Edit Distance Error-Correction & Synchronization Strings

(Also on H-S-S (ICALP '18).)
Background: \[ ED(x, y) = |x| + |y| - 2 \text{LCS}(x, y) \]

\[ x \Rightarrow y \text{ can be achieved with } (|x| - |\text{LCS}(x, y)|) \]
\[ \text{deletions} + (|y| - |\text{LCS}(x, y)|) \]
\[ \text{insertions} . \]

---

Limits (for Edit Distance Error Correction)

Unique decoding (large \( q = 1 \leq n \)): \( \text{Rate} \leq 1 - (8 + 8) \).

List decoding: \( \text{Rate} \leq 1 - 8 \).

---

Contrast with Hamming Errors

\( \gamma = 8 \); locations of insertion = location of deletions.

Unique: \( \text{Rate} \geq 1 - 2\delta \) [Achieved algorithmically by RS codes, AB1 codes etc.]

List: \( \text{Rate} \Rightarrow 1 - \delta \) [Achieved by Folded RS codes, ...]
Constructions for ED codes

- 1990's: Schulman, Zuckerman: $R > 0$, $\delta > 0$, $\gamma > 0$.  
  $E = 2^{\gamma (\delta + n)}$.

- 2010's: Gururwami: $R = \max \left\{ 1 - O(\sqrt{\delta n}) \right\}$.  
  $R(1 - \delta)^{\frac{1}{2}}$

- 2017: Hanepler + Shahrasbi: Unique $R \to 1 - (\delta + \frac{\delta}{\gamma})$!

2018: H-S-S: List: $R \to 1 - S$ (no $\delta$)

- $S$ influences list size, alphabet size

General idea of [SZ 90's]:
Start with good ECC, add "index" to each letter.

- $E: \Sigma^k \to \Sigma^n \implies \hat{E}: \Sigma^k \to (\Sigma \times [n])^n$

$\hat{E}(x)_i = (E(x)_i, i)$.

Rate = $n - \log(\Sigma^n) = \frac{k \log \Sigma}{\text{log} \Sigma} - \frac{\log \Sigma}{\text{log} \Sigma}$

Need $\Sigma$ comparable with $n$ for positive rate.

$n^\alpha$ for rate $1 - \text{const.}$
Analysis of Indexing

Each insertion ⇒ erasure
Each deletion ⇒ erasure

Except insertion + deletion at same location ⇒ error

⇒ \# erasures + 2\# errors = (8+8)n.

Unique decoder corrects this if distance > \frac{(8+8)n}{\log L+E} \text{ of base code}.

⇒ Rule: \(1 - (8+8)\frac{n}{L+E} - \frac{\log n}{\log L+E}\) achievable.

\[ HS '17 \]

Observations:

1. Indexing throws away sequence of output.
2. Perhaps some idea can try to use this information.
3. Use "hash" of index, rather than entire index.

"Synchronization Shing" \( (S_1, \ldots, S_n) \in \mathcal{F} \)

with \(|\mathcal{F}| \) hopefully small.
What makes synchronization string good?


   - Verifiable? Constructible?

2. E-RSD: Verifiable! Constructible! Useful?


**Defn: E-Self-Matching. Given \( S \subseteq \mathbb{D}^n \)**

- \( M \) - a partial matching from \( \mathbb{L} \) to \( \mathbb{L} \) is **S-valid**
  - if 1. monotone: \( M = \{ (i_1, j_1), \ldots, (i_t, j_t) \} \)
    - with \( i_2 < i_{t+1} \) & \( j_2 < j_{t+1} \) \& \( t \neq t \).
  
  2. Self-repeating: \( S(i_t) = S(j_t) \)
  
  3. Non-identity: \( i_t \neq j_t \) \& \( t \).

- \( S \) is E-self-matching if every valid matching has \( |M| = t \leq E \cdot n \).
E-Self-Matchings are Useful (for list-decoding)

\[ HSS'18 \]: \& S - e-self-matching \implies \& edit-distance + \& e\textsuperscript{*} list-recoverable codes \implies \& codes.

**List-Recoverable Code**: \( C \subseteq \Sigma^n \) is \((S, \varepsilon, L)\)-list-recoverable if for every sequence of sets \( A_1, A_2, \ldots, A_n \subseteq \Sigma \) with \( |A_i| \leq \varepsilon \), the list \( \sum_{w \in C} |w \in A_i| \geq (1-\varepsilon)n \) \implies \( L \) words within \( \varepsilon \) fraction error, where

\[
\varepsilon = \frac{1}{|A_i|}.
\]

**Theorem** [Algorithmic version due to [Guruswami-Rudra]+ many others]: For \( S, L, \varepsilon, \Sigma \) s.t. \( \forall \Sigma \subseteq \Sigma \) codes of \( C \subseteq \Sigma^n \) rate \( (1-\varepsilon) \), that is \((S, \varepsilon, L)\)-list-recoverable in poly-time.
\[ E \text{-Self-Matching} + (S, e, L) - \text{Dist recoverability} \Rightarrow \text{Edit-dist. code} \]

\[ \exists \sigma \in \mathcal{E}^n \quad C \subseteq \mathcal{E}^n \quad \tilde{C} \subseteq (\mathcal{E} \times \mathcal{E})^n \]

\[ W = (W_1 \ldots W_n) \in C \quad \Rightarrow \quad \tilde{W} = (W_i, S_i)_{i=1}^n \in \tilde{C} \]

\[ \text{Rate} \geq \text{Rate}(C) - \frac{\log P}{\log 2} \quad \left[ \leq \geq 2^{\frac{2\epsilon}{\log 2}} \right] \]

\[ \text{Algorithm: Given } A \subseteq \mathcal{E}^m, \quad C \subseteq \mathcal{E}^n \]

\[ X_1, \ldots, X_m \subseteq \mathcal{E} \]

Run for \( k = 1 \) to \( \infty \) times

1. Find LCS of \( T_i \) & \( S \)
2. if \( S_i \) matched to \( T_j \), then add \( X_j \) to \( A_i \)
3. remove \( (T_j, X_j) \) pair

repeat
Analysis

- 8-fractions of words deleted - can't do anything!
- Need to prove other "error" \( \leq \varepsilon \cdot n \)
  - Error \( \leq (W_i, S_i) \) transmitted, but not \( W_i \& A_i \).
  - How can this happen?
    1. \((W_i, S_i)\) assigned to \( A_i \) for \( i' + c \).
        \[ \# \text{ such errors} \leq \varepsilon' \cdot l \cdot n \leq \frac{\varepsilon n}{2} \]
    2. \((W_i, S_i)\) left unmatched.
       - Say \( \alpha \cdot n \) symbols unmatched at end.
       - Then \( |M_i| \geq \alpha \cdot n \neq i \)

\[ \Rightarrow l \cdot \alpha \cdot n \leq (1 + \varepsilon) \cdot n \]

\[ \Rightarrow \alpha \leq \left( \frac{1 + \varepsilon}{l} \right) \cdot n \leq \frac{\varepsilon n}{2} \]

\[ \Rightarrow \text{Use } l = \frac{2(1 + \varepsilon)}{\varepsilon} \iff \varepsilon' = \frac{\varepsilon}{2l} \]
\[ E \rightarrow \text{RSD}\ S : \]

- \( S \subseteq \ell_1 \) is \( E \rightarrow \text{RSD} \) (relative suffix distance?)
  
  if \( \forall i < j < k \)

  \[ \text{LCS}(S[i, j], S[j, k]) \leq E \cdot (k - i) \]

- Obs: \( E \rightarrow \text{RSD} \) non-verifiable.

- Thm: \( \forall E \in \Gamma \forall u \in S \subseteq \ell_1 \) that is \( E \rightarrow \text{RSD} \)

- Proof: (algorithmic) Lovasz local lemma (+ conditional expectation).

\[ E \rightarrow \text{RSD}\ S \Rightarrow E \rightarrow \text{self-matching} \quad [HS'17] \]

Suppose \( i_1, i_2, \ldots, i_e \)

\( j_1, \ldots, j_e \)

is a valid matching

\[ a \leq \text{LCS}(S[i_1, i_a], S[j_1, j_{a+1}]) \leq E \cdot (j_{a+1} - i_1) \]

\( \Rightarrow \) Top intervals & bottom intervals non-overlapping

\[ E \leq E - 2n. \]