# 1　符号化

## 1.1　アルファベットと言語

符号理論で用いるいくつかの言い回しを導入しておく.
有限集合 $A$ に対し, その元の有限列全体の成す集合

$$A^+ := \bigcup_{n\in\mathbb{N}} A^n$$

を考える. $A$ の元の事を**文字**, $A$ の事を**アルファベット**, $A^+$ の元の事を**語**と言う.

**例 1.** base, ball 等は $\{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e},\mathrm{f},\mathrm{g},\mathrm{h},\mathrm{i},\mathrm{j},\mathrm{k},\mathrm{l},\mathrm{m},\mathrm{n},\mathrm{o},\mathrm{p},\mathrm{q},\mathrm{r},\mathrm{s},\mathrm{t},\mathrm{u},\mathrm{v},\mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z}\}$ をアルファベットとする語である.

**例 2.** 0, 1, 01, 1011, 011011 等は $\{0,1\}$ をアルファベットとする語である.

**例 3.** ･･, －, ･･･, －－－ 等は $\{\cdot, -, \_\}$ をアルファベットとする語である.

語の結合を表現する二項演算子 $\cdot : A^+ \times A^+ \to A^+$ を次の様に定める. [1]

$$(a_1,\ldots,a_n)\cdot(b_1,\ldots,b_m) := (a_1,\ldots,a_n,b_1,\ldots,b_m)$$

**例 4.** base · ball = baseball, ball · base = ballbase (非可換なことに注意)

**例 5.** 01 · 1011 = 011011

**例 6.** ･･ · － = ･･ －

利便性の為に「足しても何も変化しない語」として**空語** $\phi$ を $A^+$ に追加したものを $A^*$ で表す[2]:

$$A^* := \{\phi\} \cup \bigcup_{n\in\mathbb{N}} A^n$$

$u,v,w \in A^*$ について

$$w = u\cdot v$$

である時 $u(v)$ は $w$ の**接頭語** (**接尾語**) であると言い
空語 $\phi$ は任意の語 $w$ に対し常に**接頭語** (**接尾語**) であるが特にそれ以外のものを**真接頭語** (**真接尾語**) と言う.

**例 7.** base は baseball の接頭語, ball は baseball の接尾語.

**例 8.** 0, 01 等は 011011 の接頭語, 11, 011 等は 011011 の接尾語.

語に含まれる文字の数

$$|(a_1,\ldots,a_n)| := n,\ |\phi| := 0$$

を**語長**と呼ぶ.

**例 9.** baseball の語長は 8.

**例 10.** 011011 の語長は 6.

$A^*$ の部分集合の事を**言語**と呼ぶ.

**例 11.** $\{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e},\mathrm{f},\mathrm{g},\mathrm{h},\mathrm{i},\mathrm{j},\mathrm{k},\mathrm{l},\mathrm{m},\mathrm{n},\mathrm{o},\mathrm{p},\mathrm{q},\mathrm{r},\mathrm{s},\mathrm{t},\mathrm{u},\mathrm{v},\mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z}\}^*$ のなかで “英単語” になるもの全を $L$ と書けば, $L$ は言語. baseball $\in L$ だが ballbase $\notin L$.

**例 12.** $\{41,42,43,44,45,46,47,48,49,4A,4B,4C,4D,4E,4F,50,51,52,53,54,55,56,57,58,59,5A\}$ は $\{0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F\}^*$ をアルファベットとする言語.

言語 $L_1$, $L_2$ が $L_1 \subset L_2$ を満たす時 $L_2$ は $L_1$ の**拡大**であるという.

[1] この二項演算について $A$ は (非可換な) 半群となる.
[2] $A^*$ は $\phi$ を単位元にもつモノイドとなる.

## 1.2 符号化の定義

$\Sigma$, $A$ を有限集合とする. $\Sigma^n$ から $A^+$ への写像 $\varphi$ を**符号化**と言う[3]. その像 $\varphi(L)$ は $A$ をアルファベットにもつ言語を成す事に注意せよ. $A$ の事を**符号アルファベット**と言う. $\varphi(L)$ の語を $\varphi$ の**符号語**という.

特に $\phi : \Sigma \to A^+$ が与えられれば

$$\Sigma^n \ni (x_1, x_2, \ldots, x_n) \mapsto \phi(x_1) \cdot \phi(x_2) \cdot \cdots \cdot \phi(x_n) \in A^+$$

によって任意の $\Sigma^n$ からの符号化があたえられる. 以後, 特に断りの無い限り符号化は $\Sigma$ からの写像を考える.

**例 13.**

| 情報源文字 ($\Sigma$) | モールス信号 $A = \{ \cdot, - \}$ | ASCII 符号 $A = \{0, \ldots, 9, A, \ldots, F\}$ | 冗長な符号 $A = \{A, \ldots, Z\}$ |
|---|---|---|---|
| a | ・－ | 41 | ALFA |
| b | －・・・ | 42 | BRAVO |
| c | －・－・ | 43 | CHARLIE |
| d | －・・ | 44 | DELTA |
| e | ・ | 45 | ECHO |
| f | ・・－・ | 46 | FOXTROT |
| g | ――・ | 47 | GOLF |
| h | ・・・・ | 48 | HOTE |
| i | ・・ | 49 | INDIA |
| j | ・――― | 4A | JULIETT |
| k | －・－ | 4B | KILO |
| l | ・－・・ | 4C | LIMA |
| m | －－ | 4D | MIKE |
| n | －・ | 4E | NOVEMBER |
| o | ――― | 4F | OSCAR |
| p | ・――・ | 50 | PAPA |
| q | ――・－ | 51 | QUEBEC |
| r | ・－・ | 52 | ROMEO |
| s | ・・・ | 53 | SIERRA |
| t | － | 54 | TANGO |
| u | ・・－ | 55 | UNIFORM |
| v | ・・・－ | 56 | VICTOR |
| w | ・―― | 57 | WHISKEY |
| x | －・・－ | 58 | XRAY |
| y | －・―― | 59 | YANKEE |
| z | ――・・ | 5A | ZULU |

有限集合 $\Sigma$ の元の数が同じであれば全射で自然に対応するから抽象的議論においては $\Sigma$ として具体的に何を選ぶかは重要で無い場合が多い. $\#\Sigma = m$ なる符号を $m$ **元符号**と呼ぶ.

## 1.3 符号長

符号語の長さが全て等しい時, その符号は**等長**であるという. 上述の例で言えば ASCII 符号が等長符号である.

**定義 1.** 符号 $\varphi : \Sigma \to A^+$ と $\Sigma$ を情報源アルファベットに持つ情報源 ($\Sigma$-値確率変数 $X$) について

$$E[|\varphi(X)|]$$

を平均符号長と言う.

明らかに等長符号の平均符号長はその長さと等しい.

**例 14.** $\Sigma = \{a, b, c, d\}$ を情報源アルファベットに持つ情報源

| $\Sigma$ | $p_\sigma$ |
|---|---|
| $a$ | 0.125 |
| $b$ | 0.125 |
| $c$ | 0.25 |
| $d$ | 0.5 |

[3] $\Sigma^n$ をアルファベットに持つ言語 $L \subset \Sigma^+$ からの写像に拡張することもある.

を考える. この情報源の情報量は

$$-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{2}\log_2\frac{1}{2}=\frac{3}{8}+\frac{3}{8}+\frac{2}{4}+\frac{1}{2}=1.75[\text{bit}].$$

この情報源に二種類の符号を考える.

| $\Sigma$ | $p_\sigma$ | $\varphi_1(\sigma)$ | $\lvert\varphi_1(\sigma)\rvert$ | $\varphi_2(\sigma)$ | $\lvert\varphi_2(\sigma)\rvert$ |
|---|---|---|---|---|---|
| $a$ | 0.125 | 00 | 2 | 0001 | 4 |
| $b$ | 0.125 | 01 | 2 | 001 | 3 |
| $c$ | 0.25 | 10 | 2 | 01 | 2 |
| $d$ | 0.5 | 11 | 2 | 1 | 1 |

符号 $\varphi_1$ の平均符号長は等長符号であるから 2 であり, 符号 $\varphi_2$ の平均符号長は

$$E[|\varphi_2(X)|]=4\times0.125+3\times0.125+2\times0.25+1\times0.5=1.875$$

になって, より短い.

## 1.4 復号化

符号化で符号語に変換された情報は, 再び元の情報に戻せる事が期待される.
つまり

$$\psi:\varphi(\Sigma^n)\to\Sigma^n$$

で $\psi\circ\varphi$ が恒等写像であるものが存在することを期待する ($\psi$ は**復号化**と呼ばれる) . その為には $\varphi$ が単射であることが必要十分であり, 符号化 $\varphi$ が単射でないとき**特異**, 単射のとき**非特異**という.

**例 15.** $\#\Sigma=m$ なるとき,

$$\#\{0,1\}^{[\log_2 m]}=2^{[\log_2 m]}\geq 2^{\log_2 m}=m$$

であるから非特異な長さ $[\log_2 m]$ の 2 元等長符号が存在する.

$\Sigma$ をアルファベットに持つ情報源に関して, 次の二つの事実を比較してみよう.

1. どんな分布についても, 情報量は $\log_2 m$[bit] を超えない.
2. どんな分布についても, (平均) 符号長が $[\log_2 m]$ の非特異な等長符号が存在する.

この関係から, (2 元符号の) 平均符号長と (2 を底とした) 情報量が近い値であるという予想を立てることができる. この予想は後に符号化定理と言う形で正当化される.
最大の情報量を与えない情報源については, 符号長の最適化のためには, 等長でない符号が必要となる.
等長でない符号の場合は, 実用的な復号化の為に非特異なだけでは不十分である. 例えば前述のモールス信号は非特異であるので

$$\text{a}\leftrightarrow\ \cdot\ -$$

の様に一文字の情報は復号できるが

$$\text{abc}\to\ \cdot\ -\ -\ \cdots\ -\ \cdot\ -\ \cdot\to\text{pur}$$

の様に複数の文字の情報は別の文字として解釈できてしまう.