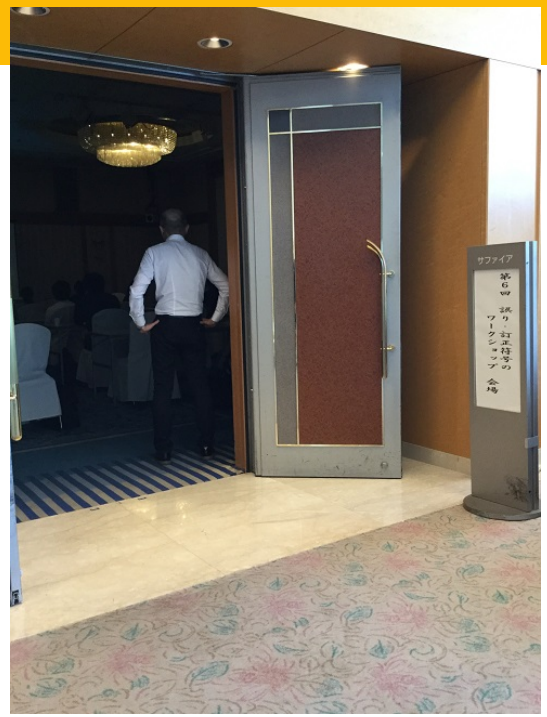
挿入削除誤り訂正符号の数学的に綺麗な性質について

Manabu HAGIWARA
hagiwara@math.s.chiba-u.ac.jp

Chiba University

2017/09, IEICE ソサエティ大会, 東京都市大学

動向



- DNA 解析
- 字句解析
- ECC-WS 2016, 2017
- ISIT2017 ... 11 papers
- DNA ストレージ

削除誤り

以下，Σ は集合.
削除誤り：

$$x_1x_2\dots x_{i-1}x_ix_{i+1}\dots x_n$$

の適当な (座標 i の) 成分が消えて

$$x_1x_2\dots x_{i-1}x_{i+1}\dots x_n$$

に写ること.

Example

00011111
$\to$ 0011111 もしくは 0001111 など.
前者は，$1 \le i \le 3$,
後者は $4 \le i \le 8$.

削除によって"空列でない"系列の長さは一つ"減る".

例

挿入誤り

挿入誤り：

$$x_1 x_2 \dots x_{i-1} x_i \dots x_n$$

の適当な座標 i と適当な成分 z により

$$x_1 x_2 \dots x_{i-1} z x_i \dots x_n$$

に写ること.

Example

系列のアルファベットが $\Sigma = \{0, 1\}$ のとき
系列 0111 は挿入誤りにより

$$00111, 01011, 01101, 01110, 10111, 01111$$

の 6 通りに変わる.

挿入誤り：補足

挿入誤り：

$$x_1 x_2 \dots x_i x_{i+1} \dots x_n$$

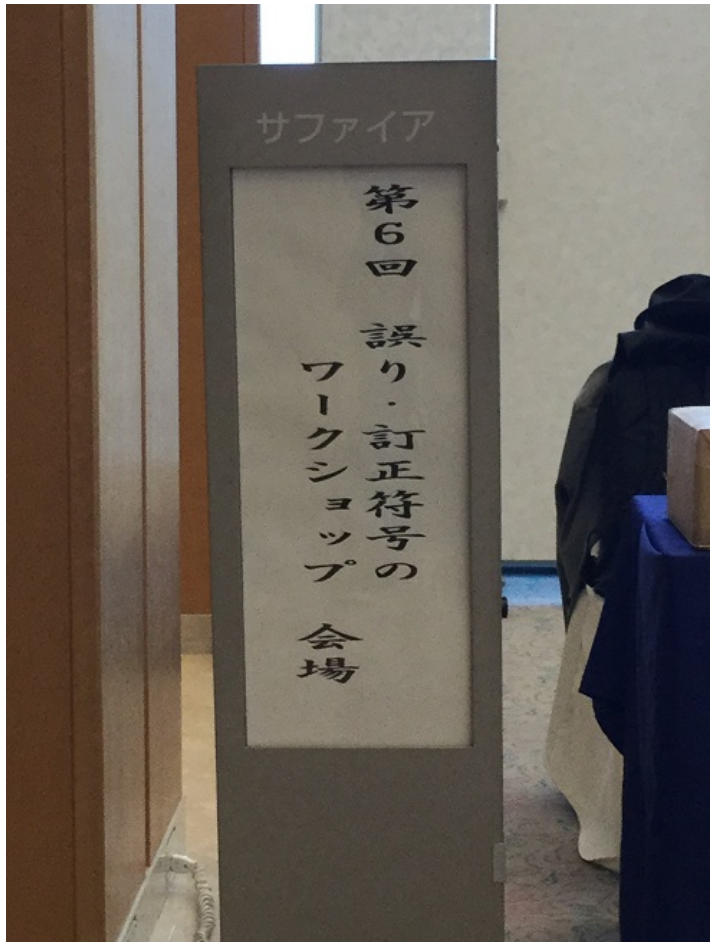
の適当な座標 i と適当な成分 z により

$$x_1 x_2 \dots x_i z x_{i+1} \dots x_n$$

に写ること.

としても OK.

例



t 重誤り

t 削除誤り：削除誤りをちょうど t 個続けたもの
t 重削除誤り：高々 t 削除誤り.

挿入誤りでも同様。

挿入／削除誤り（挿入と削除の組合せ）でも同様。

メモ：t 誤り訂正可能符号は、t 重誤り訂正可能符号ではない。
$C = \{111, 11\}$ は
2 削除誤り訂正符号だが
2 重削除誤りは必ずしも訂正できない。

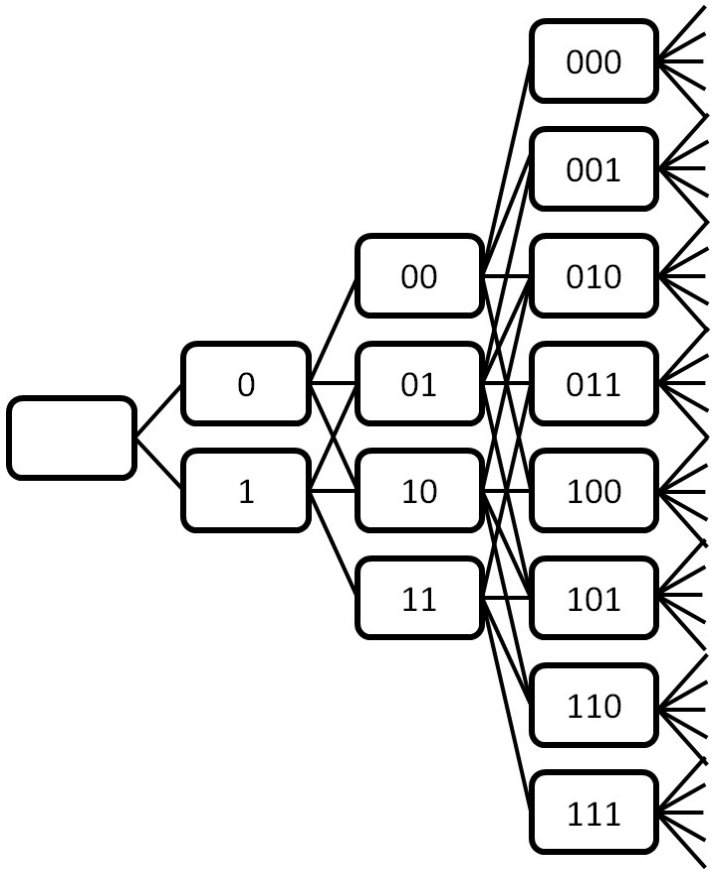
ビット反転誤り、代入誤り

ビット反転誤りは 2 挿入／削除誤りと解釈可.

例：0 を削除，そこに 1 を挿入
→ 0 から 1 へのビット反転.

$$000 \to 00 \to 010$$

Levenshtein 距離、編集距離



挿入／削除を辺にしたグラフで定まる距離を
Levenshtein 距離（もしくは編集距離）と呼ぶ.
$d_L(x, y)$ と書く。

一般のアルファベットでも

例：
economical
$\to$ economial
$\to$ economal
$\to$ conomal
$\to$ onomal
$\to$ nomal
$\to$ normal
から，

$$d_L(\textbf{economical}, \textbf{normal}) \leq 6$$

等号のためには，もっと少ない挿入／削除で economical から normal へ移れないことの証明が必要.

動的計画法による距離の言いかえ

·		e	c	o	n	o	m	i	c	a	l
	0	1	2	3	4	5	6	7	9	10	11
n	1	2	3	4	3	4	5	6	7	8	9
o	2	3	4	3	4	3	4	5	6	7	8
r	3	4	5	4	5	4	5	6	7	8	9
m	4	5	6	5	6	5	4	5	6	7	8
a	5	6	7	6	7	6	5	6	7	6	7
l	6	7	8	7	8	7	6	7	8	7	6

1 行目 :「·」「空白」そして系列を一文字ずつ.
1 列目 :「·」「空白」そして系列を一文字ずつ.
2 行目 :「空白」そして $1, 2, \ldots$.
2 列目 :「空白」そして $1, 2, \ldots$.
残りの要素 : $(1, j)$ 項目 $=$ $(i, 1)$ 項目 $\to$ (i, j) 成分 $:=$ $(i-1, j-1)$ 成分 $\neq \to$ $(i-1, j)$ 成分と $(i, j-1)$ 成分の小さいほうに 1 加えた値, としている.
一番右下の成分が系列間の Levenshtein 距離に一致.

Levenshtein 距離からわかること

Theorem

t 重削除誤り訂正符号は t 重挿入／削除誤り訂正符号.
t 重挿入誤り訂正符号は t 重挿入／削除誤り訂正符号.

次の考察からわかる：

t 重削除誤り訂正符号
⇔任意の符号語間の Levenshtein 距離が $2t+1$ 以上

挿入も同様.

挿入球の表面積

以降，Σ を有限集合.

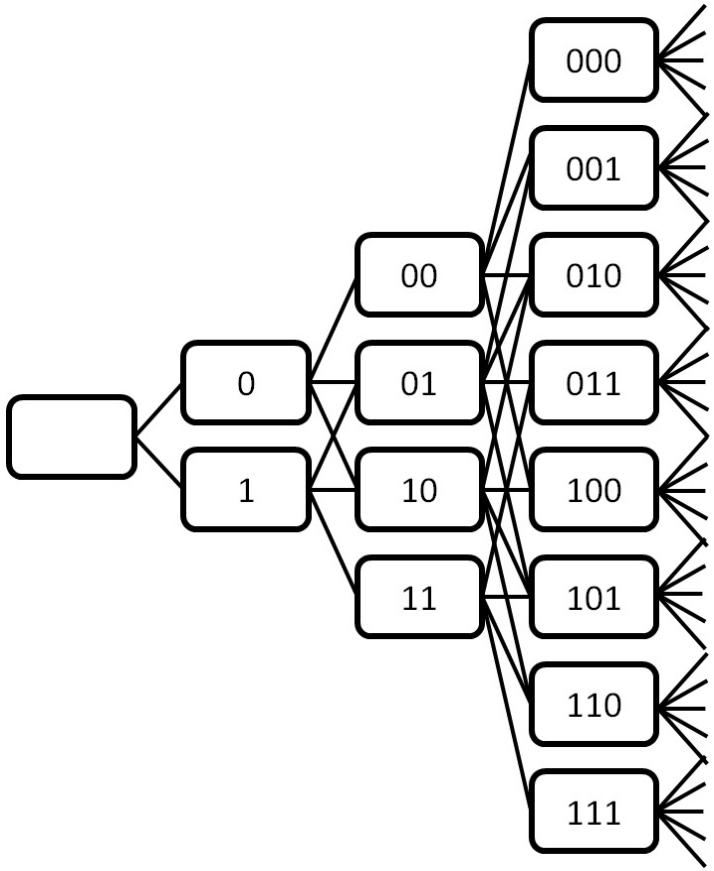
Theorem

n と m を正整数
系列 $y \in \Sigma^m$ に対し，
$B(y, n)$ を y に n 挿入で得られる系列全体.

このとき，n, m を固定した上で，
$B(y, n)$ の濃度（集合の要素数）は y に依らない.

$$|B(y, n)| = \sum_{k=0}^{n} \binom{m+n}{k} (q-1)^k$$

例：挿入球の表面積



右に出る辺の数に注目。

左はそうならない。

表面積の考え方：ビットで

証明は 2 ステップ。

1． y の長さ m にしか依らないことを帰納法で。
$B(y,n)$ を分割。もともとの y がどこに散らばったか。できるだけ左に寄せて捉える。
例：$y=000, n=5$
$00000010 \in B(y,n)$ のうち y は左の 3 ビットと一意に考えられる。

y_1 の位置に注目。その左は y_1+1 の繰り返し（一通り）。
例：$B(10,2)=$
$\{1000,1001,1010,1011,1100,1101,1110\} \sqcup \{0100,0101,0110\} \sqcup \{0010\}$

$y_2 \ldots$ に対して帰納法の仮定を適用。

2． $B(000\ldots0,n)$ を求める。
$B(y,n)$ は 1 が高々 n 個までの $n+m$ ビット全体。

$$|B(y,n)| = \sum_{k=0}^{n} \binom{m+n}{k}$$

例：挿入球の表面積

Example

$\Sigma = \{0, 1, 2\}$ とする． $B(00, 2) = \{0000\}$
$\sqcup \{1000, 2000, 0100, 0200, 0010, 0020, 0001, 0002\}$
$\sqcup \{1100, 1200, 2100, 2200, 1010, 1020, 2010, 2020, 1001, 1002,$
$2001, 2002, 0110, 0120, 0210, 0220, 0101, 0102, 0201, 0202,$
$0011, 0012, 0021, 0022\}$
とくに

$$|B(00, 2)| = 29$$

である．一方，

$$\sum_{k=0}^{2} \binom{4}{k} 2^k = \binom{4}{0} + \binom{4}{1} 2 + \binom{4}{2} 2^2 = 1 + 4 + 24 = 29$$

である．

線形符号と非線形符号

Theorem

C：有限体上の $[n, k]$ 線形符号

符号化率 $k/n > 1/2$
$\Rightarrow$ C には訂正できない 1 削除誤りがある.

1 重削除誤りを訂正できる符号
⇒符号化率 $\leq 1/2$.

一方，次で述べる Levenshtein 符号は
非線形な 1 重削除誤り訂正符号であり，
その符号化率は符号長を伸ばすと 1 に収束.

おまけ：予想『t 重削除誤り訂正 $[n, k]$ 符号は $k/n < 1/(t+1)$』

Levenshtein 符号 (VT 符号)

Definition (Levenshtein 符号)

$$L_{n,a} := \{\mathbf{x} \in \{0,1\}^n \mid x_1 + 2x_2 + \cdots + nx_n \equiv a \pmod{n+1}\}$$

Theorem

Levenshtein 符号は 1 重挿入／削除誤り訂正符号.

おまけ：法を $2n+1$ にすると 1 ビット反転も直せちゃう。

Example

$L_{5,0}$ は次の 6 つの系列からなる符号である：

$$\{00000, 11100, 00111, 10001, 01010, 11011\}$$

各系列に対し，削除前，削除後を列挙する．

- 00000 : 0000
- 11100 : 1100, 1110
- 00111 : 0111, 0011
- 10001 : 0001, 1001, 10000
- 01010 : 1010, 0010, 0110, 0100, 0101
- 11011 : 1011, 1111, 1101

以上から，1 重削除誤り生成可能．
さらに 1 重挿入／削除誤り訂正符号．

Levenshtein 符号の訂正能力の捉え方

挿入結果を

$x_1x_2x_3\ldots x_{n-1}x_n0$

$x_1x_2x_3\ldots x_{n-1}0x_n$

$\ldots$

$x_10x_2x_3\ldots x_{n-1}x_n$

$0x_1x_2x_3\ldots x_{n-1}x_n$

$1x_1x_2x_3\ldots x_{n-1}x_n$

$x_11x_2x_3\ldots x_{n-1}x_n$

$\ldots$

$x_1x_2x_3\ldots x_{n-1}1x_n$

$x_1x_2x_3\ldots x_{n-1}x_n1$

で得られる系列と捉える。

上から下へ、符号の定義式の値が高々 1 だけ増える。

違いはちょうど n。（つまり $n+1$ 通り）

定義式の値は $n+1$ で法をとると全通り現れる。

Helberg の符号と Fibonacci 数列

Levenshtein のアイデアを拡張
例：２重挿入／削除誤り訂正符号を与える条件式

$$x_1 + 2x_2 + 4x_3 + 7x_4 \cdots + f_n x_n \equiv a \pmod{f_{n+1}}$$

ここで，$f_n := F_{n-2} - 1$ であり，F_n はフィボナッチ数列である．

組織符号化のできる 1 重挿入／削除誤り訂正符号

Theorem

p と n を正の整数とし，$p < n < 2^{p-1} + p$ をみたすとする．また $\lambda_1 := 1, \lambda_2 := 2, \ldots, \lambda_p := 2^{p-1}$ かつ $\lambda_t := 2^{p-1} + t - p$（$p < t \leq n$）とおく．このとき

$$C := \{\mathbf{x} \mid \sum_{1 \leq i \leq n} \lambda_i x_i \equiv 0 \pmod{2^{p+1}}\}$$

は 1 重挿入／削除誤り訂正符号であり，組織符号化ができる．情報ビットを $(m_1, m_2, \ldots, m_{n-p})$ はパリティビット $r_1, r_2, \ldots, r_p$ をつけることで $(r_1, r_2, \ldots, r_p, m_1, m_2, \ldots, m_{n-p})$ に符号化できる．

この定理から，情報ビットが $n - p$ ビットであり，符号語は 2^{n-p} 個となることがわかる．
ちなみにこの符号の符号化率は，Levenshtein 符号と同様，符号長を伸ばすと 1 に収束する．

完全符号，とくに削除誤り

Definition

長さ n のビット列からなる符号 C が 1 削除誤りに対して**完全**

$\Leftrightarrow$

各符号語に 1 削除誤りを施して得られる系列全体が長さ $n-1$ のビット列全体に一致すること．

Theorem

任意の正整数 n に対し，$n+1$ 種の符号 $L_{n,a}$（ここで $a = 0, 1, \ldots$）は単一削除誤り訂正符号であり，かつ，完全である．

完全符号，とくに挿入誤り

同様に挿入誤りに対する完全符号も定義できる.

Theorem

単一挿入誤りに対する長さが正の完全符号は

$$\{00, 11\}$$

だけ。
（*Up to* 記号）

ちなみ符号長に 0 も許せば,
どんなアルファベット Σ に対しても
空集合は（1 に限らず任意の t 挿入誤りに対して）完全符号.

挿入誤りを捉えなおす: BPSK

$(\alpha, \beta, 1)$
$\Downarrow$ 法線ベクトル $(0, 1, -1)$ を持つ平面で鏡映
$(\alpha, 1, \beta)$
$\Downarrow$ 法線ベクトル $(1, -1, 0)$ を持つ平面で鏡映
$(1, \alpha, \beta)$
$\Downarrow$ 法線ベクトル $(1, 0, 0)$ を持つ平面で鏡映
$(-1, \alpha, \beta)$
$\Downarrow$ 法線ベクトル $(1, -1, 0)$ を持つ平面で鏡映
$(\alpha, -1, \beta)$
$\Downarrow$ 法線ベクトル $(0, 1, -1)$ を持つ平面で鏡映
$(\alpha, \beta, -1)$

バランス隣接削除と挿入

Definition (BAD...Balanced Adjacent Deletion)

01 もしくは 10 を削除

Example

0110010 :
(01)10010 $\to$ 10010
01(10)010 $\to$ 01010
0110(01)0 $\to$ 01100
01100(10) $\to$ 01100

BAD と BAI（バランス隣接挿入）から
Levenshtein 同様の距離構造
が得られる。

単一 BAD 誤り訂正符号

BAD 誤り訂正符号を削除誤り訂正符号と同様に定義。

Example

$$C'_{5,1} = \{10000, 11000, 00110, 01110, 10101, 11101\}$$

符号語：誤り後
10000 : 000
11000 : 100
00110 : 010, 001
01110 : 110, 011
10101 : 101
11101 : 111

実は完全符号。

単一 BAD 誤り訂正符号の構成法

$$C'_{n,b} := \{x \in \{0,1\}^n \backslash \{\mathbf{0},\mathbf{1}\} \mid x_1 - 0x_2 + 2x_3 - x_4 + 3x_5 - 2x_5 \cdots \equiv b \pmod{n}\}$$

係数について：
奇数位置は１，２，３，…と増えていく。
偶数位置は０，－１，－２，…と減っていく。

ちなみに，
符号化率は，Levenshtein 符号と同様，
符号長を伸ばすと１に収束する．

挿入球の表面積

Theorem

n と m を整数、$0 \leq m \leq n$.
系列 $y \in \{0,1\}^m$ に対し,
$S(y,n)$ を y に n 回の BAI で得られる系列全体.

このとき，n, m を固定した上で,
$S(y,n)$ の濃度（集合の要素数）は y に依らない.

$$|S(y,n)| = \binom{m+2n}{n}$$

二重隣接削除と挿入

Definition (DAD...Double Adjacent Deletion)

00 もしくは 11 を削除

Example

0110010 :
0(11)0010 $\to$ 00010
011(00)10 $\to$ 01110

DAD と DAI（二重隣接挿入）から
Levenshtein 同様の距離構造
が得られる。

単一 DAD 誤り訂正符号

$$D'_{n,c} := \{x \in \{0,1\}^n \backslash \{\mathbf{0},\mathbf{1}\} \mid x_1 + 0x_2 + 2x_3 + x_4 + 3x_5 + 2x_5 \cdots \equiv c \pmod{n}\}$$

係数について：
奇数位置は１，２，３，…と増えていく。
偶数位置は０，１，２，…と増えていく。
ちなみに，
符号化率は，Levenshtein 符号と同様，
符号長を伸ばすと 1 に収束する．

挿入球の表面積

Theorem

n と m を整数、$0 \leq m \leq n$.
系列 $y \in \{0,1\}^m$ に対し,
$S'(y,n)$ を y に n 回の DAI で得られる系列全体.

このとき，n, m を固定した上で,
$S'(y,n)$ の濃度（集合の要素数）は y に依らない.

$$|S'(y,n)| = \binom{m+2n}{n}$$