§3 符号

送信したい文章を2進数で送ることを考える.

例えば math という文を送るために. アルファベットの集合を

$X=\{a,b,\cdots,z\}$ とすると $|X|=26$ で $2^4<26\leq 2^5$ より.

$a:00000$, $b:00001$, $\cdots$, $z:11001$ とすればよい.

このように文字の集合に対し. 2進数を対応させることを符号化という.

定義3.1. 単射の写像 $\sigma: X \longrightarrow \bigcup_{n\in\mathbb{N}}\{0,1\}^n$ を

Xの符号という.

σは自然に文の符号$\tilde{\sigma}$に拡張できる. 例えば.

$\tilde{\sigma}(ab)=\sigma(a)\sigma(b)=0000000001$

$\tilde{\sigma}(math)=\underbrace{01100}_{m}\underbrace{00000}_{a}\underbrace{10011}_{t}\underbrace{00111}_{h}$ である.

符号化した文を元に戻すことを復号という.

命題3.2 符号σが復号できるための必要十分条件は

$\tilde{\sigma}: \bigcup_{n\in\mathbb{N}} X^n \longrightarrow \bigcup_{n\in\mathbb{N}} \{0,1\}^n$ が単射であることである。

$\because$ 復号できる $\Leftrightarrow$ 文章をもとへ戻せる

$\Leftrightarrow$ $\tilde{\sigma}$を $\tilde{\sigma}: \bigcup X^n \longrightarrow \tilde{\sigma}(\bigcup X^n)$ と

制限したものに逆写像が存在

$\Leftrightarrow$ $\tilde{\sigma}$が単射 //

例 復号できない例として

$\sigma: a \mapsto 0,\ b \mapsto 1,\ c \mapsto 01$ がある。このとき

符号化された文(符号語) 01 の復号は ab か c かわからない

$\rightsquigarrow$ $\tilde{\sigma}(ab) = 01 = \tilde{\sigma}(c)$ より $\tilde{\sigma}$が単射でない。

以下では $\tilde{\sigma}$が単射なもののみ符号という。

定義3.3 $\sigma: X \rightarrow \bigcup_{n\in\mathbb{N}} \{0,1\}^n$: 符号とする。$x \in X$に対し

$\sigma(x) \in \{0,1\}^k$ であるとき、$\sigma(x)$の符号語長は kである

という。

$|\sigma(x)| = k$ で表す. $\tilde{\sigma}$ についても同様に定義する.

定義 3.4.

${}^{\forall}x \in X$ に対し $|\sigma(x)|$ が一定である符号を等長符号という.

命題 3.5.

等長符号は復号可能

(∵) $x = x_1 \cdots x_k$, $y = y_1 \cdots y_l \in \bigcup_{n \in \mathbb{N}} X^n$ とし. $\tilde{\sigma}(x) = \tilde{\sigma}(y)$ とする.

また $|\sigma(x_i)| = m$ とすると.

$|\tilde{\sigma}(x)| = mk$, $|\tilde{\sigma}(y)| = ml$ より $k = l$. さらに.

$\tilde{\sigma}(x) = \sigma(x_1)\sigma(x_2)\cdots\sigma(x_k)$

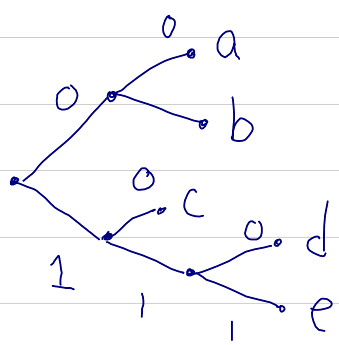
$\tilde{\sigma}(y) = \sigma(y_1)\sigma(y_2)\cdots\sigma(y_k)$ より $\sigma(x_i) = \sigma(y_i)$

σ は単射より $x_i = y_i$ ∴ $x = y$ //.

定義 3.6 どの符号語 $\sigma(x)$ も他の符号語 $\sigma(y)$ の接頭語になっていないとき. σ をプレフィックス符号という.

例. $\sigma(a)=01$　$\sigma(b)=\underbrace{01}_{\sigma(a)}1$ とすると. $\sigma(a)$ は $\sigma(b)$ の接頭語になっている.

例. プレフィックス符号は次のように作れる. $X=\{a.b.c.d.e\}$ のとき.



のようにすれば プレフィックスになる.

例えば. $\sigma(d)=110$ の接頭語は 11 と 1 だが.

これが他の符号語 $\sigma(x)$ になるには. 樹形図の途中が

x に対応するしかない. しかし. 作り方からそれはない

命題 3.7

プレフィックス符号は復号可能.

∵ $x=x_1\cdots x_k$, $y=y_1\cdots y_l \in \bigcup_{n\in\mathbb{N}} X^n$, $\tilde{\sigma}(x)=\tilde{\sigma}(y)$ とする.

$$\begin{aligned}\tilde{\sigma}(x)&=\sigma(x_1)\sigma(x_2)\cdots\sigma(x_k)\\ &\parallel\\ \tilde{\sigma}(y)&=\sigma(y_1)\sigma(y_2)\cdots\sigma(y_k)\end{aligned}$$

であるが.

$|\sigma(x_1)| \neq |\sigma(y_1)|$ であれば、どちらかがもう一方の接頭語になり
矛盾.

$\therefore |\sigma(x_1)| = |\sigma(y_1)|$ $\therefore \sigma(x_1) = \sigma(y_1)$ となり $x_1 = y_1$.

以下これを続ければ $x = y$ となる //

命題 3.8

$X = \{x_1, \cdots, x_n\}$ とし x_i が使用される確率を p_i とする.

このとき 符号語長の期待値 $L = \sum_{i=1}^{n} p_i |\sigma(x_i)|$ を

$S_2(p) \leq L \leq S_2(p) - 1$ とする符号 σ が存在する、

ただし $S_2(p) = \sum -p_i \log_2 p_i$

(∵) $p = (p_1, \cdots, p_n)$ に対し $p_1 \geq p_2 \geq \cdots \geq p_n > 0$ を仮定し、

$\frac{1}{2^{k_i - 1}} > p_i \geq \frac{1}{2^{k_i}}$ をみたす k_i をとる. この k_i は

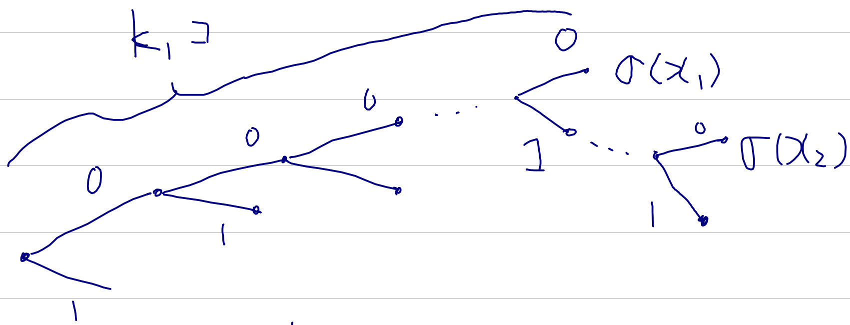
$k_1 \leq k_2 \leq \cdots \leq k_n$ をみたす また.

$\sum_{i=1}^{n} \frac{1}{2^{k_i}} \leq p_i \leq 1$ をみたすことに注意.

ここで σ を次のように作る.

まず. $|\sigma(x_i)|=k_i$ とし. 樹形図を上から埋める形で作る.

すなわち



としていく.

例えば. $\sigma(x_1)=\overbrace{0\cdots0}^{k_1\text{コ}}$ であるが. k_1ケタの2進数は 2^{k_1}コある

ので. 残り $2^{k_1}-1$コある. さらに $|\sigma(x_2)|=k_2\geq k_1$ より.

k_2ケタの2進数を考え, 符号語を1つわりふると. 残りは

$2^{k_2}-2^{k_2-k_1}-1$ である. 次に $\sigma(x_3)$ をやると残りは

$(2^{k_2}-2^{k_2-k_1}-1)(2^{k_3-k_2})-1$

$=2^{k_3}-2^{k_3-k_1}-2^{k_3-k_2}-1$ となる これを $\sigma(x_n)$ までやると.

$2^{k_n}-2^{k_n-k_1}-2^{k_n-k_2}-\cdots-1=2^{k_n}-2^{k_n}\left(\sum_{i=1}^{n}\frac{1}{2^{k_i}}\right)$

$\geq 2^{k_n}-2^{k_n}=0$ となる. ∴符号が作れる.

ここで $L=\sum p_i k_i$ だが. $\frac{1}{2^{k_i-1}}>P_i\geq\frac{1}{2^{k_i}}$ より

$-k_i+1\geq\log_2 P_i\geq -k_i$ なので.

$-\log_2 p_i \leq k_i \leq -\log_2 p_i + 1$ なので

$$\sum -p_i \log_2 p_i \leq \sum p_i k_i \leq \sum p_i(-\log_2 p_i + 1)$$
$$= -\sum p_i \log_2 p_i + 1.$$

$\therefore\ S_2(p) \leq L \leq S_2(p) + 1$ となる。//

$\rightsquigarrow$ 効率のよい符号 $\rightsquigarrow$ データ圧縮に応用.

定義3.9. 等長符号 σ を考える. $x, y \in X$ に対し,

$\sigma(x) = x_1 x_2 \cdots x_n$ ($x_i \in \{0,1\}$, y も同様) とするとき

x と y の距離を

$d(x,y) = \#\{i \mid 1 \leq i \leq n,\ x_i \neq y_i\}$ で定義する.

$\rightarrow d(\sigma(x), \sigma(y))$ とかくときもある.

これをハミング距離という.

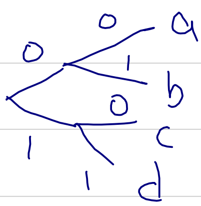
定義3.10 σ は上と同様とする.

$\sigma(x)$ に k 個以下の誤りがあるとする.

この誤りを誤りがあると判断できるとき. σ を k-誤り検出符号といい,

この誤りを訂正できるとき、k-誤り訂正符号という。

例(1) $X=\{a,b,c,d\}$ とし ϕ を



によって定義する。

00が1つ誤り01になると、01も符号語なので誤りを検出できない。

(2) $a:000$ $b:011$ $c:101$ $d:110$ で ϕ を定義すると1-誤り検出になる

<u>命題3.11.</u> $M=\min\{d(x,y)\mid x,y\in X,\ x\neq y\}$ とする。

(1) $k<M$ に対し、ϕ は k-誤り検出

(2) $k\le \frac{M-1}{2}$ に対し ϕ は k-誤り訂正

∵(1) $\phi(x)$ に $l\ (\le k)$ 個の誤りがあるとき、$l<M$ より

これは他の符号語にならない。∴誤りが検出できる

(2)

また、$\sigma(x)$ に l $(\leq k)$ 個の誤りがあり、それを x' とすると、

$d(\sigma(x), x') = l$ であるので $\forall y \in X$、に対し、

$d(\sigma(y), x') \geq M - l \geq l+1$ である。∴最も近い $\sigma(x)$ にすれば訂正できる。

例、$X = \{a, b, c, d, e, f, g, h\}$ とし、

a : 0000000
b : 0010111
c : 0101101
d : 0111010
e : 1001011
f : 1011100
g : 1100110
h : 1110001

$C_1 C_2 C_3 (C_1+C_2)(C_2+C_3)(C_1+C_3)(C_1+C_2+C_3)$

で σ を定義すると、$M=4$ より

3誤り検出、1誤り訂正