Institute for Communications Engineering
School of Computation, Information and Technology
Technical University of Munich

Sequence Reconstruction over Exact-t Adversarial Channel Repetitions: An Average Case Analysis

Vivian Papadopoulou[1,2], V. Arvind Rameshwar[3], Antonia Wachter-Zeh[1]

[1]Technical University of Munich, Institute for Communications Engineering, Coding and Cryptography Group (COD)
[2] Imperial College London, Department of Electrical Engineering, Information Processing and Communications Lab (IPCLab)
[3] India Urban Data Exchange (IUDX)

Motivation & Previous Work

In his paper[1], Levenshtein analyzed the problem of reconstructing an uncoded sequence, transmitted through an adversarial channel that introduces $\boldsymbol{t}$ distinct errors of various types (i.e., substitutions, insertions, deletions).

Through a worst-case analysis, he provides the minimum number of noisy observations required at the transmitter to guarantee successful reconstruction per error type.

For the substitution case, his formulas are a result of calculating the members of the maximal error ball intersection (worst-case).

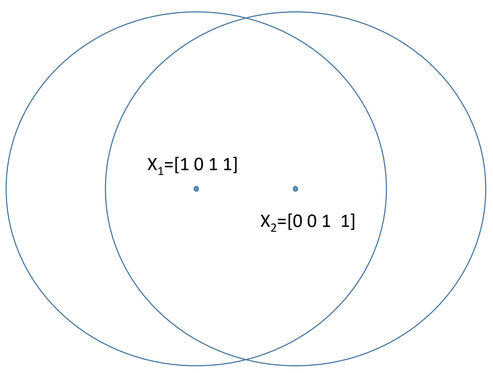

Fig.1: Levenhstein's work in a simple example

However, through his analysis, the *average* number of noisy observations required for successful reconstruction, remains unknown.

System Setup

As the average-case analysis of the sequence reconstruction problem proves to be difficult, we relax Levenshtein's constraints in the following system setup:

Assume a transmitter $\boldsymbol{T_x}$ that in interested in communicating to a receiver $\boldsymbol{R_x}$ an uncoded binary message $\boldsymbol{u}$ of length $\boldsymbol{k}$, through an adversarial channel $\boldsymbol{C}$, that introduces exactly $\boldsymbol{t}$ substitution errors (i.e., bit flips), as seen below:

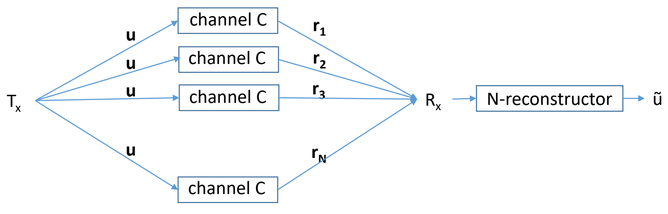

Fig. 2: System Setup

Methodology

Unlike Levenshtein, we view this problem as a (0,1) – matrix counting problem. We *assume* that the *error pattern* of each received sequence is available at $\boldsymbol{R_x}$. Therefore, $\boldsymbol{R_x}$ can construct a (0,1)-matrix of dimensions $\boldsymbol{N\,x\,k}$ with $\boldsymbol{N}$ rows containing the $\boldsymbol{N}$ error patterns corresponding to the noisy received sequences .
Assuming that the reconstructor of choice is a *majority-vote-decoder,* we further restrict our matrix as seen in the following example:

$$\boldsymbol{N}\begin{cases}\\\\\\\\\\\end{cases}\overbrace{\begin{bmatrix}1&1&0&0&0\\0&1&1&0&0\\0&0&1&1&0\\0&0&0&1&1\\1&0&0&0&1\end{bmatrix}}^{\boldsymbol{k}}$$

We would like to impose on the matrix the following constraints:

- $\boldsymbol{wt_H(row) = t}$
- $\boldsymbol{wt_H(col) \leq \left\lfloor \frac{N-1}{2} \right\rfloor}$

Simulations & Results

In general, counting the exact number of (0,1)-matrices with specific row-sum and column-sum properties is a problem found and studied in the literature[2],[3]. However, allowing these sums to vary per column (or per row, or both), makes the problem hard and the exact count remains unknown.

Through this work, we want to count (or at least estimate) the subset of the total number $\boldsymbol{M}$ of (0,1)-matrices of dimension $\boldsymbol{N\,x\,k}$ that fulfil our constraints, via sampling and recursive simulations.

The general upper bound of the targeted value to estimate, will be given by the adjusted Levenshtein formula:

$$N_{Lev} = 2\binom{k-1}{t-1}$$

The total number $\boldsymbol{M}$ of $\boldsymbol{k\,x\,N}$ matrices given some fixed $\boldsymbol{k, N, t}$ is:

$$\boldsymbol{M} = \binom{\boldsymbol{k}}{\boldsymbol{t}}^{\boldsymbol{N}} \boldsymbol{N}!$$

The lower bound of the estimation on the average number of views $\boldsymbol{N}$ required for successful reconstruction (for small $\boldsymbol{k, t}$ values):

$$\boldsymbol{E_{maj}} = \sum_{\boldsymbol{N=1}}^{\boldsymbol{N_{Lev}}} \left(\boldsymbol{1} - \frac{\widehat{\boldsymbol{G_{sim}}}}{\boldsymbol{M_{sim}}}\right)$$

When $(\boldsymbol{k, t})$ is large, and sampling can no longer provide sharp estimates, one can still estimate a lower bound on $\boldsymbol{E_{maj}}$ via the following recursive formula:

$$\boldsymbol{G_N(k,t) \geq G_N(k-3,t) + G_N\,(k-3,t-1)a_N}$$

Through this formula one can obtain a lower bound of the number of (0,1)-matrices that meet the defined constraints similarly to the sampling case.
Lastly, for the cases where $\boldsymbol{t = 1}$ and $\boldsymbol{t = k - 1}$ we prove that

$$\boldsymbol{E_{maj} = N_{Lev} = 3}$$

Future Work

- Detector-agnostic analysis / majority detector optimality
- Tighten the obtained lower bounds
- New bounds on $\boldsymbol{E_{maj}}$ (potentially via covering codes)
- Extension to other types of errors (?)
- Extension to the $\boldsymbol{q}$-ary case (?)

References

[1] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. Inf. Theory,* vol. 47, no. 1, pp. 2–22, Jan. 2001.

[2] , E. R. Canfield, B. D. McKay, "Asymptotic enumeration of dense 0-1 matrices with equal row sums and equal column sums", *The electronic journal of combinatorics*, *vol. 12*, no. R29, 2005.

[3] A. Liebenau, N, Wormald, "Asymptotic enumeration of digraphs and bipartite graphs by degree sequence", Random Structures & Algorithms, vol. *62*, pp. 259-286, 2023.