“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**
(const. 0 **ではない)**

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

「事象 P が起こる」という “情報の価値” $I(P)$

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

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)

情報源のエントロピー

情報源 $\mathcal{S} = (S, P)$ の
1 文字から得られる情報量の期待値

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

: 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、**

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

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

$\longrightarrow$ **まず一般に、符号** $\mathcal{C}$ **に対し、**
平均符号長 $L(\mathcal{C})$ **と** $H(\mathcal{S})$ **とを比べよう**

定理

$\mathcal{S} = (S, P)$：情報源

$\mathcal{C}$：$\mathcal{S}$ の一意符号 (瞬時符号)

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

(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)**

- $L(\mathcal{C}) = H(\mathcal{S})$ **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、
$L(\mathcal{C})$ **の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(再掲)

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

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**

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

Shannon-Fano 符号

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

各 i **に対し、** $\ell_i := \left\lceil \log_r \left(\frac{1}{p_i} \right) \right\rceil$

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

$\longrightarrow |\mathcal{C}(s_i)| = \ell_i$ **となる** r **元瞬時符号が存在**

この $\mathcal{C}$ **について、**

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

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

次の話題:

情報通信にはノイズ (雑音) が妨げとなる

雑音の入る通信路を介して情報通信を行なう際、

通信途中での誤りに如何に対処するか?

誤りに対処しつつ、如何に効率的に伝達するか?

$\longrightarrow$ **符号理論 (誤り訂正符号)**

誤り訂正符号 (お話)

通信途中でのノイズ (雑音) による誤りに
どう対処するか

- **誤りの発生を防ぐ (物理技術による対処)**
  - ⋆ **導線の品質を高める**
  - ⋆ **ノイズを遮蔽する　　等々**
- **誤りが起きても致命的にならないように**
  (フェイルセーフの発想・社会技術)
- **多少の誤りなら検出・訂正できる仕組み**
  $\longrightarrow$ **数理技術により実現 (誤り訂正符号)**

誤り訂正符号 (お話)

情報通信中に
誤り (らしきこと) に出遭ったときに
どうするか ？

例: 「じゃあヨツバ駅で待ち合わせね」

- 聞き直す (より安全なプロトコルの採用)
- 見当を付ける (誤りの自動訂正)

誤り訂正符号 (お話)

情報通信中に
誤り (らしきこと) に出遭ったときに
どうするか ？

例: 「じゃあヨツバ駅で待ち合わせね」

- **聞き直す (より安全なプロトコルの採用)**
- **見当を付ける (誤りの自動訂正)**

誤り訂正符号 (お話)

情報通信中に
　　誤り (らしきこと) に出遭ったときに
　　どうするか ？

例:　「じゃあヨツバ駅で待ち合わせね」

- 聞き直す (より安全なプロトコルの採用)
- 見当を付ける (誤りの自動訂正)

誤り訂正符号 (お話)

情報通信中に
誤り (らしきこと) に出遭ったときに
どうするか ？

例: 「じゃあヨツバ駅で待ち合わせね」

- **聞き直す (より安全なプロトコルの採用)**
- **見当を付ける (誤りの自動訂正)**

誤り訂正符号 (お話)

例: 「じゃあヨツバ駅で待ち合わせね」

$\longrightarrow$ きっとヨツヤ駅だろう

- 何故誤りだと気付くことが出来るのか

  $\longrightarrow$ 「ヨツバ」という駅がないから

- 何故正しく見当を付けることが出来るのか

  $\longrightarrow$ 似た名前の駅が他にないから

$\longleftarrow$ 文字列の殆どは駅名ではない

誤り訂正符号 (お話)

例: 「じゃあヨツバ駅で待ち合わせね」
$\longrightarrow$ きっとヨツヤ駅だろう

- 何故誤りだと気付くことが出来るのか
  $\longrightarrow$ 「ヨツバ」という駅がないから
- 何故正しく見当を付けることが出来るのか
  $\longrightarrow$ 似た名前の駅が他にないから

$\longleftarrow$ 文字列の殆どは駅名ではない

誤り訂正符号 (お話)

例: 「じゃあヨツバ駅で待ち合わせね」
$\longrightarrow$ きっとヨツヤ駅だろう

- 何故誤りだと気付くことが出来るのか
  $\longrightarrow$ 「ヨツバ」という駅がないから
- 何故正しく見当を付けることが出来るのか
  $\longrightarrow$ 似た名前の駅が他にないから

$\longleftarrow$ 文字列の殆どは駅名ではない

誤り訂正符号 (お話)

例: 「じゃあヨツバ駅で待ち合わせね」
$\longrightarrow$ きっとヨツヤ駅だろう

- 何故誤りだと気付くことが出来るのか
$\longrightarrow$ 「ヨツバ」という駅がないから

- 何故正しく見当を付けることが出来るのか
$\longrightarrow$ 似た名前の駅が他にないから

$\longleftarrow$ 文字列の殆どは駅名ではない

誤り訂正符号 (お話)

全国には 1 万弱の駅があるらしい

組合せ上は五十音 3 文字で表現可能

$(50^3 = 125000 > 10000)$

全ての駅名が五十音 3 文字なら**効率**は良いが、

五十音 3 文字の組合せが

全て正しい駅名だったら、

さっきの「誤りの自動訂正」は出来なかった

$\longrightarrow$ 冗長性を利用して安全性を確保した

誤り訂正符号 (お話)

全国には 1 万弱の駅があるらしい

組合せ上は五十音 3 文字で表現可能

$(50^3 = 125000 > 10000)$

全ての駅名が五十音 3 文字なら**効率**は良いが、

五十音 3 文字の組合せが
全て正しい駅名だったら、

さっきの「誤りの自動訂正」は出来なかった

$\longrightarrow$ **冗長性**を利用して**安全性**を確保した

誤り訂正符号 (お話)

「聞き直す」ことが出来ない通信の場合は、
(例: 惑星探査機からの通信)

「誤りの自動訂正」が出来るように
予め適切に冗長性を持たせて通信する

- 誤り訂正性能は高く
- とは言えなるべく効率的に

$\longrightarrow$ 有限体上の線型代数・代数幾何などの
数理構造を利用

誤り訂正符号

誤り検出:

符号 $\mathcal{C}^* : S^* \longrightarrow T^*$ で、
受信した語 $y \in T^*$ が $y \notin \mathrm{Im}\mathcal{C}^*$ なら誤り

誤り訂正:

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

誤り訂正符号

受信語 $y \notin \mathrm{Im}\mathcal{C}^* \subset T^*$ 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

誤り訂正符号

受信語 $y \notin \mathrm{Im}\mathcal{C}^* \subset T^*$ 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

誤り訂正符号

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する
と想定 (通信路符号)

誤り訂正符号

符号

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

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、**
これを符号と呼ぶ： $\mathcal{C} \subset V$