挿入や削除を訂正する符号のサイズの上界・下界

安永 憲司

東京工業大学

エクスパンダーグラフの新しい構成手法の確立とその応用2 @九州大学 IMI
2023.9.7

自己紹介

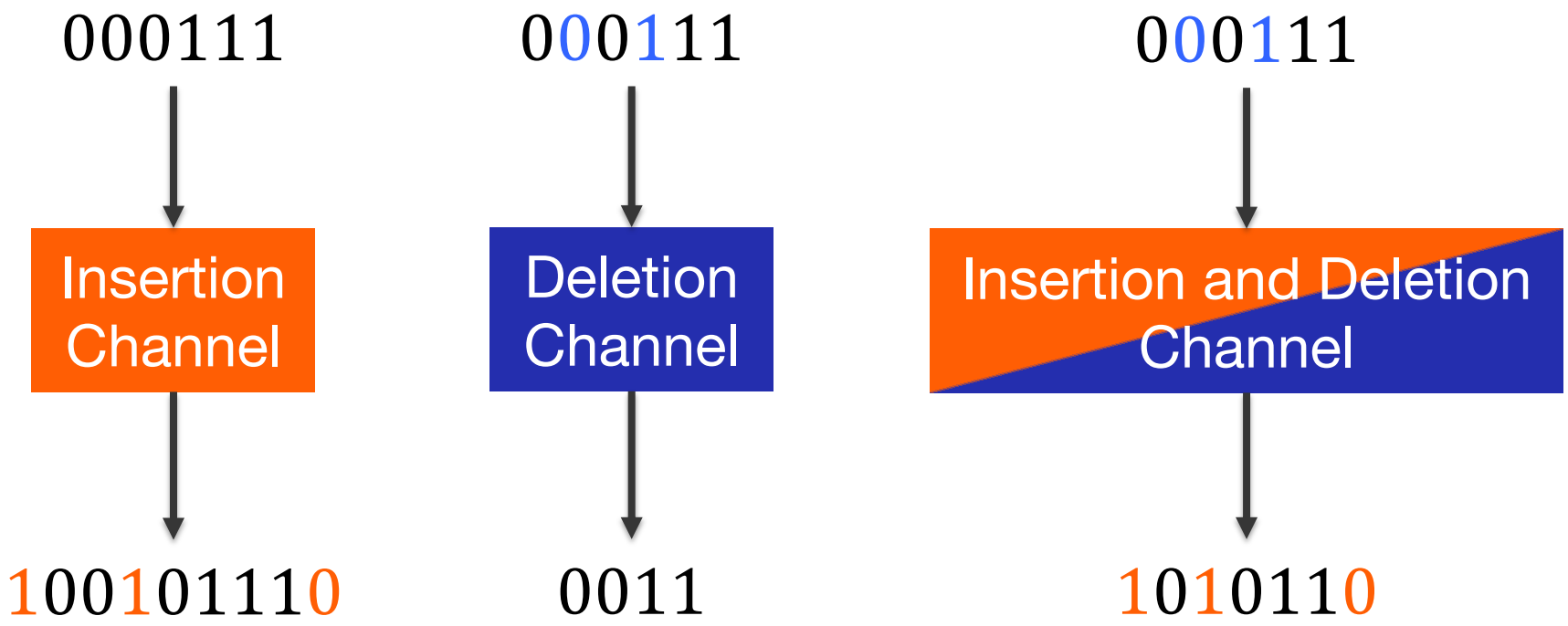
名前：安永 憲司

所属：東京工業大学　情報理工学院　数理・計算科学系

研究分野：理論計算機科学

- 主に，暗号理論
- 細々と，誤り訂正符号
- これから，計算量理論

挿入と削除


000111
Insertion Channel
100101110
000111
Deletion Channel
0011
000111
Insertion and Deletion Channel
1010110

Levenshtein 距離（編集距離）

$d_L(\boldsymbol{x}, \boldsymbol{y})$:= min { $\boldsymbol{x}$ を $\boldsymbol{y}$ に変換するのに必要な挿入・削除数 }

- 例. $d_L(000, 111) = 6$,　$d_L(101, 010) = 2$
- $|\boldsymbol{x}| = |\boldsymbol{y}| = n$ のとき，$0 \leq d_L(\boldsymbol{x}, \boldsymbol{y}) \leq 2n$
- $d_L(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x}| + |\boldsymbol{y}| - 2\mathrm{LCS}(\boldsymbol{x}, \boldsymbol{y})$　($\mathrm{LCS}(\boldsymbol{x}, \boldsymbol{y})$: $\boldsymbol{x}$ と $\boldsymbol{y}$ の最長共通部分系列の長さ)

符号 C の最小 Levenshtein 距離 :　$d_L(C) := \min_{\boldsymbol{c}_1 \neq \boldsymbol{c}_2 \in C} d_L(\boldsymbol{c}_1, \boldsymbol{c}_2)$

$d_L(C) \geq d$ のとき，C は合計 $t \leq \left\lfloor \frac{d-1}{2} \right\rfloor$ 個の挿入・削除を訂正できる

- $C \subseteq \Sigma^n$ のとき，$d_L(C)$ は偶数で，　$t \leq \frac{d_L(C)}{2} - 1$ 個を訂正できる
- 相対最小距離は $\frac{d_L(C)}{2n} \in [0,1]$
- 符号の相対最小距離 $\geq \delta$　➔　δ 割合未満の挿入・削除を訂正

最良の符号サイズ $A_q(n,d)$

$A_q(n,d) \coloneqq \max\{\ |C| : \exists C \subseteq \Sigma^n \text{ s.t. } |\Sigma| = q, d_L(C) \geq d\ \}$

- $A_q(n,d)$ の上界 ➔ 符号として存在しない領域
- $A_q(n,d)$ の下界 ➔ 符号として存在しうる領域

<u>漸近的な評価</u>

- 符号列 $\{C_n\}_n$ の符号長 $n \to \infty$ の場合の振る舞いを評価
- $A_q(n,d)$ を達成する符号 $C \subseteq \Sigma^n,\ |\Sigma| = q$ に対し，

  符号化率 $R = \frac{\log_q|C|}{n}$ と相対最小距離 $\delta = \frac{d_L(C)}{2n}$ の

  トレードオフを明らかにしたい

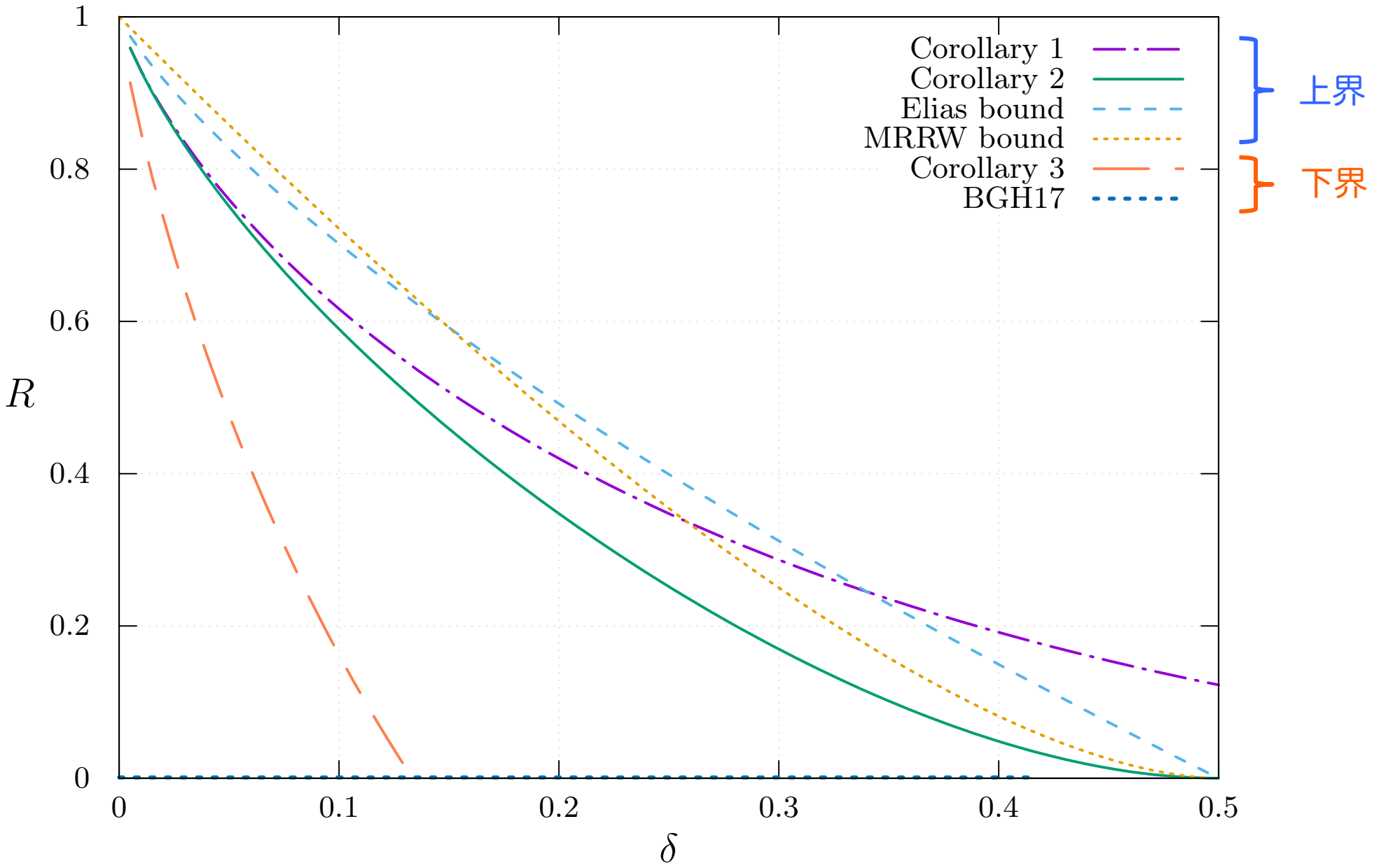
エクスパンダーグラフとの関わり

- Hamming 距離（通常の符号）
  - 定数レート符号で線形時間符号化・復号
  - 符号の最小距離の増幅
  - リスト復号可能符号との等価性

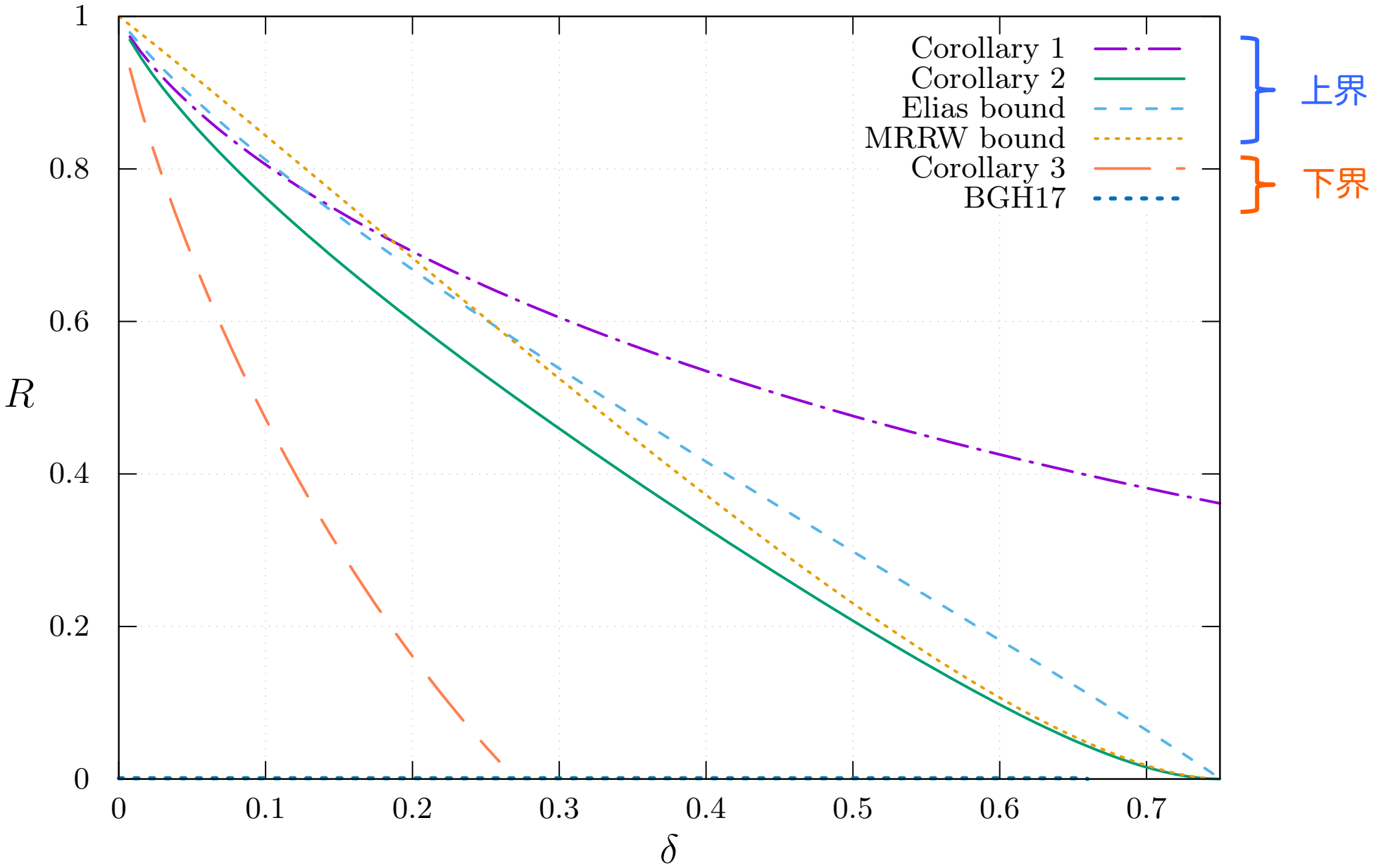
- Levenshtein 距離（挿入・削除訂正符号）
  - 直接的に関係する研究はこれまでない
  - 線形時間符号化・復号は未解決（距離を測るのに $O(n^2)$）
  - 最小距離が増幅するかも不明

主結果のトレードオフ

符号化率と相対最小距離のトレードオフ：$q = 2$



符号化率と相対最小距離のトレードオフ：$q = 4$



基本的な事実（1/2）

挿入球 $I_t(\boldsymbol{x}) \coloneqq \{\, \boldsymbol{x} \in \Sigma^n$ に t 挿入してできる文字列 $\boldsymbol{y} \in \Sigma^{n+t} \,\}$

- $|I_t(\boldsymbol{x})| = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i \coloneqq I_q(n,t) \approx q^{(n+t)H_q\left(\frac{t}{n+t}\right)}$

削除球 $D_t(\boldsymbol{x}) \coloneqq \{\, \boldsymbol{x} \in \Sigma^n$ から t 削除してできる文字列 $\boldsymbol{y} \in \Sigma^{n-t} \,\}$

- $|D_t(\boldsymbol{x})| \leq \binom{n}{t}$
- $\boldsymbol{x}$ のラン数 $= r(\boldsymbol{x}) \geq 2t$ のとき，$\sum_{i=0}^{t} \binom{r(x)-t}{i} \leq |D_t(\boldsymbol{x})| \leq \binom{r(x)+t-1}{t}$
  - 例．$r(0000) = 1,\ r(0011) = 2,\ r(0101) = 4$
- $R_q(n,r) \coloneqq \{\boldsymbol{x} \in \Sigma^n : r(x) = r\},\ |R_q(n,r)| = \binom{n-1}{r-1} q(q-1)^{r-1}$

基本的な事実（2/2）

挿入削除球 $L_{t,s}(\boldsymbol{x}) \coloneqq \{\, \boldsymbol{x} \in \Sigma^n$ に t 削除・s 挿入でできる文字列 $\boldsymbol{y} \in \Sigma^{n-t+s}\}$

- $L_{t,0}(\boldsymbol{x}) = D_t(\boldsymbol{x}),\ L_{0,t}(\boldsymbol{x}) = I_t(\boldsymbol{x})$

$|L_{t,t}(\boldsymbol{x})|$ の緊密な評価式は未解決問題

- $\left|L_{t,t}(\boldsymbol{x})\right| \leq |D_t(\boldsymbol{x})| \cdot I_q\,(n-t, t)$　が上界では最も良い？？

二重数え上げ (double counting) より，以下が成り立つ

$$\sum_{\boldsymbol{y} \in \Sigma^{n+t}} |D_t(\boldsymbol{y})| = \sum_{\boldsymbol{x} \in \Sigma^n} |I_t(\boldsymbol{x})| = q^n \cdot I_q(n, t)$$

符号サイズの上界

球充填 (sphere-packing) タイプの上界

Theorem 1. 最小 Levenshtein 距離 $d = 2(t+1)$ の $C \subseteq \Sigma^n$ に対し,

$$|C| \leq \left\lfloor \frac{q^{n+t}}{I_q(n,t)} \right\rfloor$$

証明：各 $\boldsymbol{c} \in C$ に対し，　$I_t(\boldsymbol{c}) \subseteq \Sigma^{n+t}$ は互いに交わらないため

Corollary 1. 最小 Levenshtein 距離 δn, 符号化率 R の $C \subseteq \Sigma^n$ に対し,

$$R \leq (1+\delta)\left(1 - H_q\left(\frac{\delta}{1+\delta}\right)\right) + o(1)$$

主結果その１：Elias タイプの上界

Theorem 2. 最小 Levenshtein 距離 $d < 2n$ の $C \subseteq \Sigma^n$ について,

$$t < \frac{d}{2 - d/n}$$

を満たす任意の $t \geq 0$ に対し,

$$|C| \leq \left\lfloor \frac{(n+t)d}{(n+t)d - 2nt} \cdot \frac{q^{n+t}}{I_q(n,t)} \right\rfloor$$

Corollary 2. 最小 Levenshtein 距離 δn, 符号化率 R の $C \subseteq \Sigma^n$ に対し,

$$R \leq \frac{1}{1-\delta}\Big(1 - H_q(\delta)\Big) + o(1)$$

Theorem 2 の証明

- 二重数え上げを，符号 C との共通部分に適用すると，

$$\sum_{\boldsymbol{y}\in\Sigma^{n+t}} |D_t(\boldsymbol{y}) \cap C| = \sum_{\boldsymbol{x}\in C} |I_t(\boldsymbol{x})| = |C| \cdot I_q(n,t)$$

- ランダムに $\boldsymbol{y} \in \Sigma^{n+t}$ を選ぶと，$|D_t(\boldsymbol{y}) \cap C| \geq \frac{|C| \cdot I_q(n,t)}{q^{n+t}}$ を満たす $\boldsymbol{y} \in \Sigma^{n+t}$ が存在
- $|D_t(\boldsymbol{y}) \cap C|$ は，t 挿入に対するリスト復号のリストサイズ
- [Hayashi, Yasunaga (IEEE IT 2020)] のリスト復号可能性を適用
  - $t < \frac{d}{2-d/n}$ に対し，$|D_t(\boldsymbol{y}) \cap C| \leq \frac{(n+t)d}{(n+t)d-2nt}$

Hamming 距離の符号に対する上界の適用

符号 C の最小 Hamming 距離 $\leq d$

➔ C の最小 Levenshtein 距離 $\leq 2d$

Hamming 距離での $A_q(n,d)$ の上界

➔ Levenshtein 距離での $A_q(n,d)$ の上界

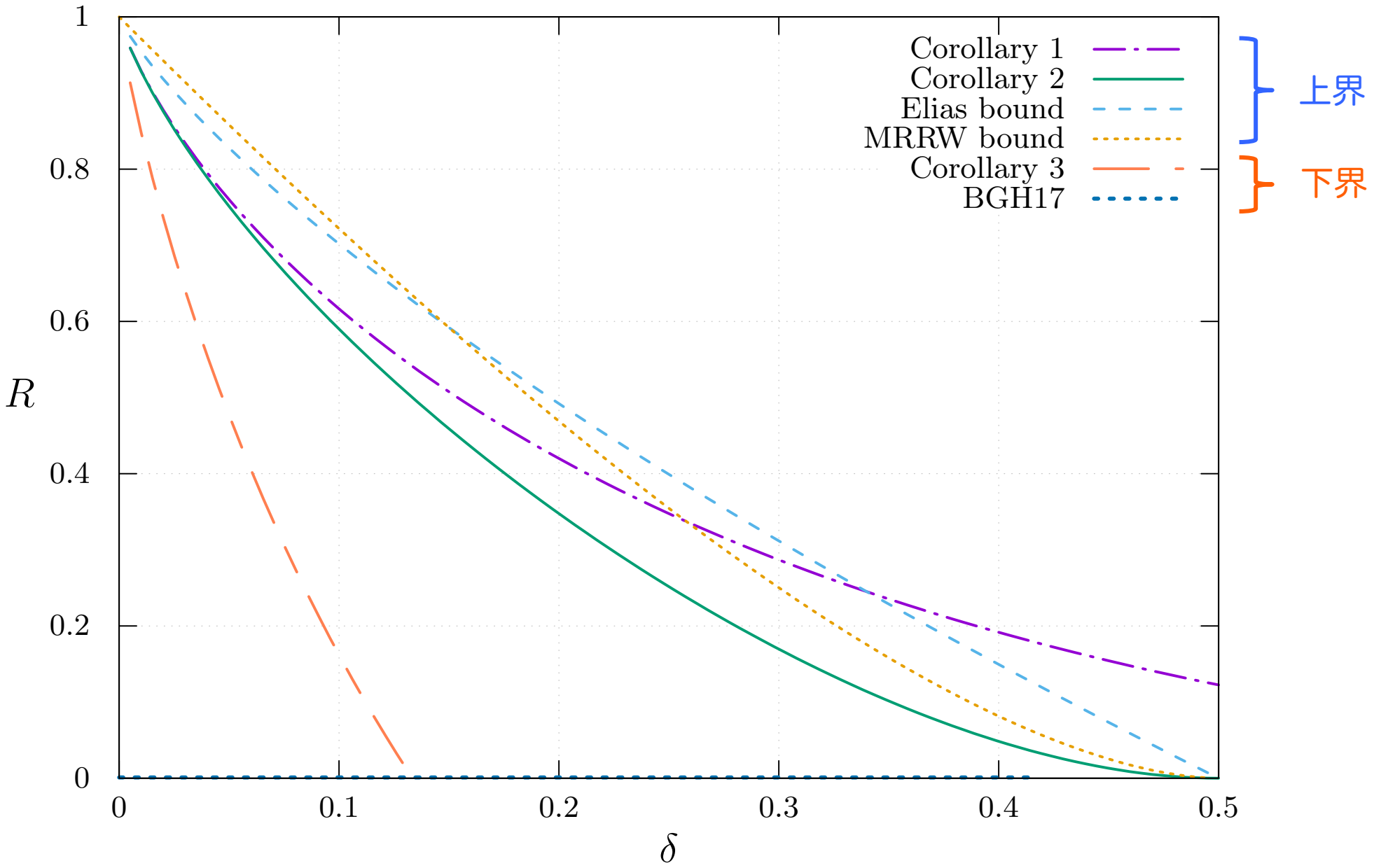
Theorem 3. 最小 Levenshtein 距離 δn, 符号化率 R の $C \subseteq \Sigma^n$ に対し,

$$R \leq 1 - H_q\left(\theta - \sqrt{\theta(\theta-\delta)}\right) + o(1) \quad \text{（Elias 限界）}$$

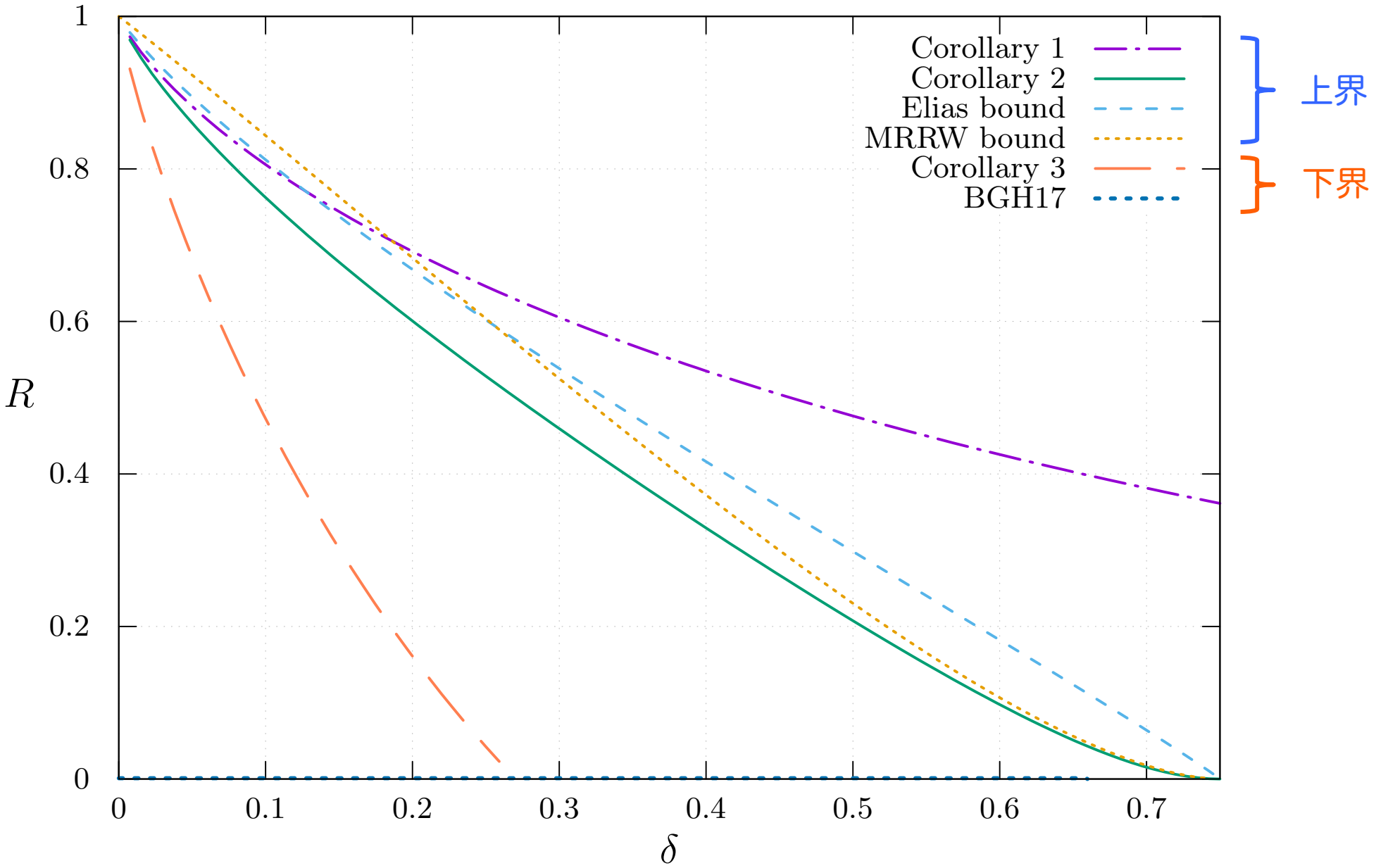
$$R \leq H_q\left(\frac{1}{q}\left(q-1-(q-2)\delta - 2\sqrt{\delta(1-\delta)(q-1)}\right)\right) + o(1) \quad \text{（MRRW 限界）}$$

ここで, $0 \leq \delta \leq \theta = 1 - \frac{1}{q}$

符号化率と相対最小距離のトレードオフ：$q = 2$



符号化率と相対最小距離のトレードオフ：$q = 4$

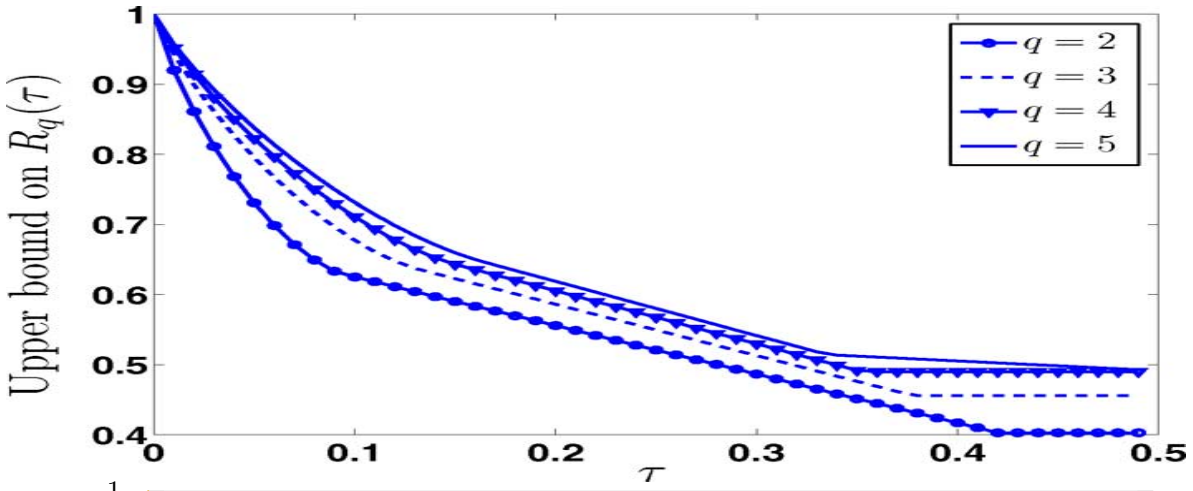


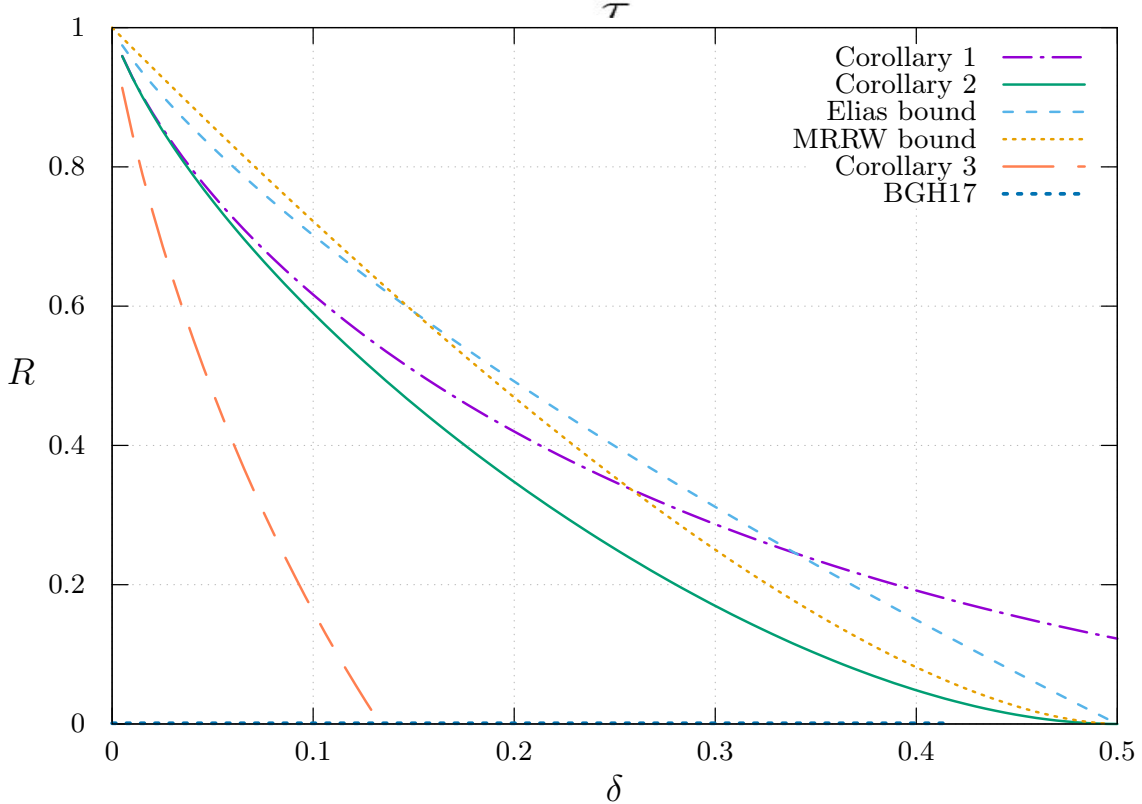
数値計算結果：上界

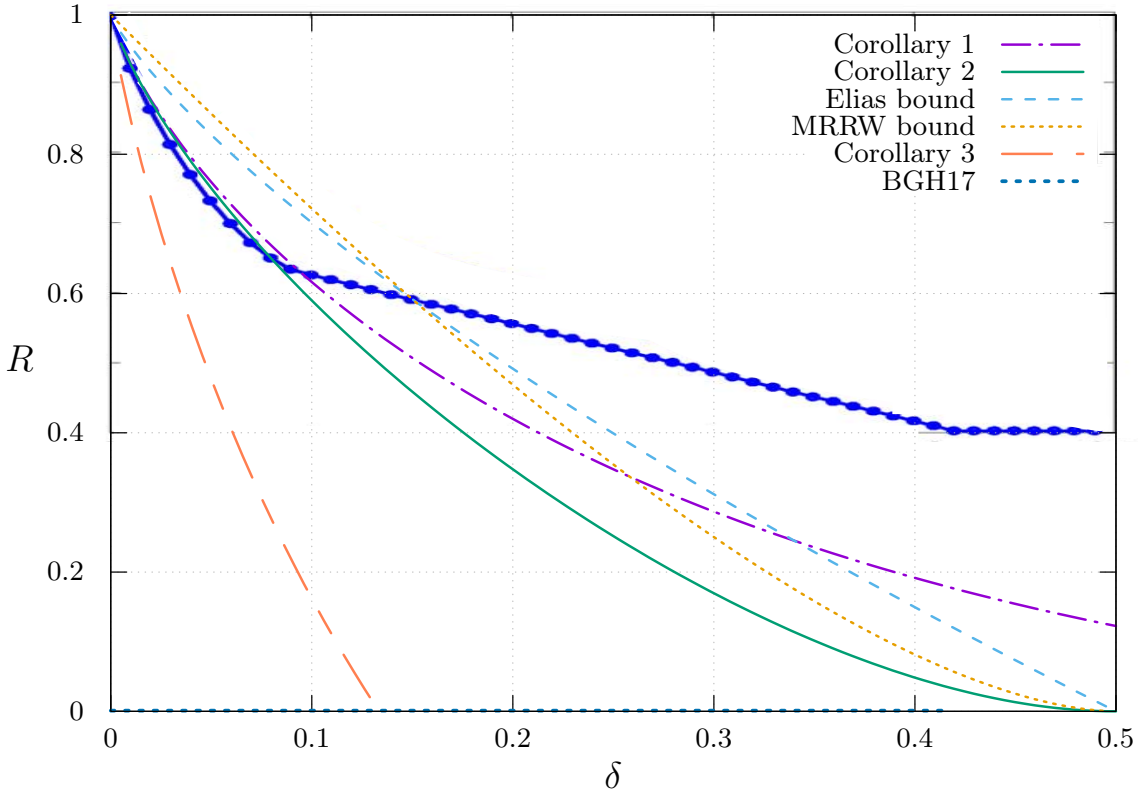
q	n	d	[Lev02]	[KK13]	Theorem 1	Theorem 2
2	20	4	97 453	55 206	95 325	181 643
2	20	10	26 456	2 535	1 295	2 452
2	20	20	190 416	1 059	32	28
2	20	30	961 048	—	5	4
2	40	4	47 498 012 376	28 192 605 878	52 357 696 560	102 167 009 660
2	40	10	1 279 636 864	9 880 934	117 292 187	228 473 245
2	40	20	1 122 371 648	3 203 459	215 900	203 859
2	40	30	13 097 807 352	298 539	3 735	1 195
2	40	40	287 193 094 240	1 048 713	231	43
4	20	4	30 003 945 118	19 289 677 788	68 719 476 736	90 174 299 388
4	20	10	360 221 648	25 002 768	306 647 351	316 287 316
4	20	20	536 774 720	1 645 315	1 258 226	79 926
4	20	30	192 278 071 952	—	34 771	71
4	40	4	$\approx 14\,843 \times 10^{18}$	$\approx 10\,332 \times 10^{18}$	$\approx 38\,997 \times 10^{18}$	$\approx 51\,574 \times 10^{18}$
4	40	10	$\approx 6\,113 \times 10^{15}$	$\approx 461\,805 \times 10^{12}$	$\approx 27\,238 \times 10^{15}$	$\approx 36\,015 \times 10^{15}$
4	40	20	$\approx 133\,526 \times 10^{12}$	$\approx 158\,374 \times 10^{9}$	$\approx 7\,269 \times 10^{12}$	$\approx 2\,172 \times 10^{12}$
4	40	40	$\approx 34\,641 \times 10^{15}$	$\approx 2\,123 \times 10^{9}$	$\approx 173\,431 \times 10^{6}$	$\approx 69 \times 10^{6}$
4	40	60	$\approx 306\,026 \times 10^{18}$	—	164 423 496	108

[KK13] A. A. Kulkarni and N. Kiyavash. Nonasymptotic upper bounds for deletion correcting codes. *IEEE Trans. Inf. Theory*, 59(8):5115–5130, 2013.

[KK13] との漸近的な比較







[Hayashi, Yasunaga (IEEE IT 2020)] のリスト復号可能性

List Decoding

- Decoder outputs a *small* list of codewords so that the list contains the transmitted codeword

- Extensively studied in Hamming metric
  - C is (t, ℓ)-list decodable (in Hamming metric)
    $\Leftrightarrow |B_H(\boldsymbol{v}, t) \cap C| \leq \ell$ for any $\boldsymbol{v} \in \Sigma^n$
    - $B_H(\boldsymbol{v}, t)$: Hamming ball of radius t centered at $\boldsymbol{v}$
    - t : list-decoding radius, ℓ : list size

- Johnson bound gives a bound on list size for $t \geq d/2$

$$\ell \leq qnd \quad \text{if} \quad t < n - \sqrt{n(n-d)}$$

q : alphabet size, d : minimum Hamming distance of C

Our Results

- Johnson-type bound in Levenshtein metric is derived
  - The result by Wachter-Zeh (ISIT 2017) has some flaws
  - Our bound is obtained by a similar approach

- The bound implies that, as long as $\ell = \mathrm{poly}(n)$,
  - $\exists$ binary code of rate $\Omega(1)$ correcting 0.707-frac. of insertions;
  - $\forall$ constant $\tau_I > 0$ and $\tau_D \in [0,1)$, $\exists q$-ary code of rate $\Omega(1)$ and $q = O(1)$ correcting τ_I-frac. of ins. and τ_D-frac. of del.

- Plotkin-type bound on code size in Levenshtein metric
  - By a simple application of Johnson-type bound

(Main Technical Lemma) Johnson-type Bound

Lemma 1

$C \subseteq \Sigma^n$ s.t. $d_L(C) = d$

For non-negative integers $t_I, t_D \leq n, N = n + t_I - t_D$, and $\boldsymbol{v} \in \Sigma^N$, let

$$\ell := |B_L(\boldsymbol{v}, t_D, t_I) \cap C|.$$

$B_L(\boldsymbol{v}, t_D, t_I)$: the set of words obtained from $\boldsymbol{v}$ by $\leq t_D$ insertions and $\leq t_I$ deletions

If $t_I < \left(\frac{d}{2} - t_D\right)\frac{n - t_D}{n - \frac{d}{2}}$, then $\ell \leq \frac{N\left(\frac{d}{2} - t_D\right)}{N\left(\frac{d}{2} - t_D\right) - t_I(n - t_D)}$

（$t_I = t, t_D = 0$ とすると）

$t < \frac{d}{2 - d/n}$ ならば，$|D_t(\boldsymbol{y}) \cap C| = \ell \leq \frac{(n+t)d}{(n+t)d - 2nt}$　（Theorem 2 で利用）

Proof Idea

- Let $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$ be the set of codewords that can be transformed to $\boldsymbol{v}$ by t_I insertions and t_D deletions

- Consider the value

  λ := Sum of pairwise distances between ℓ codewords

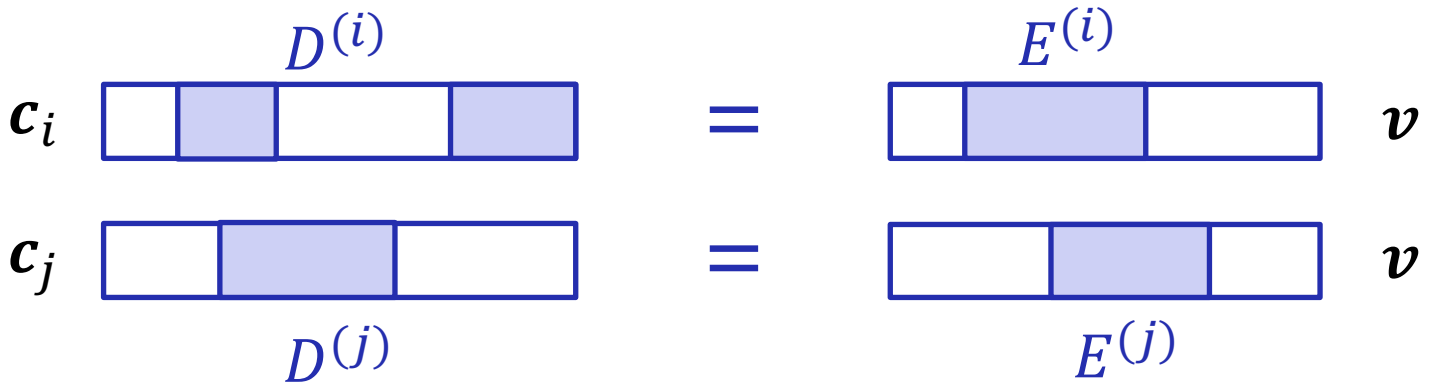
- "Double Counting" is applied to λ :

  1. Row by row → Lower bound from $d_L(\boldsymbol{c}_i, \boldsymbol{c}_j) \geq d$
  2. Column by column → Upper bound from $d_L(\boldsymbol{c}_i, \boldsymbol{c}_j) \leq d_L(\boldsymbol{c}_i, \boldsymbol{v}) + d_L(\boldsymbol{v}, \boldsymbol{c}_j)$
     - More sophisticated upper bound is used

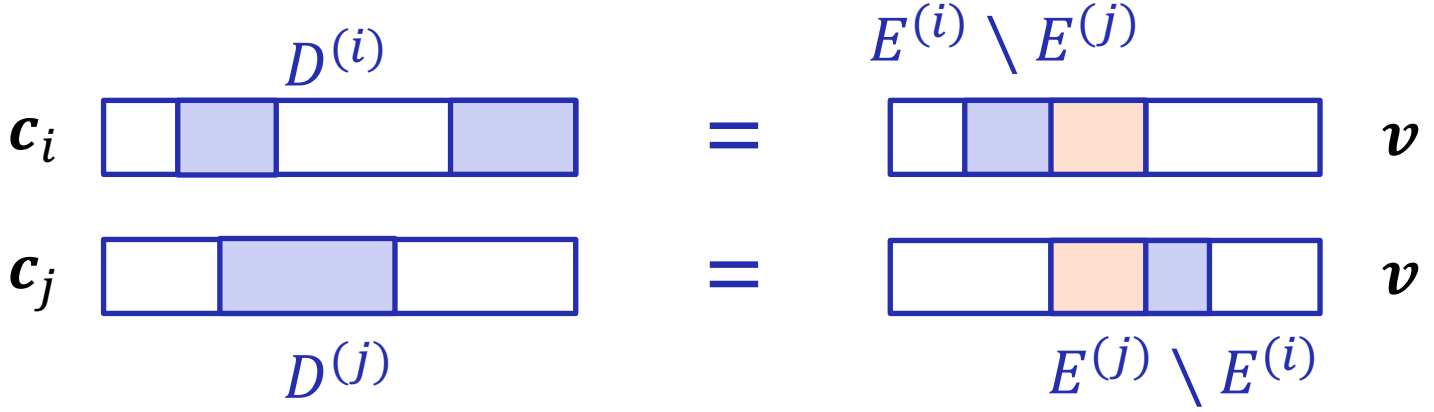
Proof of Theorem 1

- For $\boldsymbol{v} \in \Sigma^N$, let $B_L(\boldsymbol{v}, t_D, t_I) \cap C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$
- For each $\boldsymbol{c}_i$, define $D^{(i)} \subseteq [n] = \{1, \ldots, n\}$ and $E^{(i)} \subseteq [N] = \{1, \ldots, N\}$ s.t. $\boldsymbol{c}_i$ can be transformed to $\boldsymbol{v}$ by
  1. Deleting symbols in $D^{(i)}$ from $\boldsymbol{c}_i$; and
  2. Inserting symbols in $E^{(i)}$

- We can choose s.t. $\left|D^{(i)}\right| = t_D$, $\left|E^{(i)}\right| = t_I$



- $\boldsymbol{c}_i$ can be transformed to $\boldsymbol{c}_j$ by
  1. Deleting symbols in $D^{(i)}$ from $\boldsymbol{c}_i$
  2. Inserting symbols in $E^{(i)}$ to get $\boldsymbol{v}$
  3. Deleting symbols in $E^{(j)}$ from $\boldsymbol{v}$
  4. Inserting symbols in $D^{(j)}$ to get $\boldsymbol{c}_j$



- Steps 2-3 can be simplified as
  1. Deleting symbols in $D^{(i)}$ from $\boldsymbol{c}_i$
  2. Inserting symbols in $E^{(i)} \setminus E^{(j)}$ to get $\boldsymbol{v}_{|[N]\setminus(E^{(i)}\cap E^{(j)})}$
  3. Deleting symbols in $E^{(j)} \setminus E^{(i)}$ from $\boldsymbol{v}_{|[N]\setminus(E^{(i)}\cap E^{(j)})}$
  4. Inserting symbols in $D^{(j)}$ to get $\boldsymbol{c}_j$
- Thus, we have that

$$d_L(\boldsymbol{c}_i, \boldsymbol{c}_j) \leq \left|D^{(i)}\right| + \left|E^{(i)} \setminus E^{(j)}\right| + \left|E^{(j)} \setminus E^{(i)}\right| + \left|D^{(j)}\right|$$

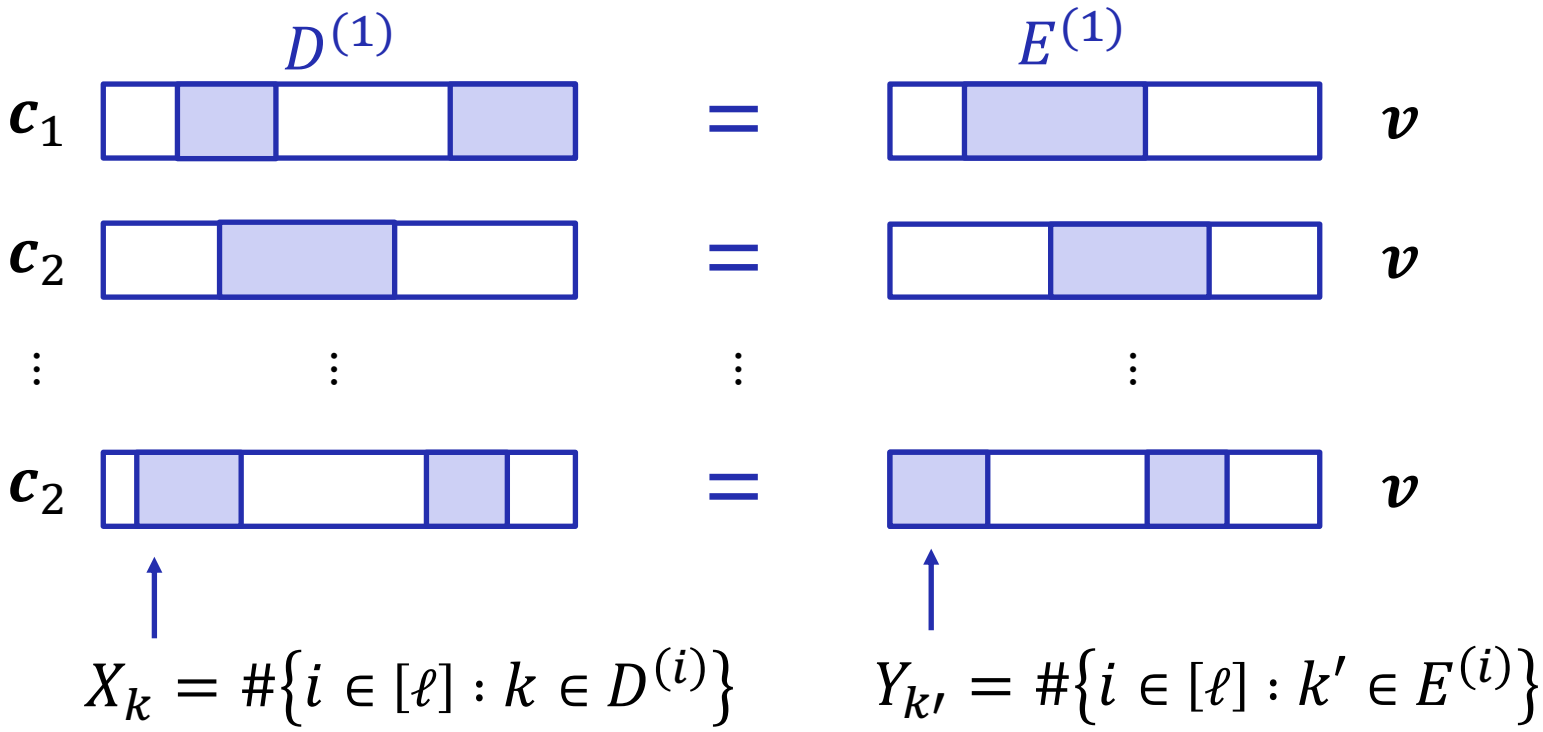
- Define $\lambda \coloneqq \sum_{i\in[\ell]} \sum_{j\in[\ell]\setminus\{i\}} d_L(\boldsymbol{c}_i, \boldsymbol{c}_j)$

- We know that
  - $\lambda \geq \ell(\ell-1)d$ (by $d_L(\boldsymbol{c}_i, \boldsymbol{c}_j) \geq d$)
  - $\lambda \leq \sum_{i\in[\ell]} \sum_{j\in[\ell]\setminus\{i\}} \left(\begin{array}{l} \left|D^{(i)}\right| + \left|E^{(i)} \setminus E^{(j)}\right| \\ +\left|E^{(j)} \setminus E^{(i)}\right| + \left|D^{(j)}\right| \end{array} \right)$

- Hence, we have

$$\ell(\ell-1)d \leq \sum_{i\in[\ell]} \sum_{j\in[\ell]\setminus\{i\}} \left(\begin{array}{l} \left|D^{(i)}\right| + \left|E^{(i)} \setminus E^{(j)}\right| \\ +\left|E^{(j)} \setminus E^{(i)}\right| + \left|D^{(j)}\right| \end{array} \right)$$

- We can show that
  - $\sum_{i\in[\ell]}\sum_{j\in[\ell]\setminus\{i\}}\left(\left|D^{(i)}\right|+\left|D^{(j)}\right|\right)=2(\ell-1)\sum_{k\in[n]}X_k$
  - $\sum_{i\in[\ell]}\sum_{j\in[\ell]\setminus\{i\}}\left(\left|E^{(i)}\setminus E^{(j)}\right|+\left|E^{(j)}\setminus E^{(i)}\right|\right)=2\sum_{k'\in[N]}Y_{k'}(\ell-Y_{k'})$



- Thus, we have

$$\ell(\ell-1)d \le 2(\ell-1)\sum_{k\in[n]} X_k + 2\sum_{k'\in[N]} Y_{k'}(\ell - Y_{k'})$$

- By using $\sum_{k\in[n]} X_k = \ell t_D$, $\sum_{k'\in[N]} Y_{k'} = \ell t_I$, we can show that

$$\ell \le \frac{N\left(\frac{d}{2} - t_D\right)}{N\left(\frac{d}{2} - t_D\right) - t_I(n - t_D)}$$

  - Both the numerator and the denominator are positive by the assumption.

QED

符号サイズの下界

平均球サイズによる下界

Tolhuizen (IEEE IT 1997). X 上の距離関数 $\rho: X \times X \to \mathbb{Z}$ に対し，

- $\boldsymbol{x}$ 中心の半径 d の球サイズ　$V_d(\boldsymbol{x}) \coloneqq |\{\boldsymbol{y} \in X : \rho(\boldsymbol{x}, \boldsymbol{y}) \leq d\}|$
- 平均球サイズ $V_d^{\mathrm{ave}} \coloneqq \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} V_d(\boldsymbol{x})$

のとき，最小距離 d の符号 C として，　$|C| \geq \frac{|X|}{V_{d-1}^{\mathrm{ave}}}$ が存在

Levenshtein (ISIT 2002). 任意の $1 \leq t \leq n$ に対し，以下を満たす最小 Levenshtein 距離 $d = 2(t+1) = 2\delta n$ の符号 C が存在.

$|C| \geq \frac{q^{n+t}}{I_q(n-t,t)^2}$　　つまり　符号化率 $R \geq 1 + \delta - 2H_q(\delta)$

グラフ理論による下界

集合 $X \subseteq \Sigma^n$ に対し，グラフ $G = (V, E)$ を

- 各 $\boldsymbol{x} \in X$ が頂点，$d_L(\boldsymbol{x}, \boldsymbol{y}) \leq d - 1$ なら辺 $(\boldsymbol{x}, \boldsymbol{y})$ が存在

と定めると，

- G の独立集合の最大サイズ $= \alpha(G)$ $\Leftrightarrow$ $A_q(n, d) = \alpha(G)$

- 独立数 (independence number) $\alpha(G)$ に関する
  グラフ理論の結果が利用できる
  - Turán の定理より GV 限界が導かれる (Tohluizen (1997))
  - Jiang, Vardy (IEEE IT 2004) による GV 限界の改良もこれ

Caro-Wei 限界による下界

Caro-Wei 限界

$$\alpha(G) \geq \sum_{x \in V} \frac{1}{1 + \deg(x)}$$

Theorem 5. 任意の $d = 2(t+1) < n$, 整数 $1 \leq r \leq n$ に対し,

$$A_q(n,d) \geq \left\lfloor \frac{\left(q^n - \sum_{i=r}^{n} \binom{n-1}{i-1} q(q-1)^{i-1}\right)^2}{I_q(n-t,t)\left(q^{n-t} \cdot I_q(n,t) - \sum_{i=r}^{n} \binom{n-1}{i-1} q(q-1)^{i-1} \cdot \left(\sum_{j=1}^{t} \binom{i-t}{j}\right)\right)} \right\rfloor$$

証明：集合 $X \subseteq \Sigma^n$ をラン数 r 以上の文字列集合とし,
Caro-Wei 限界に $\deg(x) \leq \left|L_{t,t}(\boldsymbol{x})\right| \leq |D_t(\boldsymbol{x})| \cdot I_q\,(n-t,t)$ を適用

Sala, Gabrys, Dolecek (ISIT 2014) の下界式（式自体は省略）

グラフの三角形数 T を用いた以下を利用（Δ は最大次数）

$$\alpha(G) \geq \frac{|V|}{10\Delta}\left(\log\Delta - \frac{1}{2}\log\left(\frac{T}{|V|}\right)\right)$$

Jiang, Vardy (2004) は $\frac{T}{|V|}$ の上界を与えて GV 限界を $\log n$ 倍改善

Sala, Gabrys, Dolecek (ISIT 2014) も漸近的に $\log n$ 倍改善??

However, the present work is focused on non-asymptotic bounds. To the best of the authors' knowledge, Theorem 1 is so far the strongest lower bound on deletion-correcting codes, with an improvement on the order of log n over all existing bounds, in both the asymptotic and non-asymptotic cases.

と記載はあるが・・・

Alon, Bourla, Graham, He, Kravitz (IEEE IT 2023) が上記を使った改善

平均挿入・削除球サイズ上界の改良による下界

Levenshtein (ISIT 2002) は

$$\left|L_{t,t}(\boldsymbol{x})\right| \leq |D_t(\boldsymbol{x})| \cdot I_q\,(n-t,t)$$

から，以下を利用

$$V_d^{\mathrm{ave}} := \frac{1}{|\Sigma^n|}\sum_{\boldsymbol{x}\in\Sigma^n}\left|L_{t,t}(\boldsymbol{x})\right| \leq \frac{I_q(n-t,t)}{q^n}\sum_{\boldsymbol{x}\in\Sigma^n}|D_t(\boldsymbol{x})| = \frac{I_q(n-t,t)^2}{q^t}$$

➔ 挿入・削除球サイズ $\left|L_{t,t}(\boldsymbol{x})\right|$ の上界の改良を目指す

$|L_{t,t}(\boldsymbol{x})|$ の上界の改良

アイディア：数え上げの重複を考える

$D_t(000111100000111)$ において，

$\boldsymbol{x} = 00011110000111$ の黒い部分を削除したものを $\boldsymbol{y}$ とおく．

オレンジのペアの片方を一つずつ削除すると，
その結果 $\boldsymbol{z}$ は $2^3 = 8$ 通りあり，すべて $\boldsymbol{y}$ の部分文字列

➔ 各 $|I_t(\boldsymbol{z})|$ において，　$|I_{t-3}(\boldsymbol{y})|$ は重複して数え上げられている

➔ $7|I_{t-3}(\boldsymbol{y})|$ は数えなくてよい

主結果その２：平均挿入・削除球サイズ上界の改良による下界

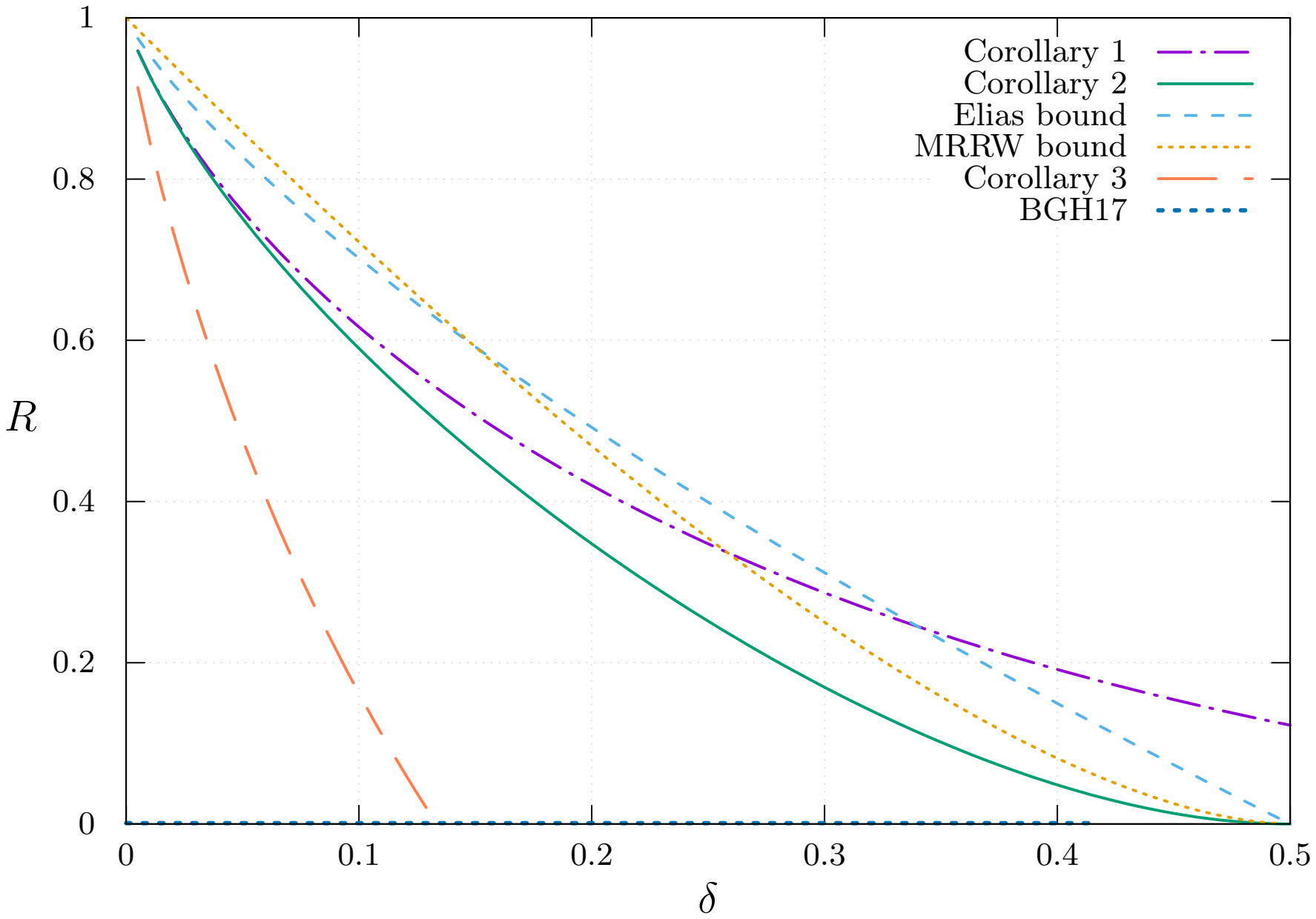
ラン数 $r(\boldsymbol{x}) \geq 2$ のときオレンジペアが１つ以上あることを利用

Corollary 3. 任意の $1 \leq t \leq n$ に対し，以下を満たす
最小距離 $d = 2(t+1)$ の符号 C が存在.

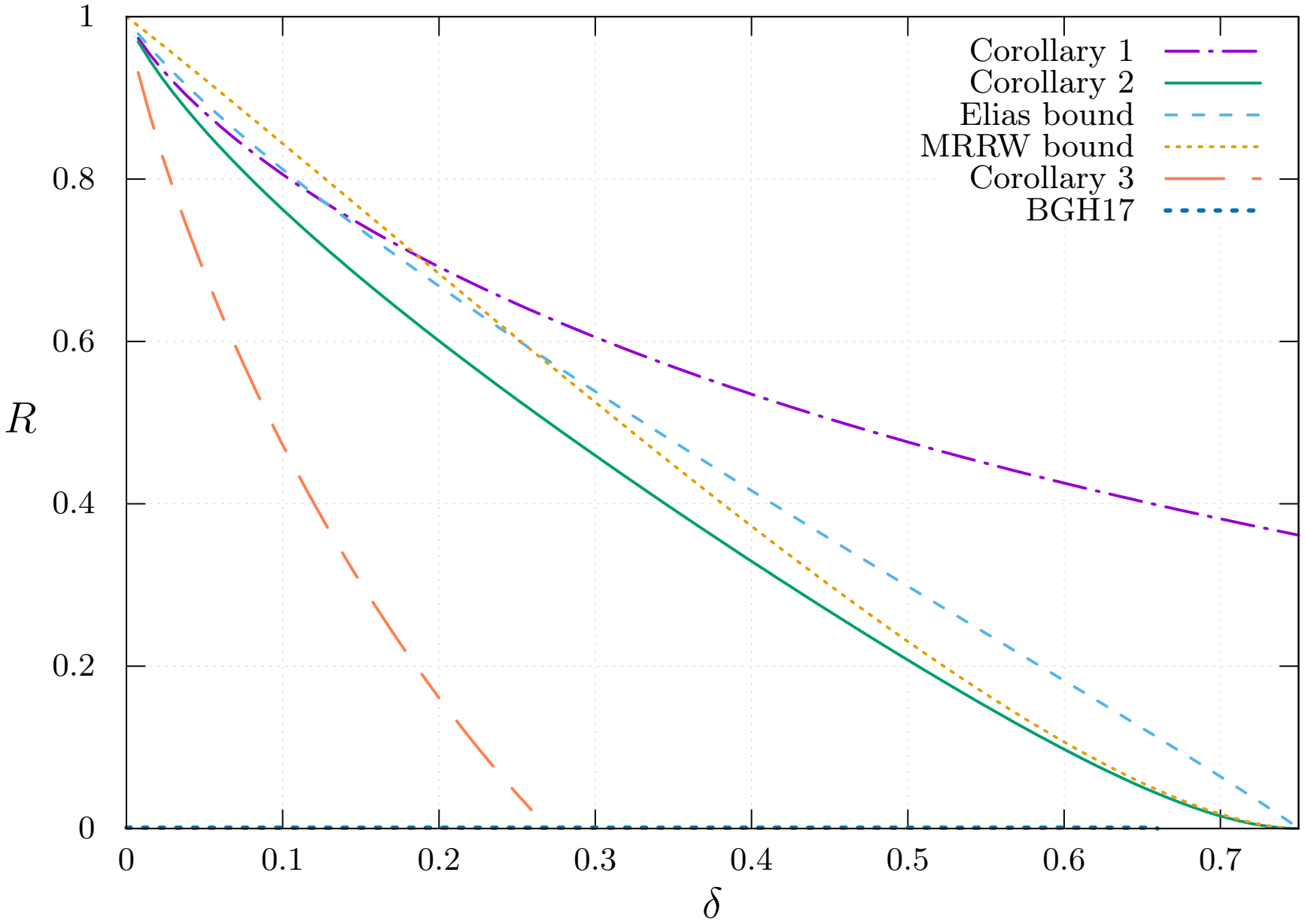
$$A_q(n,d) \geq \left\lfloor \frac{q^{n+t}}{I_q(n-t,t)^2 - (q^t - q^{-n+t+1})I_q(n-t+1,t-1)} \right\rfloor$$

符号化率は Levenshtein(2002) と同様で，$R \geq 1 + \delta - 2H_q(\delta)$

符号化率と相対最小距離のトレードオフ：$q = 2$



符号化率と相対最小距離のトレードオフ：$q = 4$



数値計算結果：上界

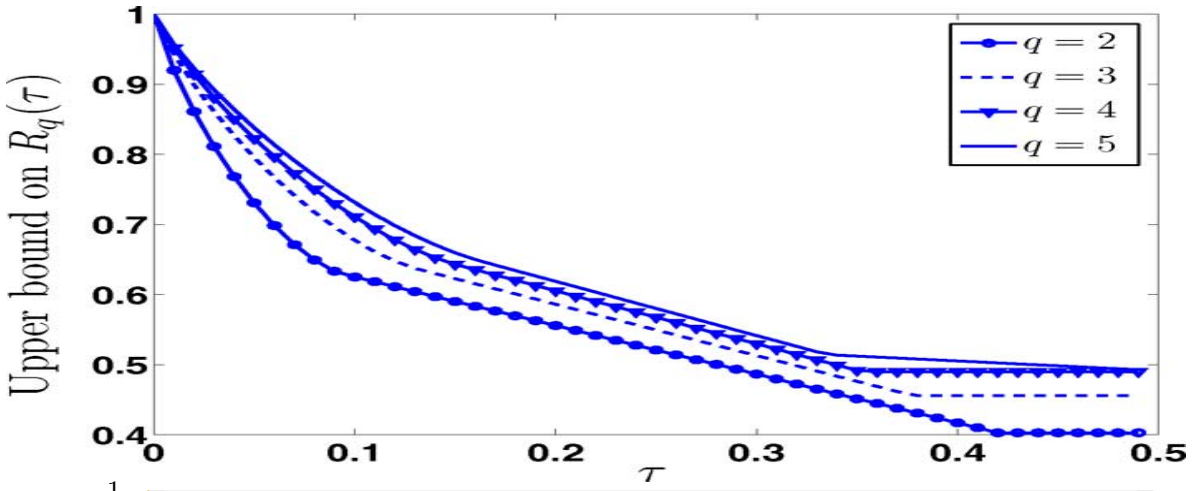
q	n	d	[Lev02]	[KK13]	Theorem 1	Theorem 2
2	20	4	97 453	55 206	95 325	181 643
2	20	10	26 456	2 535	1 295	2 452
2	20	20	190 416	1 059	32	28
2	20	30	961 048	—	5	4
2	40	4	47 498 012 376	28 192 605 878	52 357 696 560	102 167 009 660
2	40	10	1 279 636 864	9 880 934	117 292 187	228 473 245
2	40	20	1 122 371 648	3 203 459	215 900	203 859
2	40	30	13 097 807 352	298 539	3 735	1 195
2	40	40	287 193 094 240	1 048 713	231	43
4	20	4	30 003 945 118	19 289 677 788	68 719 476 736	90 174 299 388
4	20	10	360 221 648	25 002 768	306 647 351	316 287 316
4	20	20	536 774 720	1 645 315	1 258 226	79 926
4	20	30	192 278 071 952	—	34 771	71
4	40	4	$\approx 14\,843 \times 10^{18}$	$\approx 10\,332 \times 10^{18}$	$\approx 38\,997 \times 10^{18}$	$\approx 51\,574 \times 10^{18}$
4	40	10	$\approx 6\,113 \times 10^{15}$	$\approx 461\,805 \times 10^{12}$	$\approx 27\,238 \times 10^{15}$	$\approx 36\,015 \times 10^{15}$
4	40	20	$\approx 133\,526 \times 10^{12}$	$\approx 158\,374 \times 10^{9}$	$\approx 7\,269 \times 10^{12}$	$\approx 2\,172 \times 10^{12}$
4	40	40	$\approx 34\,641 \times 10^{15}$	$\approx 2\,123 \times 10^{9}$	$\approx 173\,431 \times 10^{6}$	$\approx 69 \times 10^{6}$
4	40	60	$\approx 306\,026 \times 10^{18}$	—	164 423 496	108

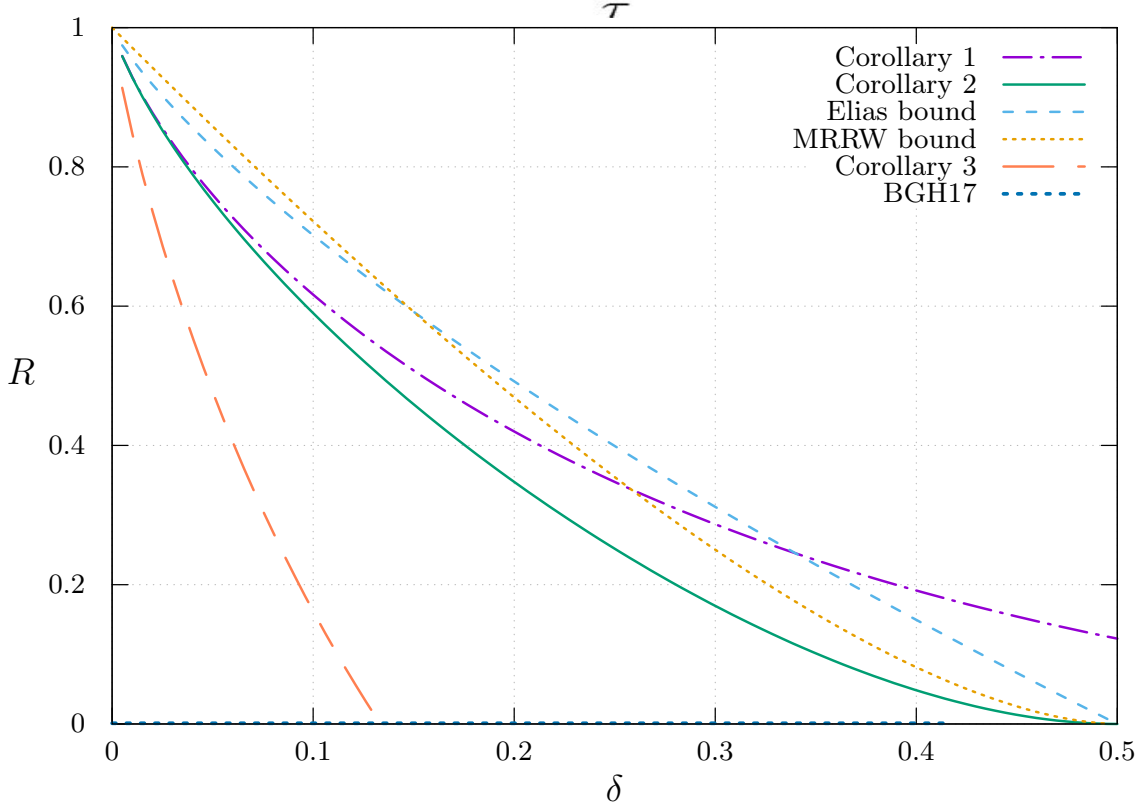
数値計算結果：下界

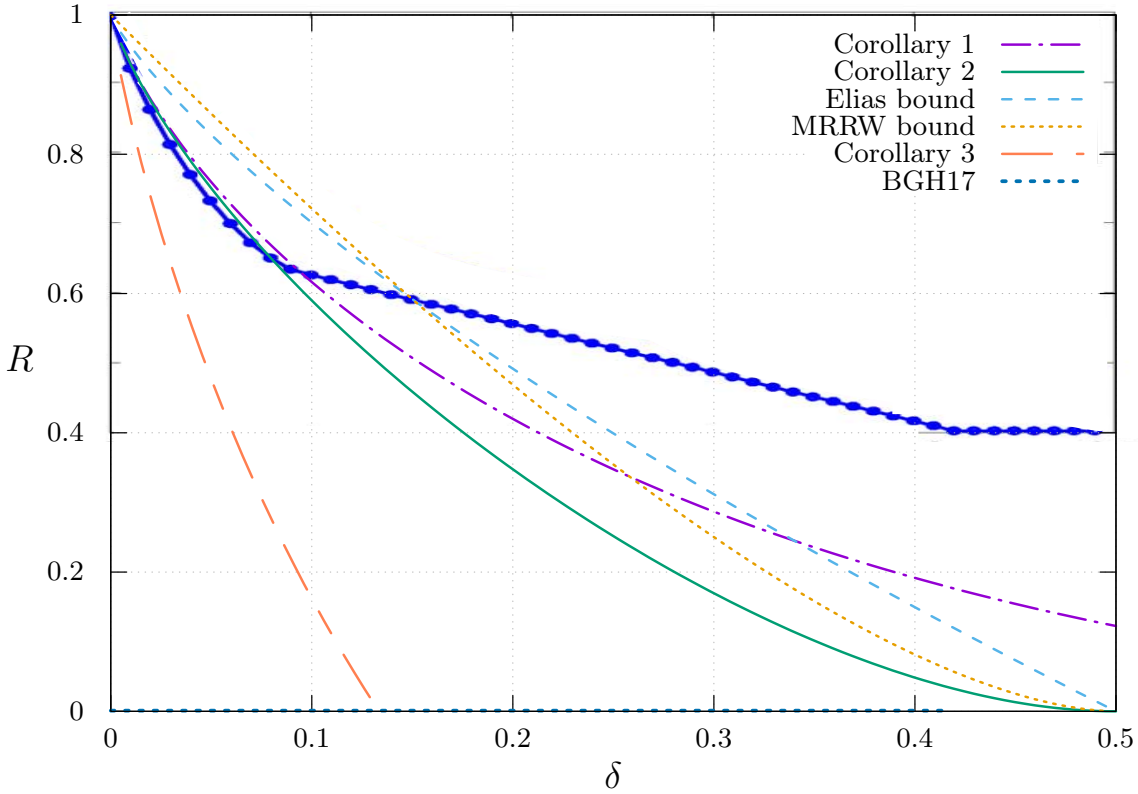
q	n	d	[Lev02]	[SGD14]	Theorem 5	Corollary 3
2	10	4	16	2	18	17
2	10	6	1	0	1	1
2	10	8	0	0	0	0
2	10	10	0	0	0	0
2	10	12	0	0	0	0
2	10	14	—	—	—	—
2	10	16	—	—	—	—
2	10	18	—	—	—	—
2	10	20	—	—	—	—
2	20	4	4755	783	4968	4777
2	20	6	94	10	94	94
2	20	8	4	0	4	4
2	20	10	0	0	0	0
2	20	12	0	0	0	0
2	20	14	0	0	0	0
2	20	16	0	0	0	0
2	20	18	0	0	0	0
2	20	20	0	0	0	0
2	20	22	0	0	0	0
2	40	4	1 308 163 745	244 663 405	1 339 190 459	1 309 722 010
2	40	6	6 524 894	881 891	6 532 808	6 526 482
2	40	8	76 814	6 032	71 601	76 818
2	40	10	1 687	68	1 396	1 687
2	40	12	60	1	42	60
2	40	14	3	0	1	3
2	40	16	0	0	0	0
2	40	18	0	0	0	0
2	40	20	0	0	0	0
2	40	22	0	0	0	0

q	n	d	[Lev02]	[SGD14]	Theorem 5	Corollary 3
4	10	4	4 364	842	4 489	4 382
4	10	6	88	14	78	88
4	10	8	4	0	3	4
4	10	10	0	0	0	0
4	10	12	0	0	0	0
4	10	14	—			
4	10	16	—			
4	10	18	—			
4	10	20	—			
4	20	4	1 181 952 838	269 953 863	1 200 316 339	1 183 224 781
4	20	6	5 608 964	1 260 800	5 257 096	5 610 710
4	20	8	66 412	11 205	52 137	66 419
4	20	10	1 558	167	937	1 558
4	20	12	64	3	27	64
4	20	14	4	0	1	4
4	20	16	0	0	0	0
4	20	18	0	0	0	0
4	20	20	0	0	0	0
4	20	22	0	0	0	0
4	40	10	5 251 871 945 006	968 893 684 250	4 033 370 043 313	5 251 878 194 182
4	40	12	4 408 536 581	5 621 632 730	28 003 006 604	4 408 537 815
4	40	14	562 976 279	46 013 071	281 585 593	562 976 323
4	40	16	10 353 270	506 907	3 886 646	10 353 271
4	40	18	263 527	7 256	70 970	263 527
4	40	20	9 010	131	1 673	9 010
4	40	22	404	2	50	404

[KK13] との漸近的な比較







まとめ

挿入・削除を訂正する最良符号のサイズの上界・下界を導出

- 球充填タイプの上界 (Theorem 1, Corollary 1)
- Elias タイプの上界 (Theorem 2, Corollary 2)
  - [HY20] のリスト復号可能性を利用
- [Lev02] を改良した下界 (Corollary 3)
  - 平均挿入・削除球サイズ上界の改良を利用
  - 漸近的には [Lev02] と等価

未解決問題

- エクスパンダーグラフの活用
- 下界の漸近的な改善（ランダム符号化は最適か？？）

参考文献

[2] B. Bukh, V. Guruswami, and J. Håstad. An improved bound on the fraction of correctable deletions. *IEEE Trans. Inf. Theory*, 63(1):93–103, 2017.

[4] D. Cullina and N. Kiyavash. An improvement to Levenshtein's upper bound on the cardinality of deletion correcting codes. *IEEE Trans. Inf. Theory*, 60(7):3862–3870, 2014.

[5] V. Guruswami, X. He, and R. Li. The zero-rate threshold for adversarial bit-deletions is less than 1/2. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 727–738. IEEE, 2021.

[7] T. Hayashi and K. Yasunaga. On the list decodability of insertions and deletions. *IEEE Trans. Inf. Theory*, 66(9):5335–5343, 2020.

[10] A. A. Kulkarni and N. Kiyavash. Nonasymptotic upper bounds for deletion correcting codes. *IEEE Trans. Inf. Theory*, 59(8):5115–5130, 2013.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[12] V. I. Levenshtein. Bounds for deletion/insertion correcting codes. In *Proceedings IEEE International Symposium on Information Theory*, page 370, 2002.

[14] F. Sala, R. Gabrys, and L. Dolecek. Gilbert-varshamov-like lower bounds for deletion-correcting codes. In *2014 IEEE Information Theory Workshop, ITW 2014, Hobart, Tasmania, Australia, November 2-5, 2014*, pages 147–151. IEEE, 2014.