問題: 次の生起確率を持つ情報源
$\mathcal{S} = (S, P), S = \{a, b\}$ **について、**

S	a	b
P	0.8	0.2

(1) 2 **次の拡大情報源** $\mathcal{S}^2 = (S^2, P^{\otimes 2})$ **に対する Huffman 符号** $\mathcal{C}_2$ **を構成し、“**1 **文字当たりの平均符号長”** $L(\mathcal{C}_2)/2$ **を求めよ。**

(2) (時間に余裕があれば) 3 **次の拡大情報源** $\mathcal{S}^3 = (S^3, P^{\otimes 3})$ **に対しても同様の計算をせよ。**

n **次の拡大情報源** $\mathcal{S}^n = (S^n, P^{\otimes n})$
の符号 $\mathcal{C}_n$ **を利用すると、**

情報源 alphabet 1 **文字当たりの平均符号長**

$$\frac{L(\mathcal{C}_n)}{n}$$

はどこまで下げられるか？

$\longrightarrow$ **情報源** S **が本来持っている “情報の量”**
より小さくはならないだろう。

“エントロピー (entropy)”

“情報の量”

「或る事象 P が起こる」

という “情報の価値” は

どう評価したら良いか ？

「事象 P が起こる」という “情報の価値” $I(P)$

要請:

(1) 生起確率 p のみに依る

$$\longrightarrow I(p) := I(P)$$

(2) 独立な事象 P_1, P_2 に対して、

$$I(P_1 \wedge P_2) = I(P_1) + I(P_2)$$

$$\longrightarrow I(p_1 p_2) = I(p_1) + I(p_2)$$

(3) $I : [0,1] \longrightarrow \boldsymbol{R}_{\geq 0}$: 連続関数

$$\longrightarrow \boxed{I(p) = C \log \frac{1}{p} = -C \log p \quad (C \geq 0)}$$

$$I(p) = C \log \frac{1}{p} = -C \log p \quad (C \geq 0)$$

定数 C の取り方

$\longleftrightarrow$ $\log$ **の底の取り方**

$\longleftrightarrow$ **“情報量” の単位の取り方**

通常 2 を底に取って ($I(\frac{1}{2}) := 1$)、

“情報量” の単位とする: bit (binary digit)

情報源 $\mathcal{S} = (S, P)$ の
1 文字から得られる平均情報量

$$\boxed{H(\mathcal{S}) := \sum_{s \in S} P(s) I(P(s))}$$

：情報源 $\mathcal{S}$ の エントロピー (entropy)

$S = \{s_1, \ldots, s_k\}, P(s_i) = p_i$ **の時は、**

$$H(\mathcal{S}) := \sum_{i=1}^{k} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{k} p_i \log p_i$$

特に $\#S = 2$ の時、

- **起きる確率 p**
- **起きない確率 $1 - p =: \overline{p}$**

$$H(p) := H(\mathcal{S}) = p \log \frac{1}{p} + \overline{p} \log \frac{1}{\overline{p}}$$

：2 値エントロピー関数

問: $y = I(p), y = H(p)$ のグラフの概形を描け。
(最大値をとる p とその時の値は ?)

n **次の拡大情報源** $\mathcal{S}^n = (S^n, P^{\otimes n})$
の符号 $\mathcal{C}_n$ **を利用すると、**

情報源 alphabet 1 **文字当たりの平均符号長**

$$\frac{L(\mathcal{C}_n)}{n}$$

はどこまで下げられるか？

$\longrightarrow$ **情報源** $\mathcal{S}$ **のエントロピー** $H(\mathcal{S})$ **より**
小さくはならないだろう。

定理

$\mathcal{S} = (S, P)$: 情報源

$\mathcal{C}$: $\mathcal{S}$ の一意符号

$$\Longrightarrow \qquad \boxed{L(\mathcal{C}) \geq H(\mathcal{S})}$$

(但し、$\log$ の底は $r := \#T$ に取る。)

$\eta := \dfrac{H(\mathcal{S})}{L(\mathcal{C})}$: $\mathcal{C}$ の 効率 (efficiency)

$\overline{\eta} = 1 - \eta$: $\mathcal{C}$ の 冗長度 (redundancy)

補題

$x_i, y_i > 0 \ (i = 1, \ldots, k)$

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} y_i = 1$$

$$\Longrightarrow \qquad \sum_{i=1}^{k} x_i \log \frac{1}{x_i} \leq \sum_{i=1}^{k} x_i \log \frac{1}{y_i}$$

(等号は $\forall i : x_i = y_i$ のとき)

- $L(\mathcal{C}) = H(\mathcal{S})$ **は実現できるのか ？**

- $\inf_{\mathcal{C}} L(\mathcal{C}) = H(\mathcal{S})$ **であるか ？**

Huffman 符号は確かに最適だが、
上からの評価は難しい。

$\longrightarrow$ **Shannon-Fano 符号**
(Kraft-McMillan の不等式の利用)

Kraft の不等式

$S = \{s_1, \ldots, s_k\}, \qquad \#T = r$

自然数列 $(\ell_1, \ldots, \ell_k)$ **に対し、**

各符号語長 $|\mathcal{C}(s_i)| = \ell_i$ **なる**

r **元 瞬時符号 が存在**

$$\Longleftrightarrow \qquad \sum_{i=1}^{k} \frac{1}{r^{\ell_i}} \leq 1$$

定理

$\mathcal{S} = (S, P)$: **情報源**

$\mathcal{C}$: $\mathcal{S}$ **の最適符号**

$$\Longrightarrow \qquad \boxed{H(\mathcal{S}) \leq L(\mathcal{C}) < 1 + H(\mathcal{S})}$$

(但し、$\log$ **の底は** $r := \#T$ **に取る。)**

Shannon の第 1 定理 (Noiseless Coding Theorem)

$\mathcal{S} = (S, P)$: 情報源

$\mathcal{S}^n = (S^n, P^{\otimes n})$: $\mathcal{S}$ の n 次の拡大情報源

$\mathcal{C}_n$: $\mathcal{S}^n$ の最適符号

$$\Longrightarrow \qquad \boxed{\lim_{n\to\infty} \frac{L(\mathcal{C}_n)}{n} = H(\mathcal{S})}$$

(但し、$\log$ の底は $r := \#T$ に取る。)

情報源を効率良く符号化する話は一段落。

雑音の入る通信路を介して

これを誤りなく (効率良く) 伝達するには?

$\longrightarrow$ **符号理論 (誤り訂正符号)**

通常、我々が通信するとき、

一般には雑音が入って正しくは伝わらない。

- **それでも正しく伝えるにはどうするか？**
  $\longrightarrow$ 繰返し言う・別の言い方をする・$\cdots$

- **「何かおかしい」と気付けるのは何故か？**
  $\longrightarrow$ あり得ないから !!

「あり得ない」とは？

符号 $\mathcal{C} : S \longrightarrow T^+$ **で、**

受信した語 $y \in T^+$ **が** $y \notin \mathrm{Im}\mathcal{C}^*$

(誤り検出)

正しくは何だったのか？

y **に “一番近い”** $x \in \mathrm{Im}\mathcal{C}^*$ **だろう !!**

(誤り訂正)

受信語 $\boldsymbol{y} \notin \mathrm{Im}\mathcal{C}^* \subset T^*$ で誤り検出

$\longrightarrow$ T^* 全部は使わない

$\longrightarrow$ 冗長度を持たせて誤り検出・訂正

とは言え

- **より効率良く (冗長度少なめ)**
- **より高い誤り対処性能を持つ**
  (誤りが沢山あっても大丈夫)

ものが望ましい。

以下では、

- **生起確率の違いを考慮しない**

- **等長符号のみを考える**
  (全ての符号語が同じ長さ)

効率良い情報源符号で符号化された文字列を、
一定の個数毎に切って、再符号化 (通信路符号)

符号 $\mathcal{C} : S \longrightarrow V := T^n$ **(**n**：符号語長)**

の像 $\mathrm{Im}\mathcal{C} =: U \subset V$ **のみが大事**

$\longrightarrow$ **寧ろ、**

像 U **をも単に** $\mathcal{C}$ **と書き、符号と呼ぶ。**

符号 $\mathcal{C} \subset V = T^n$ で、

- **受信語 $y \notin \mathcal{C}$ によって誤り検出**
- **y に “一番近い” $x \in \mathcal{C}$ が正しい、として誤り訂正**

$\longrightarrow$ **“一番近い” とは？**

$\longrightarrow$ **V に “距離” を導入**
(通常 Hamming 距離 を用いる)