum_{i=1}^{q} p_i \log_r p_i$$

- $H_r(S)$をSのr**元エントロピー(entropy)**という

エントロピーの補足

- 情報量と同じで底は重要ではない
  - $p_i = r^{\log_r p_i}$から$\log_s p_i = \log_s r \log_r p_i$となるので、

$$H_s(S) = -\sum_{i=1}^{q} p_i \log_s p_i = -\log_s r \underbrace{\sum_{i=1}^{q} p_i \log_r p_i}_{H_r(S)}$$

- 底rが明らかならば$H(S)$と記載する

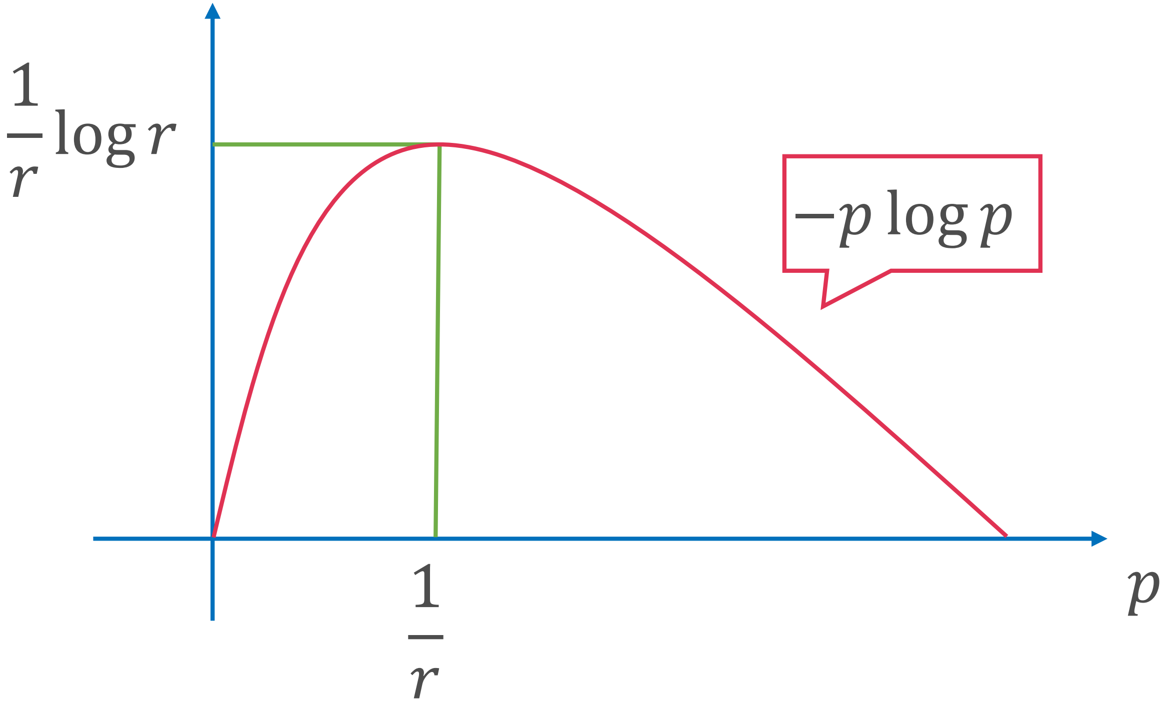
エントロピーの例

1. $-p\log_r p = \frac{-\log_r p}{{}^{1}/_{p}} \cdots (*)$
2. $x \to \infty$のとき、$x \geq \log x$から$\frac{\log x}{x} \to 0$
3. よって、$p \to 0 \Rightarrow \frac{1}{p} \to \infty$より$-\frac{-\log_r p}{{}^{1}/_{p}} = 0$

エントロピーの例

4. $(-p\log_r p)' = -\log_r p - 1$
5. $-\log_r p - 1 = 0 \Leftrightarrow \log_r p = -1$
6. よって$p = \frac{1}{r}$
7. また、$\frac{1}{r}$の付近で、
   - $p < \frac{1}{r}$のとき、$-\log_r p - 1 > 0$、つまり$(*)$は増加
   - $p > \frac{1}{r}$のとき、 $-\log_r p - 1 < 0$、つまり$(*)$は減少

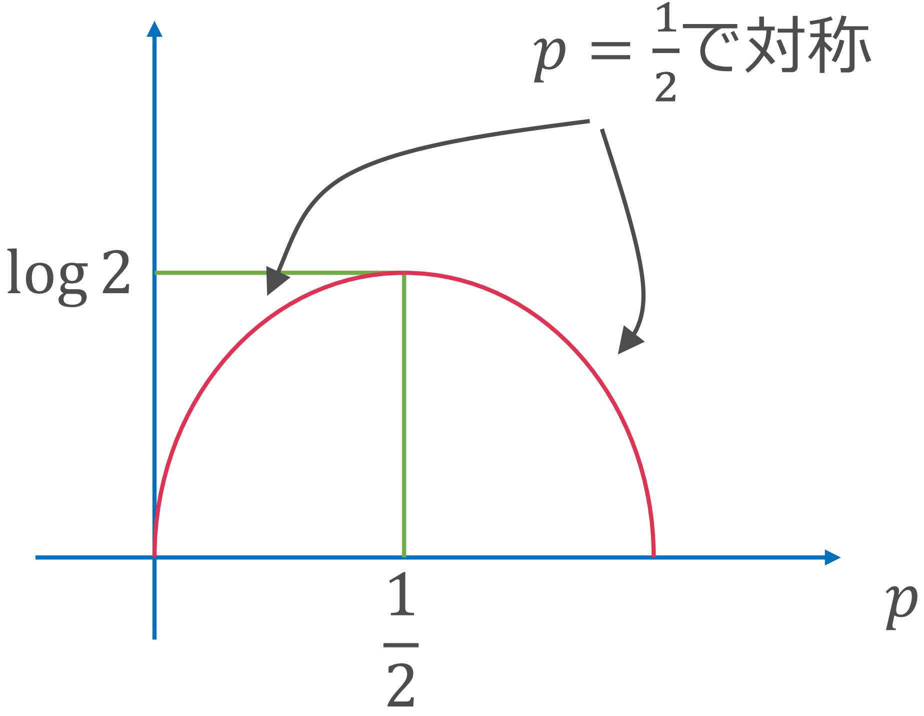
エントロピーの例



エントロピーの例

- Sを生起確率がpと$1-p(=\bar{p})$となるシンボルからなる情報源とする
- $H(S) = -p\log p - \bar{p}\log\bar{p}$

エントロピーの例



エントロピーの例

- $p = \frac{2}{3}$とする、つまり偏りがあるとすると
$$H_2(S) = \frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3 = \log_2 3 - \frac{2}{3} \approx 0.918$$
- $p = \frac{1}{2}$のときよりも情報量が少なくなる
- 実は偏りがあると情報量が小さくなる
    - 後で証明

エントロピーの例

- 偏りのないサイコロの2元エントロピーは
$$\log_2 6 \approx 2.586$$
- アルファベットの文字について一般に知られている出現頻度を用いるとエントロピーは4.03(らしい…)
- エントロピーは情報の”量”を表すのであって、メッセージの有用性については特に述べてない
  - エントロピーが低くてもランダムなアルファベットよりも人は小説を好む

エントロピーの性質

$H_r(S) \geq 0$であり、等号はあるiに対して
$p_i = 1$となるとき、かつその時のみをいう

- $p \log_r \frac{1}{p} \geq 0$から自明
  - 等号は$p = 1$か$p = 0$のときのみ成立

エントロピーの性質

- つまり、単一シンボルが常に発生したりするときなどの不確かさがないような状況だと情報は伝達されない
- 逆にエントロピーが最大となるのは？
    - これを調べる

補題

x_iと$y_i(i=1,\cdots,q)$を確率分布とする
(ただし、$y_i>0$)。すると、

$$\sum_{i=1}^{q} x_i \log_r \frac{1}{x_i} \le \sum_{i=1}^{q} x_i \log_r \frac{1}{y_i}$$

等号は全てのiについて$x_i=y_i$のとき、
かつそのときのみ成立

補題

$x_i \neq 0$ならば

$$
\begin{aligned}
&\sum_{i=1}^{q} x_i \log_r \frac{1}{x_i} - \sum_{i=1}^{q} x_i \log_r \frac{1}{y_i} \\
&= \sum_{i=1}^{q} x_i \log\left(\frac{y_i}{x_i}\right) \\
&= \frac{1}{\log_e r} \sum_{i=1}^{q} x_i \log_e \left(\frac{y_i}{x_i}\right) \left(\because \log_r x = \frac{\log_e x}{\log_e r}\right)
\end{aligned}
$$

補題

$$
\begin{aligned}
&\frac{1}{\log_e r}\sum_{i=1}^{q} x_i \log_e\left(\frac{y_i}{x_i}\right) \\
&\leq \frac{1}{\log_e r}\sum_{i=1}^{q} x_i\left(\frac{y_i}{x_i}-1\right) (\because \log_e x \leq x-1) \\
&= \frac{1}{\log_e r}\left(\sum_{i=1}^{q} y_i - \sum_{i=1}^{q} x_i\right) = 0
\end{aligned}
$$

補題

- 全てのiに対して$x_i \neq 0$のとき、等号が成立するのは
$$\sum_{i=1}^{q} x_i \log_r \left(\frac{y_i}{x_i}\right) = 0$$
から$\frac{y_i}{x_i} = 1 \Leftrightarrow x_i = y_i$のときかつそのときのみ
- $x_i = 0$のときは$x_i \log\left(\frac{1}{x_i}\right) = 0$から計算上無視できる

エントロピーの最大値

情報源Sがq個のシンボルを持つとき

$$H_r(S) \leq \log_r q$$

等号は全てのシンボルの生起確率が等しい

場合とき、かつその時のみ成立

エントロピーの最大値

- $x_i = p_i$、$y_i = \frac{1}{q}$とすれば補題の条件を満たすため、

$$H_r(S) = \sum_{i=1}^{q} p_i \log_r \frac{1}{p_i}$$

$$\leq \sum_{i=1}^{q} p_i \log_r q = \log_r q \sum_{i=1}^{q} p_i = \log_r q$$

- 等号は全てのiについて$p_i = 1/\mathrm{q}$のときみ

エントロピーの最大値

- 以上から、エントロピーが最大となるのは
  出てくるシンボルの出現の不確かさが
  最大のとき

エントロピーと平均符号長の関係

- 実は、次のような関係が証明できる

Cが情報源Sの一意復号可能なr元符号ならば、

$$L(C) \geq H_r(S)$$

エントロピーと平均符号長の関係

- Cの符号長を$l_1, \cdots, l_q$として$K \coloneqq \sum_{i=1}^{q} r^{-l_i}$とする
- マクミランの不等式(p.15)から$K \leq 1$
- $x_i = p_i$、$y_i = \frac{r^{-l_i}}{K}$とすると、$y_i > 0$かつ$\sum_i y_i = 1$となるので、補題を適用できる

エントロピーと平均符号長の関係

$$H_{r(S)} = \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{p_i}\right) \leq \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{y_i}\right) (\because \text{補題})$$

$$= \sum_{i=1}^{q} p_i \log_r\left(r^{l_i} K\right) = \sum_{i=1}^{q} p_i (l_i + \log_r K)$$

$$= \sum_{i=1}^{q} p_i l_i + \log_r K \sum_{i=1}^{q} p_i$$

エントロピーと平均符号長の関係

$\sum_{i=1}^{q} p_i = 1$なので、

$$\sum_{i=1}^{q} p_i l_i + \log_r K \sum_{i=1}^{q} p_i = L(C) + \log_r K$$

$K \leq 1$から$\log K \leq 0$となるので、

$$L(C) + \log_r K \leq L(C)$$

エントロピーと平均符号長の関係

- これは次のことを意味している
  - シンボルは平均として$H_r(S)$の情報量を持つため、1シンボルを送るたびに平均$H_r(S)$の情報量を送信していることとなる
  - すなわち、符号語Cの平均符号長は平均$H_r(S)$の情報量を送信するために、それ以上より長いか等しくなければならない
  - もし、 $H_r(S)$の情報量よりも短い場合、それはシンボルを送信できていないことに等しいと言える

$L(C) = H_r(S)$となる条件

Cが情報源Sの一意復号可能なr元符号とし、
p_iを$s_i \in S$の生起確率とする。
$L(C) = H_r(S)$となるものが存在するのは、
全ての$\log_r p_i$が整数、すなわち、ある
整数$e_i \leq 0$に対して$p_i = r^{e_i}$となるとき、
かつそのときのみである

$L(C) = H_r(S)$となる条件

- まず、$p_i = r^{e_i}$となるときに$L(C) = H_r(S)$となることを示す
- $p_i \leq 1$から$l_i \coloneqq -\log_r p_i \geq 1$
- 全てのiについて、$r^{l_i} = r^{\log_r \frac{1}{p_i}} = \frac{1}{p_i}$
- よって$\sum_{i=1}^{q} \frac{1}{r^{l_i}} = \sum_{i=1}^{q} p_i = 1$

$L(C) = H_r(S)$となる条件

- $\sum_{i=1}^{q} \frac{1}{r^{l_i}} = \sum_{i=1}^{q} p_i = 1$からマクミランの不等式の条件を満たしているので、$S$に対して一意復号可能な$r$元符号で符号長が$l_i$のものが存在する
- 平均符号長を計算すると

$$\sum_{i=1}^{q} p_i l_i = \sum_{i=1}^{q} p_i \log_{\mathrm{r}} \frac{1}{p_i} = H_r(S)$$

$L(C) = H_r(S)$となる条件

- $L(C) = H_r(S)$となるCが存在すると仮定して全てのiに対して$p_i = r^{e_i}$となる整数$e_i \leq 0$が存在することを示す
- エントロピーと平均符号長の関係を導出する際に用いたうちの2つの不等式は等号となる。

$L(C) = H_r(S)$となる条件

- すなわち

$$H_r(S) = \sum_{i=1}^{q} p_i \log_r \frac{1}{p_i} = \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{y_i}\right)$$

$$= \sum_{i=1}^{q} p_i \log_r \left(r^{l_i} K\right) = L(C) + \log_r K = L(C)$$

- 式から$\log_r K = 0 \Leftrightarrow K = 1$

$L(C) = H_r(S)$となる条件

- また、下の枠の中にある補題から$p_i = y_i$
- さらに$\sum_{i=1}^{q} p_i \log_r \frac{1}{p_i} = \sum_{i=1}^{q} p_i \log_r \left(r^{l_i} K\right)$から

$$p_i = \frac{r^{-l_i}}{K} = r^{-l_i} \Leftrightarrow \log_r p_i = -l_i$$

情報源Sがq個のシンボルを持つとき
$$H_r(S) \leq \log_r q$$
等号は全てのシンボルの生起確率が等しい
場合とき、かつその時のみ成立

$L(C) = H_r(S)$となる条件の補足

- ただし、このような場合は非常に限定的となる
- ほとんどは$L(C) > H_r(S)$

$L(C) = H_r(S)$となる例

- Sのシンボル$s_i (i = 1,2,3)$の生起確率がそれぞれ $p_1 = \frac{1}{4} = 2^{-2}, p_2 = \frac{1}{2} = 2^{-1}, p_3 = \frac{1}{4}$とすると、

$$H_r(S) = 2 \cdot \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 = \frac{3}{2}$$

で、符号を$C: s_1 \mapsto 00, s_2 \mapsto 1, s_3 \mapsto 01$とすればこの符号は一意復号可能で、

$$L(C) = 2 \cdot 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} = \frac{3}{2}$$

$L(C) = H_r(S)$とするには

- $p_i = 0$となるp_iがあれば、$L(C) > H_r(S)$でなければならない($p_i = r^{l_i}$となるl_iは存在しない)
- Sは生起確率$\frac{1}{2}, \frac{1}{2}, 0$からなるシンボルのみ持つ情報源とする。すると$H_2(S) = 2 \cdot \frac{1}{2} \log_2 2 = 1$
- このSの2元ハフマン符号の符号長は1,2,2

$L(C) = H_r(S)$とするには

- 2元ハフマン符号の符号長は
$$1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} + 2 \cdot 0 = 1.5$$
- しかし、生起確率0のものを削除すれば、
$$H_2(S) = 1$$
で、平均符号長は$1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1$

効率と冗長度

- Cが情報源Sのr元符号のとき、

$$\eta = \frac{H_r(S)}{L(C)}$$

  をCの**効率(efficiency)**という
  - $H_r(S) \leq L(C)$から$0 \leq \eta \leq 1$
- また、$1 - \eta$を**冗長度(redudancy)**という
- 効率が良ければ冗長度が減り、冗長度が良ければ冗長度が増えるという関係となる

演習問題3.1

情報源Sが生起確率

$0.3, 0.2, 0.15, 0.1, 0.1, 0.08, 0.05, 0.02$

を持つとしたときの$H_2(S), H_3(S)$を求め、

Sの2元ハフマン符号と3元ハフマン符号の

平均符号長と比較せよ

演習問題3.1

- 計算はCPython 3.8.0でやった https://wandbox.org/permlink/94PvuXuqAJNDRq9M
  - $H_2(S) \approx 2.681, H_3(S) \approx 1.691$
- Sの2元ハフマン符号と3元ハフマン符号の平均符号長はそれぞれ
$$L(C_2) = 2.72, L(C_3) = 1.77$$
(演習問題2.7参照)

演習問題3.2

$q \geq 2$に対し、q個のシンボルを持つ
情報源の生起確率の例を与え、そのSの
2元瞬時符号Cで、$L(C) = H_2(S)$を
実現するものを与えよ

演習問題3.2

- 生起確率を$2^{-1}, 2^{-2}, \cdots, 2^{-(q-2)}, 2^{-(q-1)}, 2^{-(q-1)}$とし、瞬時符号として、

$$C = \{0,10,110,\cdots,\underbrace{1\cdots10}_{q-2},\underbrace{1\cdots10}_{q-1},\underbrace{1\cdots11}_{q-1}\}$$

を考える

演習問題3.2

- このとき、

$$
\begin{aligned}
&H_2(S) \\
&= \sum_{i=1}^{q-1} 2^{-i} \log_2 2^{i} + 2^{-(q-1)} \log_2 2^{q-1} \\
&= \sum_{i=1}^{q} i2^{-i} + (q-1)2^{-(q-1)} \\
&= L(C)
\end{aligned}
$$

演習問題3.3

情報源Sの生起確率を$0.4, 0.3, 0.1, 0.1, 0.06, 0.04$
としたときのSのエントロピーを計算し、
Sの2元ハフマン符号との効率を計算せよ

演習問題3.3

- 計算はCPython 3.8.0でやった https://wandbox.org/permlink/R7ymefXdMmMmdQj1
  - $H_2(S) \approx 2.144$
- Sの2元ハフマン符号の符号長は$L(C) = 2.2$ (演習問題2.3参照)

演習問題3.3

- 以上から、効率は

$$\eta = \frac{2.144}{2.2} \approx 0.975$$