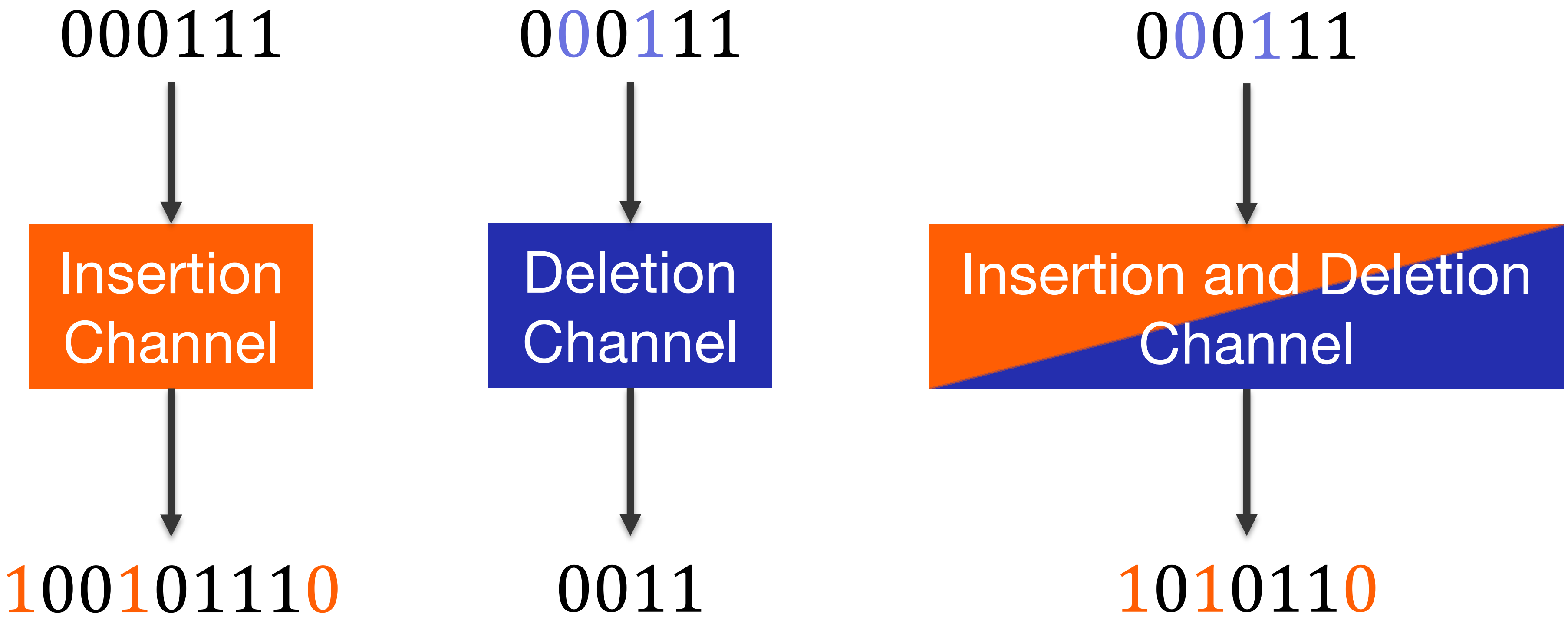
挿入と削除に対するリスト復号

安永憲司（大阪大学）

林智弘（金沢大学）との共同研究

第７回誤り訂正符号のワークショップ@盛岡市清温荘
2018.9.3

Insertions and Deletions



000111

Insertion Channel

100101110

000111

Deletion Channel

0011

000111

Insertion and Deletion Channel

1010110

Levenshtein distance

- $d_L(\boldsymbol{x},\boldsymbol{y})$:= min { #(ins./del.) to transform $\boldsymbol{x}$ into $\boldsymbol{y}$ }
  - Ex. $d_L(000,111)=6,\quad d_L(101,010)=2$

- Minimum Levenshtein distance of a code C :

$$d_L(C) \coloneqq \min_{\boldsymbol{c}_1 \neq \boldsymbol{c}_2 \in C} d_L(\boldsymbol{c}_1,\boldsymbol{c}_2)$$

- If $d_L(C) \geq d$, C can (uniquely) correct total t insertions/deletions for $t < d/2$

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 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 $\ell = \mathrm{poly}(n)$,
  - ∃ binary code of rate $\Omega(1)$ correcting 0.707-frac. of insertions;
  - ∀ 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

List Decoding in Levenshtein metric

- C is (t_I, t_D, ℓ)-list decodable

⇔ $\exists$ decoder s.t. $\forall \boldsymbol{c} \in C$, when $\leq t_I$ insertions and $\leq t_D$ deletions occur, the decoder outputs a list of size $\leq \ell$ that contains $\boldsymbol{c}$

⇔ $|B_L(\boldsymbol{v}, t_D, t_I) \cap C| \leq \ell$ for any $\boldsymbol{v} \in \Sigma^*$

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

(Main Theorem) Johnson-type Bound

Theorem 1

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

For non-negative integers $t_I, t_D, N \in [n - t_D, n + t_I]$, and $\boldsymbol{v} \in \Sigma^N$, let $\ell := |B_L(\boldsymbol{v}, t_D, t_I) \cap C|$ be the maximum list size when $\boldsymbol{v}$ is received.

Let t_I', t_D' be the maximum integers s.t. $t_I' - t_D' = N - n$, $t_I' \le t_I, t_D' \le t_D$

If $\underline{\frac{d}{2} > t_D' + \frac{t_I'(n-t_D')}{N},}$ (*) then $\ell \le \frac{N\left(\frac{d}{2} - t_D'\right)}{N\left(\frac{d}{2} - t_D'\right) - t_I'(n-t_D')}$

Proof Idea

- Let $\{\boldsymbol{c}_1, \dots, \boldsymbol{c}_\ell\}$ be the set of codewords that can be transformed to $\boldsymbol{v}$ by $\leq t_I$ insertions and $\leq t_D$ deletions
  - W.l.o.g, we assume that every $\boldsymbol{c}_i$ can be transformed to $\boldsymbol{v}$ by exactly t_I' insertions and t_D' deletions

- Consider the value

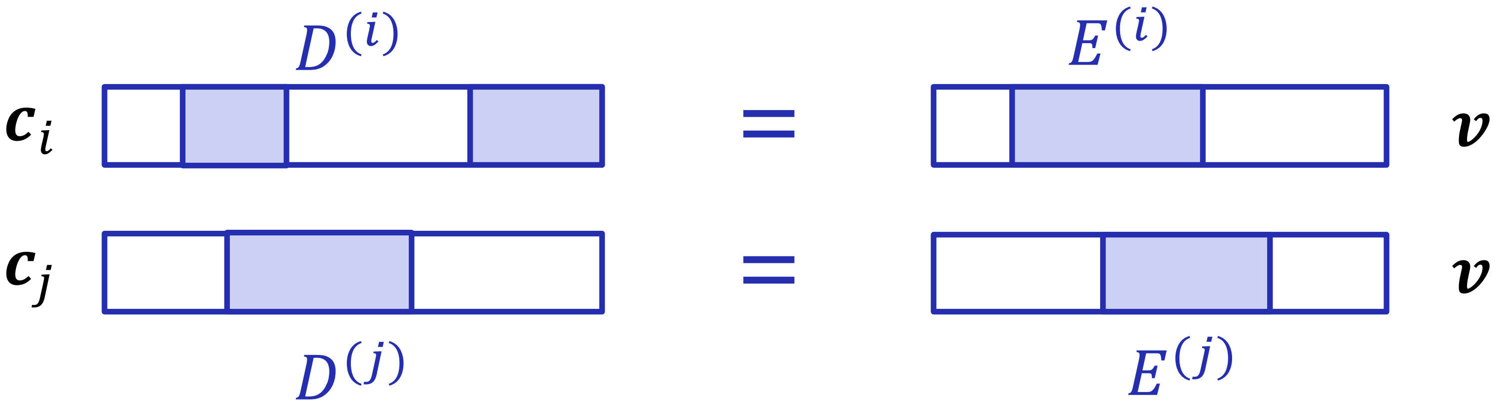
  λ := sum of pairwise distances between ℓ codewords

- "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

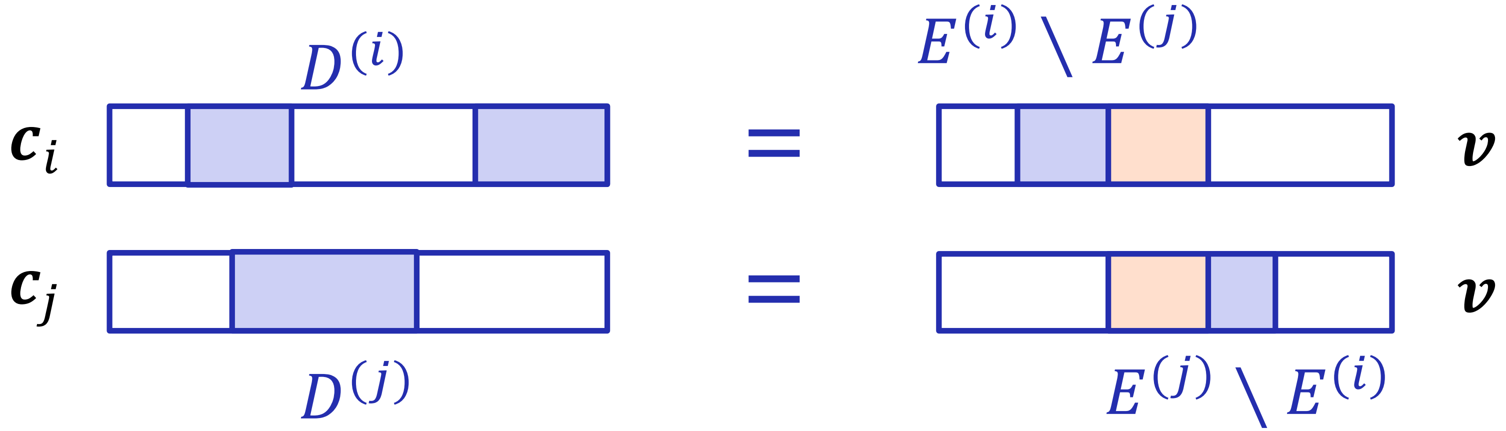
- For $\boldsymbol{v} \in \Sigma^N$, let $B_L(\boldsymbol{v}, t_D, t_I) \cap C = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_\ell\}$
- For each $\boldsymbol{c}_i$, define $D^{(i)} \subseteq [n] = \{1, \dots, n\}$ and $E^{(i)} \subseteq [N] = \{1, \dots, N\}$ s.t. $\boldsymbol{c}_i$ can be transformed to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

- Note that $\left|D^{(i)}\right| = t'_D$, $\left|E^{(i)}\right| = t'_I$



- $\boldsymbol{c}_i$ 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

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

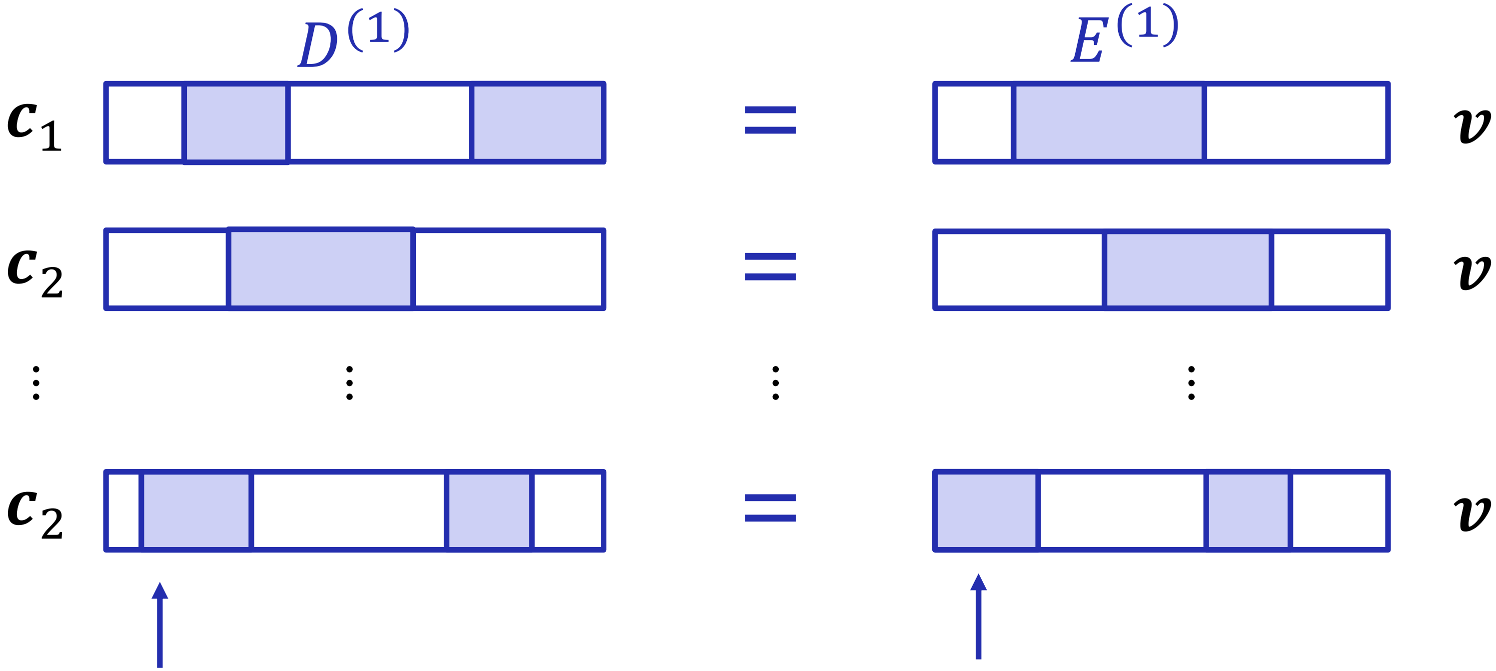
- Define λ $:= \sum_{i\in[\ell]} \sum_{j\in[\ell]\setminus\{i\}} d_L\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right)$

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

- Hence, we have

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

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



$X_k = \#\left\{i \in [\ell] : k \in D^{(i)}\right\}$ $\quad$ $Y_{k'} = \#\left\{i \in [\ell] : k' \in E^{(i)}\right\}$

- Thus, we have

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

- By using $\sum_{k\in[n]} X_k = \ell t'_D$, $\sum_{k'\in[N]} Y_{k'} = \ell t'_I$, we can show that

$$\ell \leq \frac{N\left(\frac{d}{2} - t'_D\right)}{N\left(\frac{d}{2} - t'_D\right) - t'_I(n - t'_D)}$$

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

QED

Discussion

- NOTE : $(*)$ is a condition for $\underline{t_I' \text{ and } t_D'}$, not for $\underline{t_I \text{ and } t_D}$

  Numbers of errors　　Their upper bounds

- We can see that $(*)$ is equivalent to

$$t_I' < \frac{\left(\frac{d}{2} - t_D'\right)(n - t_D')}{n - \frac{d}{2}} \quad \cdot \cdot \cdot \quad (**)$$

  - RHS of $(**)$ is monotonically decreasing on t_D'

⇨ If $(**)$ is satisfied for $t_I' = t_I,\ t_D' = t_D$,
then $(**)$ is satisfied for all $t_I' \leq t_I,\ t_D' \leq t_D$

- Hence, the following $(***)$ can be used for bounds on t_I and t_D

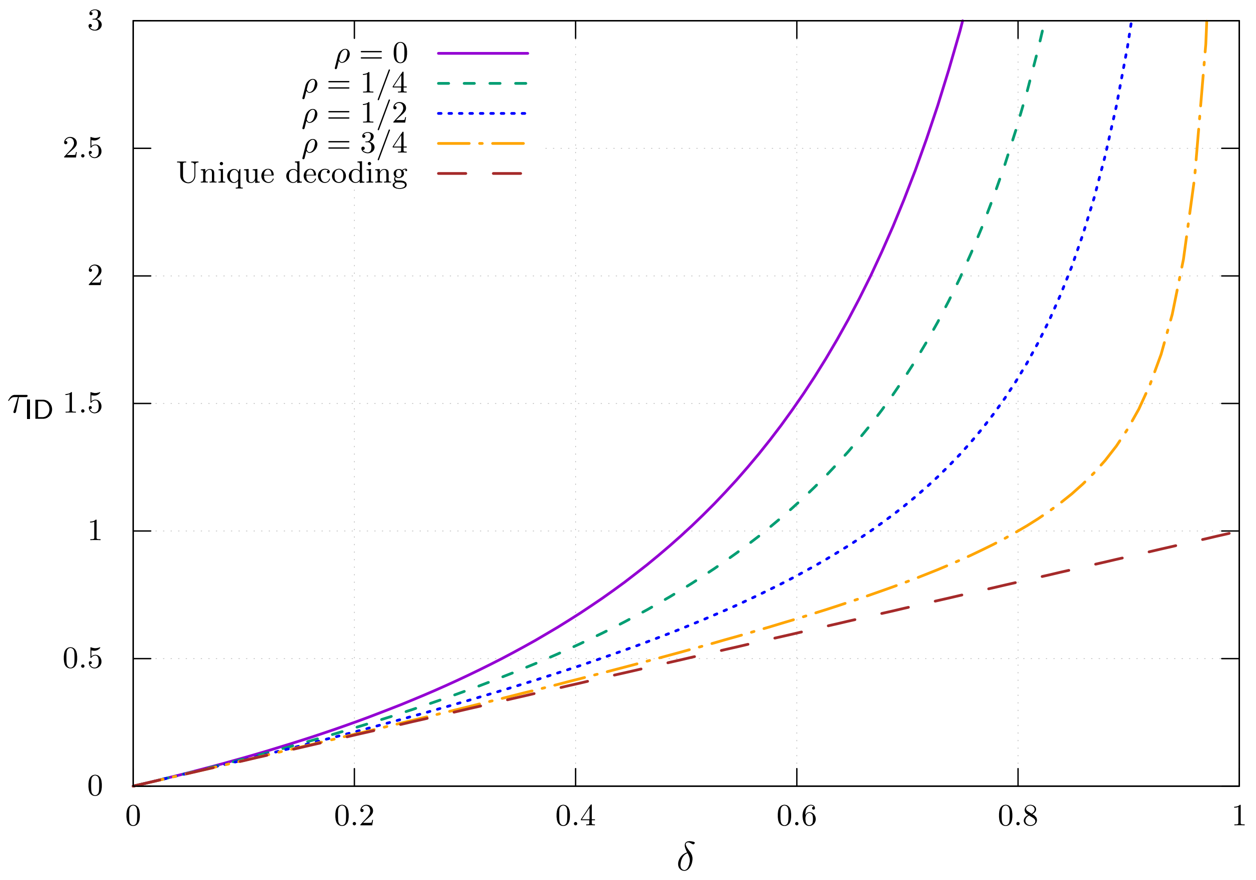
$$t_I < \frac{\left(\frac{d}{2} - t_D\right)(n - t_D)}{n - \frac{d}{2}} \quad \cdot \cdot \cdot \quad (***)$$

Bounds on t_I and t_D

- Condition $(***)$ is equivalent to
  - $\frac{t_I+t_D}{n} < \delta + \frac{(1-\rho)^2\delta^2}{1-\delta} := \tau_{ID}(\delta,\rho)$
  - $\frac{t_I}{n} < \frac{(1-\rho)\delta\,(1-\rho\delta)}{1-\delta} := \tau_I\,(\delta,\rho)$
  - $\frac{t_D}{n} < \frac{1+\delta-\sqrt{(1-\delta)(4\tau_{\text{ins}}+1-\delta)}}{2} := \tau_D(\delta,\tau_{\text{ins}})$

  where $\delta := \frac{d}{2n}, t_D := \rho\left(\frac{d}{2}\right), t_I := \tau_{\text{ins}}n$

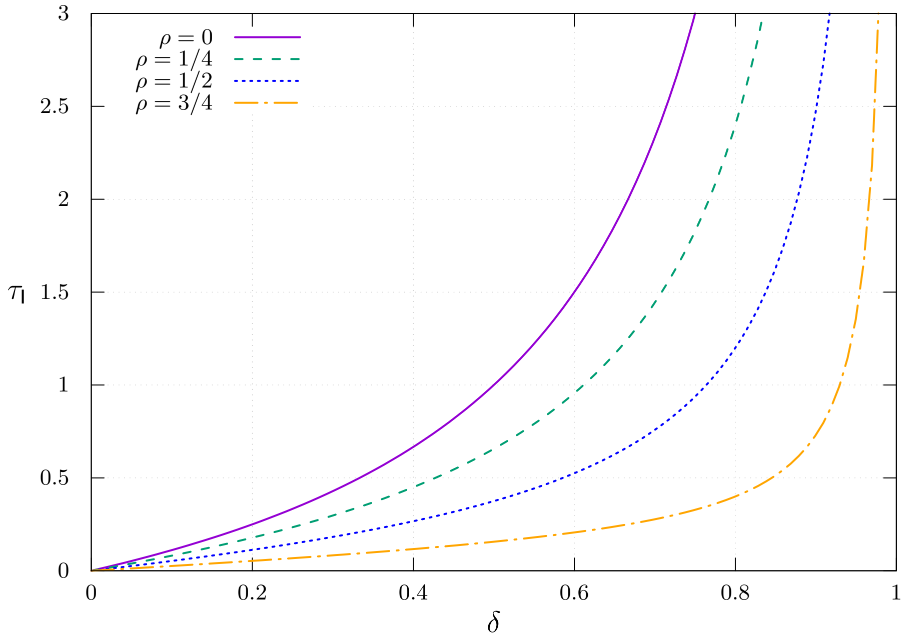
  for $\delta \in [0,1), \rho \in [0,1), \tau_{\text{ins}} \geq 0$

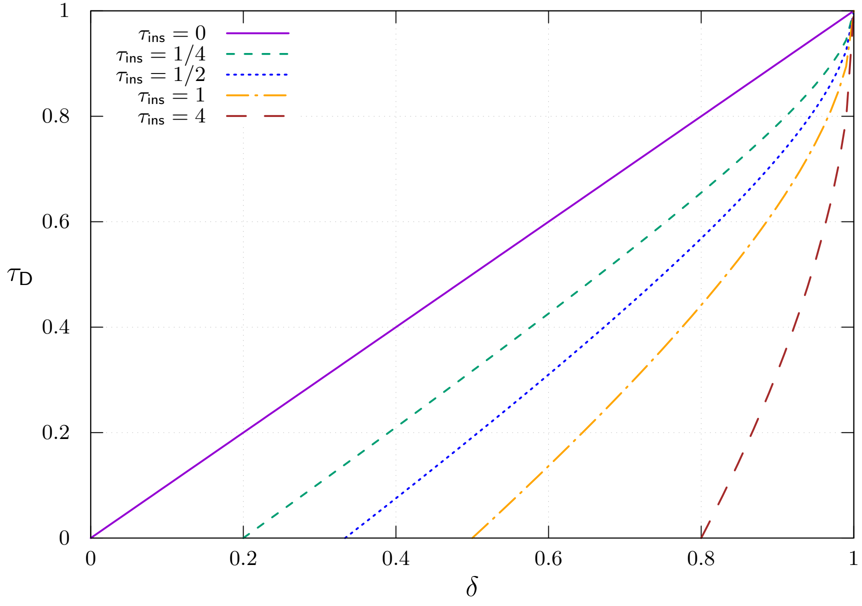


$$\delta := \frac{d}{2n}$$

$$t_D := \rho\left(\frac{d}{2}\right)$$

$$t_I := \tau_{\text{ins}} n$$





Corollary 1

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

For non-negative integers $t_I, t_D \in \left[0, \frac{d}{2}\right), N \in [n - t_D, n + t_I]$,

let $\ell := \max_{\boldsymbol{v} \in \Sigma^N} |B_L(\boldsymbol{v}, t_D, t_I) \cap C|$.

Let $\delta := \frac{d}{2n}$ and $t_D := \rho\left(\frac{d}{2}\right)$ for $\rho \in [0,1)$.

If $\frac{t_I + t_D}{n} < \delta + \frac{(1-\rho)^2\delta^2}{1-\delta} = \tau_{ID}(\delta, \rho)$, then $\ell \le (n + t_I)d$.

Specifically, for any $\tau_I^* > 0, \tau_D^* \in [0,1)$,
if ∃code with $\delta \in [0,1)$ satisfying (A), the code is
$(\tau_I^* n, \tau_D^* n, \ell)$-list decodable for $\ell \le 1 + \frac{\tau_I^*}{\delta - \tau_D^* - \tau_I^*}$.

$$\tau_I^* + \tau_D^* < \tau_{ID}\left(\delta, \frac{\tau_D^*}{\delta}\right) \Leftrightarrow \delta > \frac{\tau_I^* + \tau_D^*(1 - \tau_D^*)}{\tau_I^* + 1 - \tau_D^*} \quad \cdot \cdot \cdot \text{(A)}$$

Existence of list-decodable codes

- [Bukh, Guruswami, Hastad (IEEE IT 2017)] :
  $\forall q \geq 2, \exists q$-ary code of $\delta \approx 1 - \frac{2}{q+\sqrt{q}}$ and rate $\Omega(1)$
  ($\exists$ binary code of $\delta \approx 0.414$)

  → $\exists$ binary code of rate $\Omega(1)$ list-decodable
  0.707-frac. of ins. (or 0.414-frac. of del.)

- [Guruswami, Wang (IEEE IT 2017)] : $\forall \varepsilon > 0, \exists q$-ary code of $\delta = 1 - \varepsilon,\ q = O(\varepsilon^{-3})$ and rate $\Omega(1)$

  → $\tau_I^* > 0, \tau_D^* \in [0,1)$, $\exists q$-ary code of $q = O(1)$ and
  rate $\Omega(1)$ list-decodable against
  τ_I^*-frac. of insertions and τ_D^*-frac. of deletions

Recent Results

- Efficient encoding and decoding for list-decoding of radius approaching $\tau_I(\delta, 0)$ [Hayashi, Yasunaga (arXiv 2018)]
  - Concatenated code with outer Reed-Solomon code
  - Also, possible for deletion only, but not for both ins. & del.
- [Haeupler, Shahrasbi, Sudan (ICALP 2018)] :
  - $\forall \tau_I > 0,\ \tau_D \in (0,1),\ \varepsilon > 0,\ \ \exists q$-ary code of rate $1 - \tau_D - \varepsilon$, $q = O(1)$ list-decodable for τ_I-frac. ins. & τ_D-frac. del.
    - Optimal with respect to rate $1 - \tau_D - \varepsilon$ and radius $(\tau_{I,}\tau_D)$
  - Efficient encoding and decoding are also presented
  - Based on *synchronization strings* [Haeupler, Shahrasbi (STOC 2017)]

    → q should be large, difficult to construct binary codes

Plotkin-type Upper Bound on Code Size

Theorem 2

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

Suppose $\exists \boldsymbol{v} \in \Sigma^N$ that is a supersequence of every $\boldsymbol{c} \in C$.

If $\frac{d}{2n} \geq 1 - \frac{n}{N}$, then $|C| \leq \frac{Nd}{Nd-2(N-n)n}$.

Proof. Apply Theorem 1 for $t_I = N - n, t_D = 0$, and the fact that $\ell := \max_{\boldsymbol{v} \in \Sigma^N} |B_L(\boldsymbol{v}, t_D, t_I) \cap C| = |C|$. QED

- A trivial supersequence for $\boldsymbol{v}$ is
  $$\boldsymbol{v}' = 12 \cdots q12 \cdots q \cdots \cdots 12 \cdots q \in [q]^{qn}$$
- But, Theorem 2 for $\boldsymbol{v}'$ can be obtained by Plotkin bound in Hamming metric and the fact that $d_H(\boldsymbol{c_i}, \boldsymbol{c_j}) \geq d_L(\boldsymbol{c_i}, \boldsymbol{c_j})/2$

➔ Theorem 2 is effective if non-trivial supsersequence $\boldsymbol{v}$ exists

Conclusion

- Our results
  - Johnson-type bound on list decodability of insertions and deletions
    - ∃ binary code of rate $\Omega(1)$ list-decodable 0.707-frac ins.
    - $\tau_I^* > 0, \tau_D^* \in [0,1)$, $\exists q$-ary code of $q = O(1)$ and rate $\Omega(1)$ list-decodable against τ_I^*-frac. Ins. and τ_D^*-frac. del.
  - Plotkin-type upper bound on code size

- Open problems
  - Efficient list-decoding for both insertions & deletions
  - Alphabet-size dependent Johnson-type bound
  - Plotkin-type bound without assuming supersequence