訂正可能な削除割合の限界

安永 憲司（金沢大学）

2017.2.23
ICAワークショップ@唐津市

削除訂正

- 010010 → 01000, 001101 → 00111 （１個削除）
  010010 → 0110,　001101 → 0110　（２個削除）

- $(r+1)$ 回繰り返し符号は、r 個削除訂正可能
  - $rep_{r+1}(s) = \underbrace{s_1 \cdots s_1}_{r+1\text{ 回}} \, s_2 \cdots s_2 \; \cdots \; s_t \cdots s_t$
  - $s' = \quad s_2' \quad$
    この位置は必ず s_2
  - レート $\frac{t}{(r+1)t} = \frac{1}{r+1}$, 削除割合 $\frac{r}{(r+1)t} = O\left(\frac{1}{t}\right)$

削除訂正符号

- $[k]=\{1,2,\dots,k\}$ 上の p 削除訂正符号 $C\subseteq[k]^n$
$\Leftrightarrow \forall c_1,c_2\in C, \mathrm{LCS}(c_1,c_2)<(1-p)n$

- $\mathrm{LCS}(c_1,c_2)$：最長共通部分系列の長さ (Longest Common Subsequence)
  - $\mathrm{LCS}(111222,121212)=4$
  - $\mathrm{LCS}(111222333,123123123)=5$
- c_1,c_2 を最小の削除数で一致させたときの長さが $\mathrm{LCS}(c_1,c_2)$

  → このときの削除数 $=n-\mathrm{LCS}(c_1,c_2)$

  → $n-\mathrm{LCS}(c_1,c_2)>pn$ なら p 削除訂正可能

削除訂正割合の限界

- 正レート符号で p 削除訂正可能な p の上限は？
  - $|C| \geq \exp(\Omega_k(n))$ で p 削除訂正 →冗長度 $O_k(1)$
- $p^*(k) = \limsup\{\, p \in (0,1)$: $\exists[k]$上の正レート p 削除訂正符号 $\}$
  - $p^*(k) \leq 1 - 1/k$
  - $p^*(2) \geq 0.17$
    - ランダム符号は 0.788n < LCS < 0.8263n
  - $p^*(k) \geq 1 - O(1/\sqrt{k})$ [Guruswami, Wang 2015]
    - $k \to \infty$ のとき $\mathrm{E}[\mathrm{LCS}(c_1, c_2)] \sim \frac{2}{\sqrt{k}} n$ [Kiwi, Loebl, Matousek 2004]

紹介する論文

- Boris Bukh, Venkatesan Guruswami, Johan Hastad: An improved bound on the fraction of correctable deletions. ECCC TR15-117.
 （SODA 2016 とその改良）

- 主結果：
  - $\forall k \geq 2, \;\; p^*(k) \geq 1 - \frac{2}{k+\sqrt{k}}$
  - $\forall \varepsilon > 0$, レート $r(\varepsilon, k) > 0$ の明示的な k 元符号が存在し、$\mathrm{LCS}(c_1, c_2) < \frac{2}{k+\sqrt{k}} + \varepsilon$
    - ２元のとき、 $\sqrt{2} - 1 - \varepsilon > 0.414 - \varepsilon$ 削除訂正可能

準備

- $w \in [k]^n$ の部分系列 (subsequence) : w から削除して得られる系列
- $w \in [k]^n$ の部分語 (subword) : w 内の連続した系列
- w 内の部分系列 w' のスパン $\mathrm{span}_w(w')$: w' を含む w の最短部分語の長さ
  - $\mathrm{span}_{1121112}(212) = 5$, $\mathrm{span}_{11222333}(122) = 3$
- w_1 と w_2 の共通部分系列 (common subsequence) (w'_1, w'_2): $w'_1 = w'_2$
- w_1 と w_2 の最長共通部分系列の長さ $\mathrm{LCS}(w_1, w_2)$
- $C \subseteq [k]^n$ に対し、$\mathrm{LCS}(C) = \max_{c_1 \neq c_2 \in C} \mathrm{LCS}(c_1, c_2)$

準備

- w_1 と w_2 の共通部分系列 (w'_1, w'_2) のスパン：
  $\mathrm{span}(w'_1, w'_2) = \mathrm{span}_{w_1}(w'_1) + \mathrm{span}_{w_2}(w'_2)$

- 事実. w_1 と w_2 の任意の共通部分系列 (w'_1, w'_2) に対し、 $\mathrm{span}(w'_1, w'_2) \geq b\,\mathrm{len}(w'_1) - c$ ならば、

$$\mathrm{LCS}(w_1, w_2) \leq \frac{2n + c}{b}$$

  - 証明：
    - $b\,\mathrm{len}(w'_1) \leq \mathrm{span}(w'_1, w'_2) + c \leq 2n + c$
    - $\mathrm{LCS}(w_1, w_2) = \max_{(w'_1, w'_2)} \mathrm{len}(w'_1) \leq \frac{2n+c}{b}$

削除訂正符号のアルファベット削減

- $\forall \varepsilon > 0, K = K(\varepsilon) \gg k$ に対し、
  $C_1 \subseteq [K]^n$ with $\mathrm{LCS}(C_1) \ll \varepsilon n$ を
  $C_2 \subseteq [k]^N$ with $\mathrm{LCS}(C_2) \approx \frac{2}{k+\sqrt{k}} N$ へ変換
  - C_1 の各シンボルを、サイズ K の符号で連接符号化
  - 内符号の構成法により最終的な符号の性質が変わる

- きれいな構成法 (定理 4) : $\mathrm{LCS}(C_2) \approx \frac{2}{k+1} N$
- 汚れた構成法 (定理 3) : $\mathrm{LCS}(C_2) \approx \frac{2}{k+\sqrt{k}} N$

きれいな構成法

定理４

$C_1 \subseteq [K]^n$ with $\mathrm{LCS}(C_1) = \gamma n, k \geq 2$ に対し
ある $T = T(K, \gamma, k) \leq 32 \left(\frac{2k}{\gamma}\right)^K$ と $\tau \colon [K] \to [k]^T$ が存在し、
C_1 の各シンボルを τ で連接符号化して得られる
$C_2 \subseteq [k]^N, N = nT$ は、
２つの異なる $c, \tilde{c} \in C_2$ の共通部分系列 s に対し以下を満たす:

$$\mathrm{span}(s) \geq (k+1)\,\mathrm{len}(s) - 4\gamma k N$$

特に、 $\mathrm{LCS}(C_2) \leq \left(\frac{2+4\gamma k}{k+1}\right) N < \left(\frac{2}{k+1} + 4\gamma\right) N$

きれいな構成法の符号化法 (1/2)

- 整数 L を割り切る整数 A に対し、振幅 A の語：
$$f_A = (1^A 2^A \dots k^A)^{L/A}$$
  - $\mathrm{len}(f_A) = kL$
  - $L = 4, k = 2$ のとき
$f_1 = 12121212, f_2 = 11221122, f_4 = 11112222$
  - $L = 6, k = 3$ のとき
$$f_1 = 123123123123123123$$
$$f_2 = 112233112233112233$$
$$f_3 = 111222333111222333$$
$$f_6 = 111111222222333333$$

きれいな構成法の符号化法 (2/2)

- B/A が大きいとき f_A と f_B が長い共通部分系列を持たないことを利用

- $R \geq k,\ [K]$ 上の語 $w = \ell_1 \ell_2 \ldots$ に対し、
$$\tilde{w} = f_{R^{\ell_1 - 1}} f_{R^{\ell_2 - 1}} \ldots$$
  - $\mathrm{len}(\tilde{w}) = kL\,\mathrm{len}(w)$
  - $A, B \leq R^{K-1}$
  - $\tilde{w}$ 内のシンボル x が、 w 内のシンボル y から展開されているとき、 y は x の**親**と呼ぶ

きれいな構成法の分析

補題 5

$f_A^\infty = (1^A 2^A \ldots k^A)^*$ とし、 $kA \le B$ のとき、
f_A^∞ と f_B^∞ の共通部分系列 $s = (w'_1, w'_2)$ に対し、

$$\mathrm{span}(s) \ge \left(k + 1 - \frac{kA}{B}\right) \mathrm{len}(s) - 2(A + B)$$

- 用語の定義
  - チャンク：ℓ^A, ℓ^B の形をした部分語
  - f_A^∞ 内のチャンクが w'_1 によってスパンされる
    ⇔ w'_1 のスパンに対応する部分語が、
    そのチャンクのシンボルを１つ以上含む
    - 例. $f_A^\infty = 111122223333111122223333\ldots$
      $w'_1 = \quad 11 \qquad\qquad 31 \quad 2$
    - スパンされるチャンクは、$1111, 2222, 3333, 1111, 2222$

補題 5 の証明 (1/2)

- 方針： $s = (w'_1, w'_2)$ でスパンされるチャンク数を見積もる
- $s = k_1^{p_1} k_2^{p_2} \ldots k_t^{p_t}$ の形をしている（$k_\ell \neq k_{\ell+1}$）
- $k_\ell^{p_\ell}$ は f_A^∞ 内で $k \left\lceil \frac{p_\ell - A}{A} \right\rceil + 1$ 個以上のチャンクをスパンする
  - k_ℓ^A というチャンクは含まれる
  - $p_\ell > A$ なら、その他の k_ℓ^A を $\left\lceil \frac{p_\ell - A}{A} \right\rceil$ 個含む

補題 5 の証明 (2/2)

- f_A^∞ と f_B^∞ の両方で $k_\ell^{p_\ell}$ によってスパンされるチャンクに含まれるシンボル数は

$$\phi(p_\ell) = A\left(k\left\lceil\frac{p_\ell - A}{A}\right\rceil + 1\right) + B\left(k\left\lceil\frac{p_\ell - B}{B}\right\rceil + 1\right)$$

- $\phi(p_\ell) \geq \begin{cases} k(p_\ell - A) + B & p_\ell \leq B \\ k(p_\ell - A) + k(p_\ell - B) + B & p_\ell > B \end{cases}$
  $\geq \left(k + 1 - \frac{kA}{B}\right) p_\ell$

- $k_\ell^{p_\ell}$ と $k_{\ell'}^{p_{\ell'}}$ でスパンされるチャンクは異なるため、
  (s でスパンされるチャンクに含まれるシンボル数)
  $\geq \sum_\ell \phi(p_\ell) \geq \left(k + 1 - \frac{kA}{B}\right) \mathrm{len}(s)$

- 最初と最後のチャンクは全て入っていないかもしれないので $2(A+B)$ を引く

(証明終)

共通部分系列のマッチ

- (w'_1, w'_2): $\widetilde{w_1}$ と $\widetilde{w_2}$ の共通部分系列
- (w'_1, w'_2) 内の i 番目シンボルが well-matched
  ⇔ $w'_1[i]$ と $w'_2[i]$ の親が $[K]$ 内で同じ記号
- 共通部分系列が badly-matched
  ⇔ どのシンボルも well-matched でない
- 例. $L = 4, k = 2, R = 2, w_1 = 1332, w_2 = 2313,$

$$\widetilde{w_1} = f_{R^0} f_{R^2} f_{R^2} f_{R^1} = \overset{1}{12121212}\ \overset{3}{11112222}\ \overset{3}{11112222}\ \overset{2}{11221122}$$

$$\widetilde{w_2} = f_{R^1} f_{R^2} f_{R^0} f_{R^2} = \underset{2}{11221122}\ \underset{3}{11112222}\ \underset{1}{12121212}\ \underset{3}{11112222}$$

補題6

$w_1, w_2 \in [K]^*$ であり、 $\widetilde{w_1}$ と $\widetilde{w_2}$ の共通部分系列 $s = (w'_1, w'_2)$ が badly-matched のとき

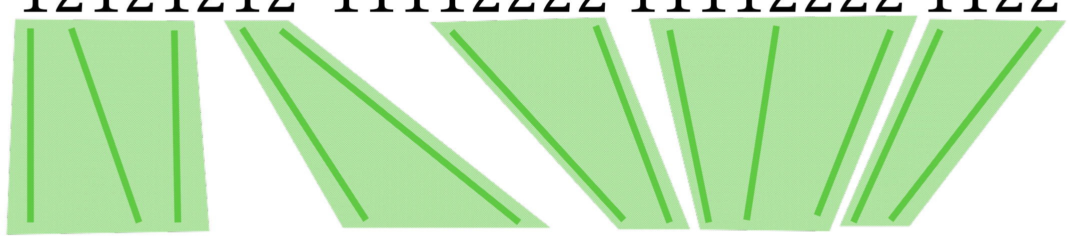
$$\mathrm{span}(s) \geq \left(k + 1 - \frac{k}{R} - \frac{8R^{K-1}}{L}\right) \mathrm{len}(s) - 16R^{K-1}$$

■ 証明：

- 以下を満たすように $s = (s_1, s_2, \ldots, s_r)$ と分割
  - 各 s_i の w'_j における親が同じであり、 r は最小

$w_1 =$ 1 3 3 2

$\widetilde{w_1} =$ 12121212 11112222 11112222 11221122



$\widetilde{w_2} =$ 11221122 11112222 12121212 11112222

$w_2 =$ 2 3 1 3

補題６の証明（続き）

- 各 s_i に対する w_1 と w_2 における親は異なる
- w_1 と w_2 における $r-4$ 個以上のシンボルを展開したものは、 w'_1 と w'_2 のスパンに含まれる
  - w_1 と w_2 の左右両端４つを除けばよい
- $kL(r-4) \leq \mathrm{span}(s)$ であり、 $r \leq \dfrac{\mathrm{span}(s)}{kL} + 4$
- 補題５より

$$
\begin{aligned}
\mathrm{span}(s) &\geq \left(k + 1 - \frac{k}{R}\right)\mathrm{len}(s) - 2(A+B)r \\
&\geq \left(k + 1 - \frac{k}{R}\right)\mathrm{len}(s) - 4rR^{K-1} \\
&\geq \left(k + 1 - \frac{k}{R}\right)\mathrm{len}(s) - 4R^{K-1}\left(\frac{\mathrm{span}(s)}{kL} + 4\right)
\end{aligned}
$$

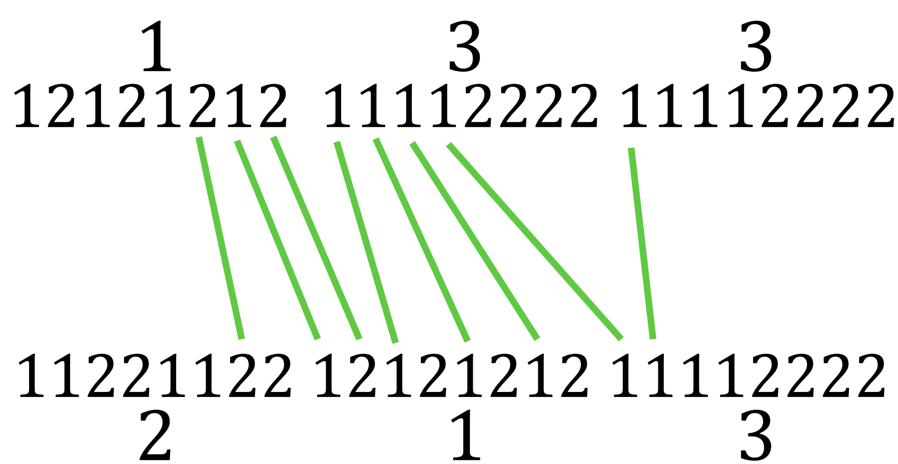
（証明終）

補題 7

$w_1, w_2 \in [K]^*$ であり、$\widetilde{w_1}$ と $\widetilde{w_2}$ の共通部分系列
$s = (w'_1, w'_2)$ に対し

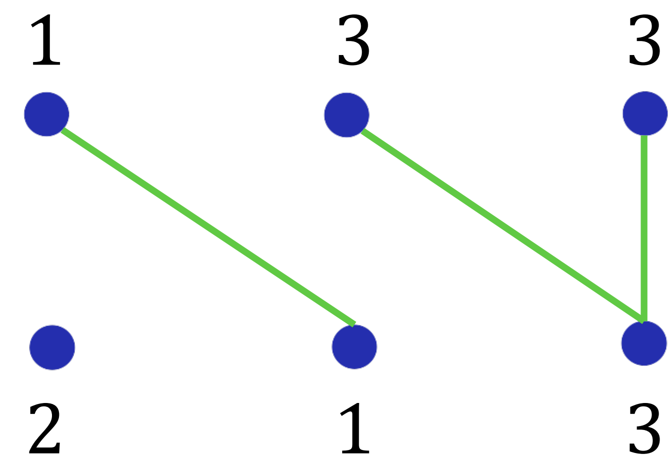
$$\mathrm{span}(s) \geq \left(k + 1 - \frac{k}{R} - \frac{8R^{K-1}}{L}\right) \mathrm{len}(s)$$
$$-2Lk(k+1)\mathrm{LCS}(w_1, w_2) - 16R^{K-1}$$

- 証明：
  - s はスパン内で長さ最大と仮定
    - $\mathrm{len}(s)$ を増やせば $\mathrm{span}(s)$ も増える



補題７の証明（続き）

- マッチしたシンボル同士を連結した二部グラフ G を構成
  - 次数は２以下にできる
- グラフの最大マッチングは、$|E(G)|/2$ 以上, $\mathrm{LCS}(w_1, w_2)$ 以下
  → $|E(G)| \leq 2\mathrm{LCS}(w_1, w_2)$



- s 内のマッチしたシンボルをすべて削除 → s'
  $\mathrm{len}(s') \geq \mathrm{len}(s) - Lk|E(G)| \geq \mathrm{len}(\mathrm{s}) - 2Lk\ \mathrm{LCS}(w_1, w_2)$
- s' は badly-matched であり、補題６より

$$\mathrm{span}(s) \geq \mathrm{span}(s')$$

$$\geq \left(k + 1 - \frac{k}{R} - \frac{8R^{K-1}}{L}\right)\mathrm{len}(s) - 2Lk(k+1)\mathrm{LCS}(w_1, w_2) - 16R^{K-1}$$

（証明終）

きれいな構成法（再掲）

定理４

$C_1 \subseteq [K]^n$ with $\mathrm{LCS}(C_1) = \gamma n, k \geq 2$ に対し
ある $T = T(K, \gamma, k) \leq 32 \left(\frac{2k}{\gamma}\right)^K$ と $\tau\colon [K] \to [k]^T$ が存在し、
C_1 の各シンボルを τ で連接符号化して得られる
$C_2 \subseteq [k]^N, N = nT$ は、
２つの異なる $c, \tilde{c} \in C_2$ の共通部分系列 s に対し以下を満たす:

$$\mathrm{span}(s) \geq (k+1)\,\mathrm{len}(s) - 4\gamma k N$$

特に、 $\mathrm{LCS}(C_2) \leq \left(\frac{2+4\gamma k}{k+1}\right) N < \left(\frac{2}{k+1} + 4\gamma\right) N$

定理４の証明

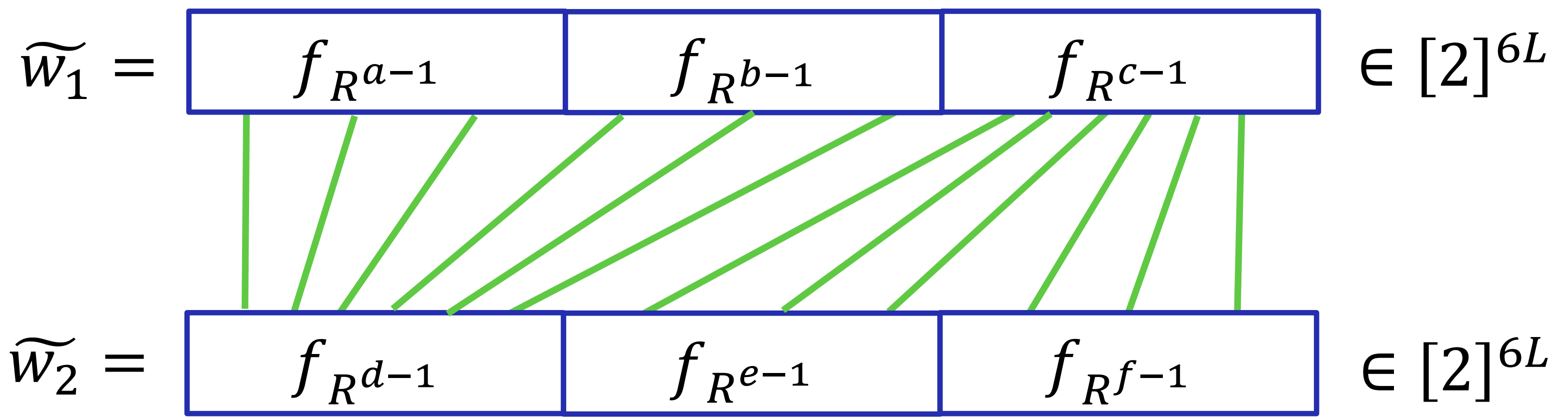
- $C_1 \subseteq [K]^n$ with $\mathrm{LCS}(C_1) = \gamma n, \varepsilon > 0, k \geq 2$ に対し
  $R = \left\lceil \frac{2k}{\gamma} \right\rceil, L = 16R^{K-1} \left\lceil \frac{1}{\gamma} \right\rceil$ とする
- $\tau: [K] \to [k]^T, T = kL$ を $\tau(\ell) = f_{R^{\ell-1}}$ と定義し
  $C_2 \subseteq [k]^N, N = nkL$ とする
  - $T = 16 \left\lceil \frac{2k}{\gamma} \right\rceil^{K-1} \left\lceil \frac{1}{\gamma} \right\rceil k \leq 32 \left(\frac{2k}{\gamma}\right)^K$ である
- 補題７より、 C_2 の任意の共通部分系列 s は

$$\begin{aligned}
\mathrm{span}(s) &\geq \left(k + 1 - \frac{k}{R} - \frac{8R^{K-1}}{L}\right) \mathrm{len}(s) - 2Lk(k+1)\mathrm{LCS}(w_1, w_2) \\
&\quad -16R^{K-1} \\
&\geq \left(k + 1 - \frac{\gamma}{2} - \frac{\gamma}{2}\right) \mathrm{len}(s) - 2(k+1)\gamma N - \gamma N \\
&\geq (k+1)\mathrm{len}(s) - \gamma N - 2(k+1)\gamma N - \gamma N \\
&= (k+1)\mathrm{len}(s) - 2(k+2)\gamma N \\
&\geq (k+1)\mathrm{len}(s) - 4k\gamma N
\end{aligned}$$

（証明終）

きれいな構成法の限界

- $k = 2$ で誤り割合 $p < 1/3$ しか達成できない理由
  - $w_1 = abc, w_2 = def \in [K]^3$ が
    $d > a, b$ および $c > e, f$ を満たすとき



  - $f_{R^{d-1}}$ は $f_{R^{a-1}}, f_{R^{b-1}}$ にマッチする（$f_{R^{c-1}}$ も同様）
    - $f_{R^{a-1}}, f_{R^{b-1}}$ の方が、変動が早いため
  - 長さ $4L$ の共通部分系列が存在
    - a, b, c, d, e, f に共通シンボルが存在していなくても
- $k > 2$ でも同様の議論より、 $p < \frac{k-1}{k+1}$ が限界

改善のアイディア (1/2)

- $f_A^\infty = (1^A 2^A \dots k^A)^*, f_B^\infty = (1^B 2^B \dots k^B)^*, A \ll B$
- スパンの短い共通部分系列がネック
  - 補題５ではチャンク内シンボル数の下界を導出
    $$\phi(p_\ell) = A\left(k\left\lceil\frac{p_\ell - A}{A}\right\rceil + 1\right) + B\left(k\left\lceil\frac{p_\ell - B}{B}\right\rceil + 1\right)$$
  - $p_\ell = B$ のときに値が小さそう
    - $s = 1^B$ とすると、f_B^∞ 内のスパンは f_A^∞ 内の半分

$$f_A^\infty = \boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\cdots$$

$$s = \boxed{1^A}\;\;\boxed{1^A}\;\;\boxed{1^A}\;\;\boxed{1^A}$$

$$f_B^\infty = \boxed{1^B}\boxed{2^B}\boxed{1^B}\boxed{2^B}\cdots$$

改善のアイディア (2/2)

- アイディア：f_B^∞ 内の 1^B の中に適度に 2 を挿入
  （ $2^B, 1^A, 2^A$ も同様）
- 長い共通部分系列を作ろうとするとき、
  その「汚れ」を含めるか、含めないか？
  - 1^A とマッチさせている間は、1^B 内の 2 とマッチさせても得しない
  - 2^A とマッチさせている間に、1^B 内の 2 とマッチさせると、 1^B 内の 1 とマッチしなくなる

→ 共通部分系列は広がらず、汚れの分だけスパンが伸びる

$$f_A^\infty = \boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\boxed{1^A}\boxed{2^A}\cdots$$

$$s = \boxed{1^A}\quad\boxed{1^A}\quad\boxed{1^A}\quad\boxed{1^A}$$

$$f_B^\infty = \boxed{\ \boxed{2}\ 1^B\ \boxed{2}\ }\boxed{2^B}\boxed{1^B}\boxed{2^B}\cdots$$

汚れた構成法（２元の場合）

- $0 < c \leq \sqrt{2} - 1$ を固定
- 長さ M 振幅 a の汚れた 1:

$$1_{M,a} = (1^a 2^{ca})^{M/(1+c)a}$$
$$= \underbrace{\boxed{1^a}\boxed{2^{ca}}\boxed{1^a}\boxed{2^{ca}}\boxed{1^a}\boxed{2^{ca}} \cdots \boxed{1^a}\boxed{2^{ca}}}_{M}$$

- （参考）きれいな構成： $f_{R^{i-1}} = (1^{R^{i-1}} 2^{R^{i-1}})^{L/R^{i-1}}$
- 汚れた構成： $g_i = (1_{R^{K+1+i},R^{K-i}}\ 2_{R^{K+1+i},R^{K-i}})^{L/R^{K+1+i}}$
  - 整数 $R = (1 + c)t$ for integer t, $L = R^{2K+1}$
  - $\mathrm{len}(g_i) = R^{K+1+i} \times 2 \times \frac{L}{R^{K+1+i}} = 2L$

汚れた構成法の分析

補題８

$w_1 = 1_{\infty,a}$ (or $2_{\infty,a}$) および
$w_2 = (1_{b_1,b_2}\ 2_{b_1,b_2})$ の部分系列 s に対し、

$$\mathrm{span}_{w_1} s + 2b_1 \geq (3+c)\mathrm{len}(s) - \frac{4ab_1}{b_2}$$

■ 証明：

- $w_1 = 1_{\infty,a}$ と仮定
  - 対称性より $w_1 = 2_{\infty,a}$ も同様に証明可能
- $w_2 = \underbrace{\boxed{1^{b_2}}\boxed{2^{cb_2}}\ \cdots\ \boxed{1^{b_2}}\boxed{2^{cb_2}}}_{b_1}\underbrace{\boxed{2^{b_2}}\boxed{1^{cb_2}}\ \cdots\ \boxed{2^{b_2}}\boxed{1^{cb_2}}}_{b_1}$
- s は $1^{b_2}, 1^{cb_2}, 2^{b_2}, 2^{cb_2}$ の部分語で構成

補題８の証明の続き (1/2)

- s をそのような部分語に分解
  - $s = 1^{s_1} 1^{s_2} 2^{t_1} 1^{s_3} 2^{t_2} \cdots$
  - i 番目の 1^* の部分語は s_i 個の 1 で構成
  - i 番目の 2^* の部分語は t_i 個の 2 で構成
- このとき $\mathrm{span}_{w_1} 1^{s_i} \geq \left(\frac{s_i}{a} - 1\right)(1+c)a$
  - 同様に $\mathrm{span}_{w_1} 2^{t_i} \geq \left(\frac{t_i}{ca} - 1\right)(1+c)a$
- s に合計 S 個の 1, $\tilde{S}$ 個の 2 が含まれるとき

$$
\begin{aligned}
\mathrm{span}_{w_1} s &\geq \textstyle\sum_i \left(\frac{s_i}{a} - 1\right)(1+c)a + \sum_i \left(\frac{t_i}{ca} - 1\right)(1+c)a \\
&\geq \frac{S}{a}(1+c)a + \frac{\tilde{S}}{ca}(1+c)a - \underbrace{\textstyle\sum (1+c)a}_{\text{部分語の数} \leq \frac{b_1}{(1+c)b_2} \times 2 \times 2} \\
&\geq (1+c)S + \frac{1+c}{c}\tilde{S} - \frac{4ab_1}{b_2}
\end{aligned}
$$

補題８の証明の続き (2/2)

- $\mathrm{len}(s) = S + \tilde{S}$ であるため
$$(1+c)S + \frac{1+c}{c}\tilde{S} + 2b_1 \geq (3+c)\left(S+\tilde{S}\right)$$
を示せばよい

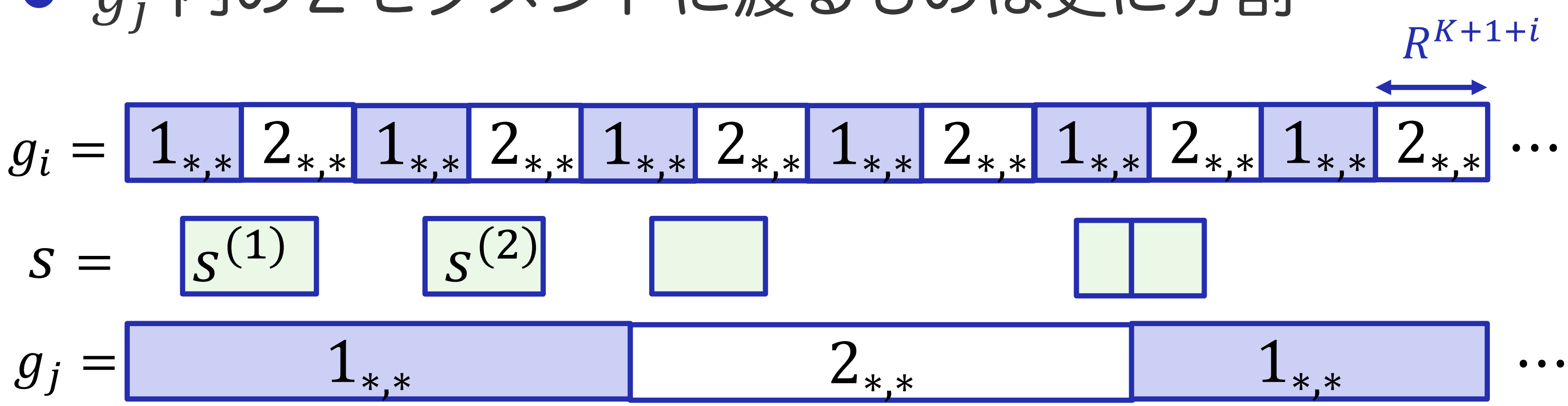
- $S, \tilde{S} \in [0, b_1]$ であり
$0 < c \leq \sqrt{2} - 1$ より $\frac{1+c}{c} > 3 + c$ であるから
$$(1+c)S + \frac{1+c}{c}\tilde{S} + 2b_1$$
$$\geq (1+c)S + (3+c)\tilde{S} + 2S$$
$$\geq (3+c)\left(S+\tilde{S}\right)$$
(証明終)

補題９

g_i, g_j $(i < j)$ の共通部分系列 s に対し、$R \geq 10$ のとき

$$\left(1 + \frac{2}{R}\right) \mathrm{span}_{g_i} s + \mathrm{span}_{g_j} s \geq (3 + \mathrm{c}) \mathrm{len}(s) - \frac{10L}{R}$$

- 証明：
  - $g_i = (1_{R^{K+1+i}, R^{K-i}}\ 2_{R^{K+1+i}, R^{K-i}})^{L/R^{K+1+i}}$
  - s を g_i 内のセグメント ($1_{R^{K+1+i}, R^{K-i}}\ 2_{R^{K+1+i}, R^{K-i}}$) との対応関係により $s^{(1)}, s^{(2)}, \cdots$ と分割
  - g_j 内の２セグメントに渡るものは更に分割



補題９の証明（続き）

- $s = \left(s^{(1)}, s^{(2)}, \cdots, s^{(p)}\right)$, $p \leq 2 + \frac{\mathrm{span}_{g_i} s}{2R^{K+1+i}} + \frac{2L}{R^{K+1+j}}$
- 各 $s^{(k)}$ は $a = R^{K-j}, b_1 = R^{K+1+i}, b_2 = R^{K-i}$ として補題８の仮定を満たす
  - $w_1 \Leftrightarrow g_j$ 内のセグメントの半分
  - $w_2 \Leftrightarrow g_i$ 内のセグメント
- 補題８より

$$\mathrm{span}_{g_j} s^{(k)} + 2R^{K+i+1} \geq (3+c)\mathrm{len}\left(s^{(k)}\right) - 4R^{i-j}R^{K+i+1}$$

- $\mathrm{span}_{g_j} s \geq \sum_k \mathrm{span}_{g_j} s^{(k)}$ と $\mathrm{len}(s) = \sum_k \mathrm{len}\left(s^{(k)}\right)$ より

$$\mathrm{span}_{g_j} s + 2pR^{K+i+1} \geq (3+c)\mathrm{len}(s) - 4pR^{i-j}R^{K+i+1}$$

- p の不等式より

$$\mathrm{span}_{g_j} s + \left(1 + 2R^{i-j}\right)\mathrm{span}_{g_i} s \geq (3+c)\mathrm{len}(s) - (1 + 2R^{i-j})(4R^{K+i+1} + 4LR^{i-j})$$

  - $R \geq 10, R^{K+i+1} \leq \frac{L}{R}, i < j$ より証明できる　（証明終）

汚れた構成法の符号化法

- $[K]$ 上の語 $w = \ell_1 \ell_2 \dots$ に対し $\hat{w} = g_{\ell_1} g_{\ell_2} \dots$

補題 10

$w_1, w_2 \in [K]^*$ に対し、 $\widehat{w_1}$ と $\widehat{w_2}$ の共通部分系列 $s = (w'_1, w'_2)$ が badly-matched のとき

$$\mathrm{span}(w'_1) + \mathrm{span}(w'_1) \geq \left(3 + c - \frac{28}{R}\right) \mathrm{len}(s) - \frac{40L}{R}$$

補題 11

$w_1, w_2 \in [K]^*$, $\widehat{w_1}$ と $\widehat{w_2}$ の共通部分系列 $s = (w'_1, w'_2)$ は

$$\mathrm{span}(s) \geq \left(3 + c - \frac{28}{R}\right) \mathrm{len}(s) - 16L\, \mathrm{LCS}(w_1, w_2) - \frac{40L}{R}$$

汚れた構成法

定理３

$C_1 \subseteq [K]^n$ with $\mathrm{LCS}(C_1) = \gamma n, k \geq 2$ に対し
ある $T = T(K, \gamma, k) \leq O\left((2k/\gamma)^{2K+2}\right)$ と $\tau: [K] \to [k]^T$ が存在し、
C_1 の各シンボルを τ で連接符号化して得られる
$C_2 \subseteq [k]^N, N = nT$ は、
２つの異なる $c, \tilde{c} \in C_2$ の共通部分系列 s に対し以下を満たす:

$$\mathrm{span}(s) \geq \left(k + \sqrt{k}\right) \mathrm{len}(s) - 5\gamma k N$$

特に、 $\mathrm{LCS}(C_2) \leq \left(\frac{2+5\gamma k}{k+\sqrt{k}}\right) N < \left(\frac{2}{k+\sqrt{k}} + 5\gamma\right) N$

よい削除訂正符号の存在性

補題 13

$\gamma, r > 0$ と整数 $K \geq 2$ が

$$2r \log K + 2h(\gamma) - \gamma \log K < 0$$

を満たすとき、すべての n に対し、
$[K]^n$上のサイズ K^{rn} の符号が存在し、
任意の異なる符号語 w, w' に対し、 $\mathrm{LCS}(w, w') < \gamma n$

■ 証明：

- $w_1, \ldots, w_{K^{rn}}$ を $[K]^n$ から独立にランダムに、戻すことなく、選んだ系列とする
- $i < j$ に対し、分布 (w_i, w_j) は、互いに異なるという条件下で $[K]^n$ から２つ独立に選ぶ分布に等しい
- 和集合上界より

$$\Pr[\mathrm{LCS}(w_i, w_j) > \gamma n\ ] \leq \binom{n}{\gamma n} K^{(1-\gamma)n} K^{-n} \leq \binom{n}{\gamma n}^2 K^{-\gamma n}$$

■ 証明の続き：

- 和集合上界より、十分大きな n に対し

$$\Pr\left[\exists w_i, w_j \in C, \mathrm{LCS}(w_i, w_j) > \gamma n\ \right] \le K^{2rn} \binom{n}{\gamma n}^2 K^{-\gamma n}$$
$$= 2^{n(2r \log K + 2h(\gamma) - \gamma \log K) + o(n)} < 1$$

(証明終)

定理 14

整数 k を固定. 任意の $\varepsilon > 0$ に対し $\tilde{r} = (\varepsilon/k)^{O(\varepsilon^{-3})}$ が存在し、無限に多くの N について、レート $\tilde{r}$ 以上の符号 $C \subseteq [k]^N$ が存在し、$\mathrm{LCS}(C) < \left(\frac{2}{k+\sqrt{k}} + \varepsilon\right) n$

■ 証明：

- 補題 13 で $\gamma = \frac{\varepsilon}{4}, r = \frac{\gamma}{6} = \frac{\varepsilon}{24}$ とすれば $C_1 \subseteq [K]^n, K \le O\left(\frac{1}{\varepsilon^3}\right), \mathrm{LCS}(C_1) \le \frac{\varepsilon n}{4}, |C_1| \ge K^{rn}$ となる
- 定理 3 を適用すると $C_2 \subseteq [k]^N, \mathrm{LCS}(C_2) \le \left(\frac{2}{k+\sqrt{k}} + \varepsilon\right) N$
レートは $\tilde{r} = rn/Tn \ge (\varepsilon/k)^{O(\varepsilon^{-3})}$

(証明終)

明示的な構成法

- Guruswami, Wang (RANDOM 2015)
  - $\forall\gamma > 0, \exists K \le O\left(\frac{1}{\gamma^5}\right)$
    レート$\Omega(\gamma^3)$, $\mathrm{LCS}(C) \le \gamma n$ の符号 $C \subseteq [K]^n$ を
    $n(\log n)^{\mathrm{poly}(1/\gamma)}$ 時間で構成できる
    - 符号化と復号も $n(\log n)^{\mathrm{poly}(1/\gamma)}$ 時間で可能

- 定理３と組み合わせれば、
  $\mathrm{LCS}(C) < \left(\frac{2}{k+\sqrt{k}} + \varepsilon\right)N$ を満たす符号 $C \subseteq [k]^N$
  を効率的に構成できる（定理 16）

効率的な復号ができる符号

- ハミング距離の大きな符号 (Reed-Solomon) と定理 16 の符号を連接符号化することで実現
  - 外符号の相対ハミング距離 ≈ 1 であれば、$\mathrm{LCS}(C) < \left(\frac{2}{k+\sqrt{k}} + \varepsilon\right) N$ が保たれる（補題 17）
- 構成法
  - 外符号：F_q 上の RS 符号
  - 内符号 C_{in} ：符号長 $m = O(\log q), |C_{in}| = q^2$
- 復号法
  - 境界線がわからないため、連続した系列をしらみつぶしに削除訂正 → 各位置に対し複数のシンボル候補
  - RS の list recovery によってリスト復号
  - 符号の性質より、一意復号となる

[Bukh, Guruswami, Hastad] のまとめ

- $p^*(k)$: 正レートk 元符号で訂正できる削除割合の上極限
  - $p^*(k) < 1 - 1/k$ が成立

- $\forall k \geq 2, \varepsilon > 0$,
  ∃ 効率的に構成可能なレート $r(q, \varepsilon) > 0$ の k 元符号で
  $1 - \frac{2}{k+\sqrt{k}} - \varepsilon$ 割合削除を $n^3\mathrm{poly}(\log n)$ 時間復号
  - $p^*(k) \geq 1 - \frac{2}{k+\sqrt{k}}$ （よって $p^*(k) = 1 - \Theta\left(\frac{1}{k}\right), k \to \infty$)
  - ２元符号で $\sqrt{2} - 1 - \varepsilon > 0.414 - \varepsilon$ 割合を訂正可能

削除訂正線形符号の限界

定理 8 [Brakensiek, Guruswami, Zbarsky (2016)]

k 削除訂正できる符号長 n の線形符号 C に対し、

$$\dim(C) \leq \frac{n}{k+1} + (k+1)^2$$

- k 削除訂正できる線形符号で達成可能な漸近的な符号化レートは $1/(k+1)$
  - $(k+1)$ 回繰り返し符号で達成可能

定理８の証明

- $C^0 = C,\ i = 1, \dots, k$ に対し
$$C^i = \{(x_{i+1}, \dots, x_n, x_1, \dots, x_i) | (x_1, \dots, x_n) \in C\}$$

補題９

すべての $0 \le i < j \le k$ に対し $\dim\left(C^i \cap C^j\right) \le j - i$

- 補題９の証明
  - $z \in C^i \cap C^j$ のとき $x, y \in C$ が存在し
    $z = (x_{i+1}, \dots, x_n, x_1, \dots, x_i) = \left(y_{j+1}, \dots, y_n, y_1, \dots, y_j\right)$
  - $j > i$ より $\left(x_{i+1}, \dots, x_{n-j+i}\right) = \left(y_{j+1}, \dots, y_n\right)$
  - $k > j$ より x, y は k 削除で同じ結果 → $x = y$ である
  - → $\forall \ell \in \{1, \dots, n\}, x_\ell = x_{\ell+j-i\ (mod\ n)}$ ($j - i$ シフトで等価)
  - 各 x_ℓ は最大で $j - i$ 種類の値 → x は 2^{j-i} 通り以下

$$\dim\left(C^i \cap C^j\right) \le j - i$$

(補題の証明終)

定理８の証明

- 補題９より

$$
\begin{aligned}
n &\geq \dim\left(\bigcup_{i=0}^{k} C^i\right) \\
&\geq \sum_{i=0}^{k} \dim\left(C^i\right) - \sum_{i=0}^{k} \sum_{j=i+1}^{k} \dim\left(C^i \cap C^j\right) \\
&= (k+1)\dim(C) - \sum_{i=0}^{k} \sum_{j=i+1}^{k} \dim\left(C^i \cap C^j\right) \\
&\geq (k+1)\dim(C) - (k+1)^3
\end{aligned}
$$

ここが間違い（未解決）

- したがって $\dim(C) \leq (k+1)^2 + n/(k+1)$

（証明終）

削除訂正符号に関する既存研究 (1/4)

- ランダム削除通信路
  - シンボル毎に確率 p で削除
  - 通信路容量は未解決
    - $p \to 0$ のとき, レートは $1 - h(p)$ を達成可能
    - $p \to 1$ のときでも正レートを達成可能
      - $(1-p)/9$ を達成 [Mitzenmacher, Drinea 2006]

削除訂正符号に関する既存研究 (2/4)

- Schulman, Zuckerman (SODA '97, IEEE IT '99)
  - 小さな定数割合削除を効率的に訂正する定数レート符号
- Guruswami, Wang (RANDOM 2015)
  - $0 < \varepsilon < 1/2$ に対し、$(1-\varepsilon)$ 割合削除を訂正するレート $\Omega(\varepsilon^2)$, アルファベットサイズ $\mathrm{poly}(1/\varepsilon)$ の符号
    - 構成・符号化・復号時間は $N^{\mathrm{poly}(1/\varepsilon)}$
  - $\varepsilon > 0$ に対し、 ε 割合削除を訂正するレート $1-\tilde{O}(\sqrt{\varepsilon})$ の２元符号
    - 構成・符号化・復号時間は $N^{\mathrm{poly}(1/\varepsilon)}$
  - $0 < \varepsilon < 1/2, 1/2-\varepsilon$ 割合削除をサイズ $(1/\varepsilon)^{O(\log\log(1/\varepsilon))}$ でリスト復号するレート$\Omega(\varepsilon^3)$ の２元符号
    - 構成・符号化・復号時間は $N^{\mathrm{poly}(1/\varepsilon)}$

削除訂正符号に関する既存研究 (3/4)

- $\mathrm{del}(N,k)$: 符号長 N の k 削除訂正符号の最大サイズ
  - 定数 k に対し、k だけに依存する定数 $a_k > 0, A_k < \infty$ が存在し

$$a_k \frac{2^N}{N^{2k}} \leq \mathrm{del}(N,k) \leq A_k \frac{2^N}{N^k}$$

  - $\mathrm{del}(N,1) = \Theta\left(\frac{2^N}{N}\right)$
    - 1 削除訂正の Varshamov-Tenengolts (VT) 符号 $\left\{(x_1,\dots,x_n) \in \{0,1\}^N \middle| \sum_{i=1}^{N} ix_i \equiv 0 \ \ mod\ (N+1)\right\}$
- Brakensiek, Guruswami, Zbarsky (SODA 2016)
  - 定数 k に対し、冗長度 $N - \log\big(del(N,k)\big) = O(k^2 \log k \log N)$ の k 削除訂正符号の効率的な構成法
  - 復号時間 $O_k(N(\log N)^4)$

削除訂正符号に関する既存研究 (4/4)

- Oblivious deletions
  - 符号語を見ずに削除パターンを決める通信路
  - ランダム削除と最悪ケース削除の間

- Guruswami, Li (arXiv 2016)
  - $\forall p \in (0,1), \exists R > 0, Enc\colon \{0,1\}^{Rn} \to \{0,1\}^{n}, Dec\colon \{0,1\}^{(1-p)n} \to \{0,1\}^{Rn} \cup \{\perp\}$ s.t. 任意の pn 個削除のパターン τ と $m \in \{0,1\}^{Rn}$ に対し
    $$\Pr[\, Dec(\tau(Enc(m))) \neq m] \leq o(1)$$
  - 明示的構成法で効率的な復号ができるレート $(1-p)/180$ の符号

任意の削除レートを訂正できる！

未解決問題

- $p^*(k)$ の特定
  - $k=2$ のとき、 0.414 と 0.5 の間
  - $p^*(k) = 1 - \Theta\left(\frac{1}{k}\right), k \to \infty$

- p 割合削除を訂正するレート $1-p-\varepsilon$,
  アルファベットサイズ $\mathrm{poly}(1/\varepsilon)$ の符号の構成

- ε 割合削除を訂正するレート $1-\varepsilon\, poly(\log(1/\varepsilon))$
  の２元符号の構成
  - レート $1-O\big(\varepsilon(\log(1/\varepsilon))\big)$ はランダム符号で存在