Efficient Reconstruction of Sequences

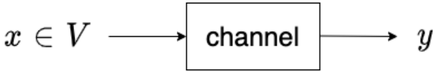
Vladimir I. Levenshtein

Presented by Ruo-Chun Tzeng

IEEE Transactions on Information Theory (2001)

Motivation

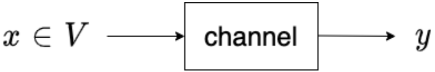
- Given a set V with minimum Hamming distance $2\tau + 1$.



- With 1 transmission of $x \in V$, $\leq \tau$ errors can be corrected.

Motivation

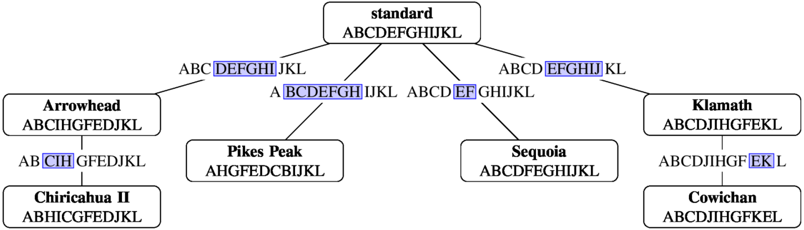
- Given a set V with minimum Hamming distance $2\tau + 1$.



- With 1 transmission of $x \in V$, $\leq \tau$ errors can be corrected.
- This paper studies the idea of correcting $> \tau$ errors by repeated transmissions of x.

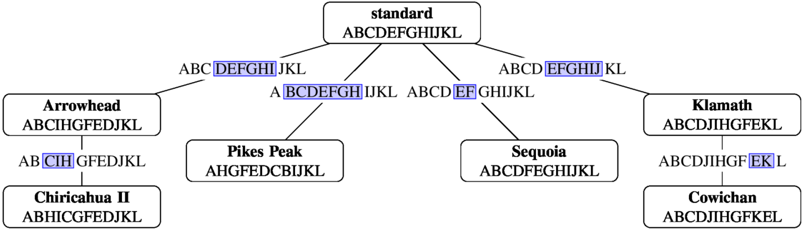
Motivation

- The problem of reconstructing the unknown x from N of its distorted sequences, $y^{(1)}, \cdots, y^{(N)}$, has many applications, e.g., DNA ancestral reconstruction:



Motivation

- The problem of reconstructing the unknown x from N of its distorted sequences, $y^{(1)}, \cdots, y^{(N)}$, has many applications, e.g., DNA ancestral reconstruction:



Question

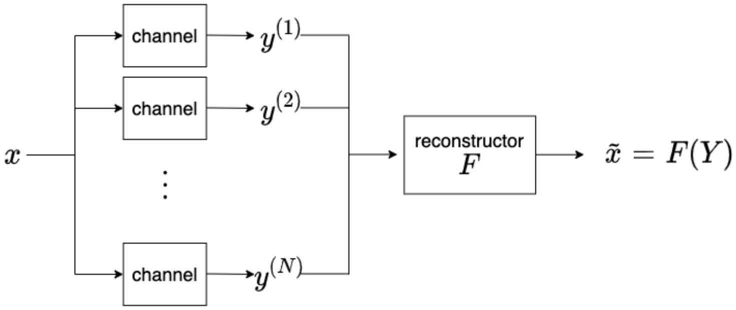
(1) What is the minimum value of N for reconstruction?

(2) How to efficiently reconstruct x from $y^{(1)}, \cdots, y^{(N)}$?

Outline

- Communication model
- Combinatorial channels ($\leq t$ errors in 1 transmission)
  - The idea for exact reconstruction
  - Recent trends
- Probabilistic channels
  - Discrete memoryless channel
  - Recent trends
- Conclusion

Communication model



- Assume $x \in V = A_q^n$ and $Y = \left[y^{(1)}, \cdots, y^{(N)}\right]$ where each $y^{(i)} \in A_r^m$.
- Measure accuracy by Hamming distance $d_H(x, F(Y)) \leq d$.

Combinatorial channel

- (n, t)-combinatorial channel: has $\leq t$ *single errors* from H in 1 transmission
  - H: the set of all single errors of the same type

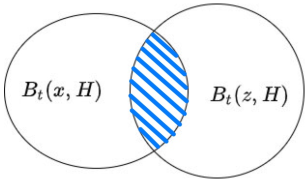
Combinatorial channel

- (n, t)-combinatorial channel: has $\leq t$ *single errors* from H in 1 transmission
- $B_t(v, H)$: the set of all words achievable from v by $\leq t$ single errors.

Combinatorial channel

- (n, t)-combinatorial channel: has $\leq t$ *single errors* from H in 1 transmission
- $B_t(v, H)$: the set of all words achievable from v by $\leq t$ single errors.
- **Idea:** $N = N_H(V, t) + 1$ for *exact reconstruction* from *distinct* $y^{(1)}, \cdots, y^{(N)}$, where

$$N_H(V, t) := \max_{v,z \in V, v \neq z} |B_t(v, H) \cap B_t(z, H)|. \tag{1}$$



Combinatorial channel

- **Idea:** $N = N_H(V,t) + 1$ for *exact reconstruction* from *distinct* $y^{(1)}, \cdots, y^{(N)}$, where

$$N_H(V,t) := \max_{v,z \in V, v \neq z} |B_t(v,H) \cap B_t(z,H)|. \quad (1)$$

- (n,t)-substitution channel:
  - $N_H(V,t) = q \sum_{i=0}^{t-1} \binom{n-1}{i} (q-1)^i$
  - $F(Y)$ exactly recovers $x_i = \text{majority}(y_i^{(1)}, \cdots, y_i^{(N)})$

Lemma $\forall a \in A_q, a \neq x_i, \left|\{v \in B_t(x,H) : v_i = a\}\right| \leq \sum_{j=0}^{t-1} \binom{n-1}{j} (q-1)^j.$

Combinatorial channel

- **Idea:** $N = N_H(V,t) + 1$ for *exact reconstruction* from *distinct* $y^{(1)}, \cdots, y^{(N)}$, where

$$N_H(V,t) := \max_{v,z \in V, v \neq z} |B_t(v,H) \cap B_t(z,H)|. \quad (1)$$

Table: Exact reconstruction results for (n,t)-combinatorial channe. All require $N = n^{\Omega(t)}$.

error-type	**case**	**reconstructor** F
substitution	all	majority
transposition	$q = 2$	thresholding
insertion	exactly t errors	[10]
deletion	exactly t errors	[10]

Combinatorial channel

- **Idea:** $N = N_H(V,t) + 1$ for *exact reconstruction* from *distinct* $y^{(1)}, \cdots, y^{(N)}$, where

$$N_H(V,t) := \max_{v,z \in V, v \neq z} |B_t(v,H) \cap B_t(z,H)|. \tag{1}$$

Table: Exact reconstruction results for (n,t)-combinatorial channe. All require $N = n^{\Omega(t)}$.

error-type	**case**	**reconstructor** F
substitution	all	majority
transposition	$q = 2$	thresholding
insertion	exactly t errors	[10]
deletion	exactly t errors	[10]

- Graph-theoretic approach [11, 15] generalizes to the problem of reconstruction within $d_H(x, F(Y)) \leq d$.

Combinatorial channel: recent trends

- Exact reconstruction for $x \in V \subseteq A_q^n$:
  - [14, 13] - insertion errors in insertion/deletion-correcting code
  - [5] - deletion errors in single-deletion code
- Practical limit on the $\#$ of repeated transmissions $\tilde{N}$:
  - [8] - list-decoding in the regime when $\tilde{N} < N$
  - [9, 4] - design of codebook V such that $N < \tilde{N}$
- Combination of different types of errors [3]
- Exact reconstruction in non-identical channels [7]

Probabilistic channel: discrete memoryless channel

- Given a DMC C with transition probability $P_C \in [0,1]^{q \times q}$ and $\delta > 0, d \geq 0$.
- Find the smallest $N = N_C(n, d, \delta)$ and a reconstructor F such that for any Y,

$$\mathbb{P}(d_H(x, F(Y)) \leq d) \geq 1 - \delta.$$

Probabilistic channel: discrete memoryless channel

- **Theorem** Let $\delta = \delta(n) > 0$ and $d = d(n) \geq 0$ be such that $\delta \to 0, d/n \to 0$ as $n \to \infty$. Then, as $n \to \infty$,

$$N_C(n, d, \delta) \to \frac{\ln \frac{n}{d+1} + \frac{\ln \delta^{-1}}{d+1}}{\ln \alpha^{-1}},$$

  where $\alpha \in (0, 1)$ depends only on P_C.

- Comparison with (n, t)-substitution channel:
  - (Combinatorial) Exact reconstruction for $t = \Theta(n)$: $N = n^{\Omega(n)}$.
  - (Probablistic) Exact reconstruction succeeds w.p. $\geq 1 - \delta$: $N = \Theta(\ln n)$.

Probabilistic channel: recent trends

- The state-of-the-art result of deletion channel (a.k.a. trace reconstruction) for $q = 2$:

	lowerbound	**upperbound**
worst-case[1]	$\tilde{\Omega}(n^{3/2})$ [2]	$e^{\tilde{\mathcal{O}}(n^{1/5})}$ [1]
average-case[2]	$\Omega(\frac{\ln^{5/2} n}{(\ln \ln n)^7})$ [2]	$e^{\mathcal{O}(\ln^{1/3} n)}$ [6]

- Circular trace reconstruction [12]

[1] Worst-case guarantee: reconstruction in high probability for any $x \in A_2^n$

[2] Average-case guarantee: reconstruction in higih probability for x drawn uniformly from A_2^n

Conclusion

- This paper initiated the study of the problem of efficient sequence reconstruction which naturally arises in many fields.
- For both combinatorial and probablistic channels, the proposed approach has inspired many future works and leaves many open problems.

Reference I

[1] Zachary Chase.
New upper bounds for trace reconstruction.
arXiv preprint arXiv:2009.03296, 2020.

[2] Zachary Chase.
New lower bounds for trace reconstruction.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Institut Henri Poincaré, 2021.

[3] Yeow Meng Chee, Han Mao Kiah, Alexander Vardy, Eitan Yaakobi, et al.
Coding for racetrack memories.
IEEE Transactions on Information Theory, 2018.

[4] Johan Chrisnata, Han Mao Kiah, and Eitan Yaakobi.
Optimal reconstruction codes for deletion channels.
In *Proc. of ISITA*. IEEE, 2020.

Reference II

[5] Ryan Gabrys and Eitan Yaakobi.
Sequence reconstruction over the deletion channel.
IEEE Transactions on Information Theory, 2018.

[6] Nina Holden, Robin Pemantle, and Yuval Peres.
Subpolynomial trace reconstruction for random strings and arbitrary deletion probability.
In *Proc. of COLT*. PMLR, 2018.

[7] Michal Horovitz and Eitan Yaakobi.
Reconstruction of sequences over non-identical channels.
IEEE Transactions on Information Theory, 2018.

[8] Ville Junnila, Tero Laihonen, and Tuomo Lehtilä.
On levenshtein's channel and list size in information retrieval.
IEEE Transactions on Information Theory, 2020.

Reference III

[9] Han Mao Kiah, Tuan Thanh Nguyen, and Eitan Yaakobi.
Coding for sequence reconstruction for single edits.
In *Proc. of ISIT*. IEEE, 2020.

[10] Vladimir I Levenshtein.
Efficient reconstruction of sequences from their subsequences or supersequences.
Journal of Combinatorial Theory, 2001.

[11] Vladimir I Levenshtein and Johannes Siemons.
Error graphs and the reconstruction of elements in groups.
Journal of Combinatorial Theory, 2009.

[12] Shyam Narayanan and Michael Ren.
Circular trace reconstruction.
In *Proc. of ITCS*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

Reference IV

[13] Frederic Sala, Ryan Gabrys, Clayton Schoeny, and Lara Dolecek.
Exact reconstruction from insertions in synchronization codes.
IEEE Transactions on Information Theory, 2017.

[14] Frederic Sala, Ryan Gabrys, Clayton Schoeny, Kayvon Mazooji, and Lara Dolecek.
Exact sequence reconstruction for insertion-correcting codes.
In *Proc. of ISIT*. IEEE, 2016.

[15] Eitan Yaakobi, Moshe Schwartz, Michael Langberg, and Jehoshua Bruck.
Sequence reconstruction for grassmann graphs and permutations.
In *Proc. of ISIT*. IEEE, 2013.