Deterministic Constructions of Edit-Distance Tree Codes for Interactive Communication

Ronak N. Desai & Hieu D. Nguyen

Department of Mathematics, Rowan University

JOINT MATHEMATICS MEETINGS
SAN DIEGO • JAN 10–13, 2018
AMERICAN MATHEMATICAL SOCIETY
MATHEMATICAL ASSOCIATION OF AMERICA

Introduction

Tree Code:

- A tree graph labeled with symbols on its edges
- At each level, two parties, Alice and Bob, alternate sending messages to each other
- We restrict the alphabet that Alice and Bob can use to be only two symbols (binary) which means we use a binary tree.
- To use an alphabet of q symbols we would need a q-ary tree.
- All possible conversation between Alice and Bob can be represented by a binary tree where a conversation is given by a path (string of symbols) starting from the root node, e.g. $s = 1364\ldots$

Error Correction:

- Suppose a third party, Eve (adversary), corrupts the conversation so that certain symbols are deleted or inserted, e.g. $\tilde{s} = 164\ldots$
- Motivating question: How to label edges in the tree so that any corrupted conversation $\tilde{s}$ can be decoded to recover the intended conversation s?
- Edit Distance: A metric that shows how different two strings of symbols are

Definition (Edit Distance). For any two strings x, y, $ED(x,y) = |x|+|y|-2\cdot LCS(x,y)$. Here $LCS(x,y)$ is the longest common substring of x and y.
Example 2. Consider two strings $x = 20120$ and $y = 0122$

- $|x|= 5$, $|y|= 4$, and $LCS(x,y) = 3$

$$ED(x,y) = 5 + 4 - 2\cdot 3 = 3$$

- 3 operations to transform one string into the other
- 20120 → 0120 → 012 → 0122
- A substitution can be thought of as an insertion immediately followed by a deletion (or vice-versa)

Edit-Distance Tree Codes (EDTC):

- EDTC is a tree code that has all of its paths sufficiently different
- Parametrized by distance parameter α that is related to edit distance

Definition (Relative Distance). If we pick four tree nodes A, B, D, E such that $B \neq D$, $B \neq E$, B is D and E's common ancestor, A is B's ancestor or B itself, the relative distance is:

$$RD(AD, BE) \equiv 2 - \frac{4\cdot LCS(AD,BE)}{|AD|+|BE|}$$

The Relative Distance is a modification of the Edit Distance that can directly be compared to the distance parameter α.

Definition (Edit-Distance Tree Code). We say that a tree code is a α-edit-distance tree code if when we consider all suitable combinations of four tree nodes A, B, D, E defined previously, the following relation holds:

$$min_{AD,BE} RD(AD, BE) > \alpha$$

The maximum value of α for which the above relation is satisfied for a given tree T is called the α-threshold of T.

Types of Deterministic Constructions

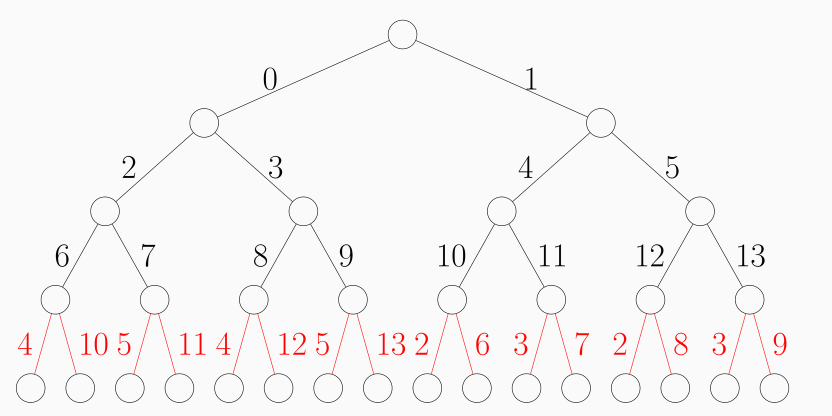
- We want an EDTC with high α and low number of different symbols
- We fix number of different symbols according to some initial depth and then extend the tree using a modification
- This modification permutes the previous row's symbols and appends onto the next row
  - Stagger (S): Place symbols from previous **two rows** in a "zig-zag" fashion
  - Inside Out (IO): Start placing symbols on next row from the symbols in previous row working inside outward
  - Greedy (G): A simplistic greedy algorithm for extending the tree that minimizes the number of occurrences of individual pairs of distinct symbols.
- **Stagger modification yielded most consistent results**

Inside Out Modification

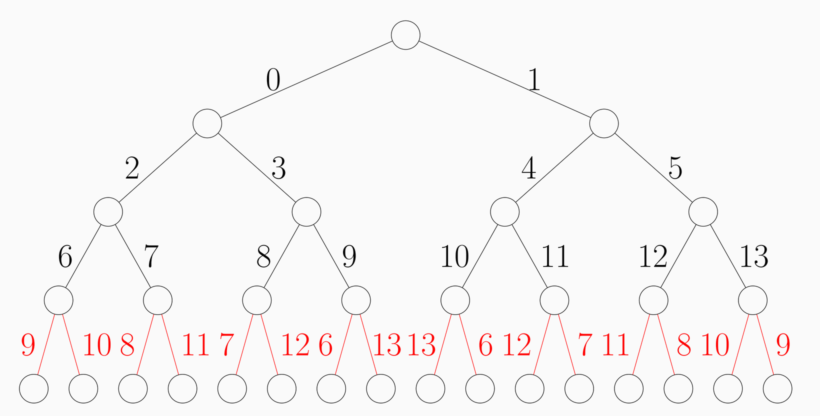
Definition (Inside Out Extension). Let $T = \{a_{j,k}\}$ be a tree code of depth d where $a_{j,k}$ denotes the edge symbol at level j and row position k and let $N = 2^d$. We extend T by one additional row $a_{d+1,k}$ defined by

$$a_{d+1,k} = \begin{cases} a_{d,\frac{N-(k-1)}{2}} & \text{if } k \text{ odd}, 1 \leq k \leq N-1 \\ a_{d,\frac{N+k}{2}} & \text{if } k \text{ even}, 2 \leq k \leq N \\ a_{d,\frac{3N-(k-1)}{2}} & \text{if } k \text{ odd}, N+1 \leq k \leq 2N-1 \\ a_{d,\frac{k-N}{2}} & \text{if } k \text{ even}, N+2 \leq k \leq 2N \end{cases}$$

Stagger

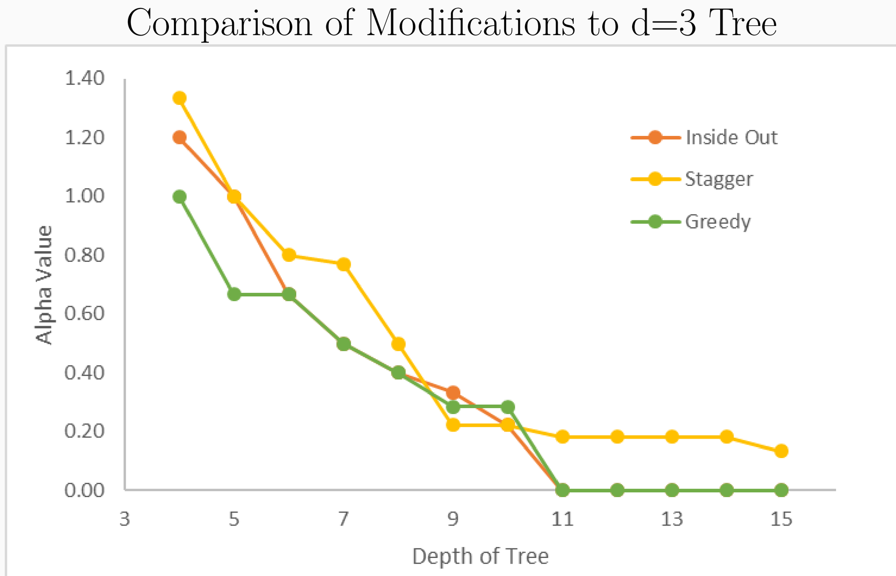


Inside Out

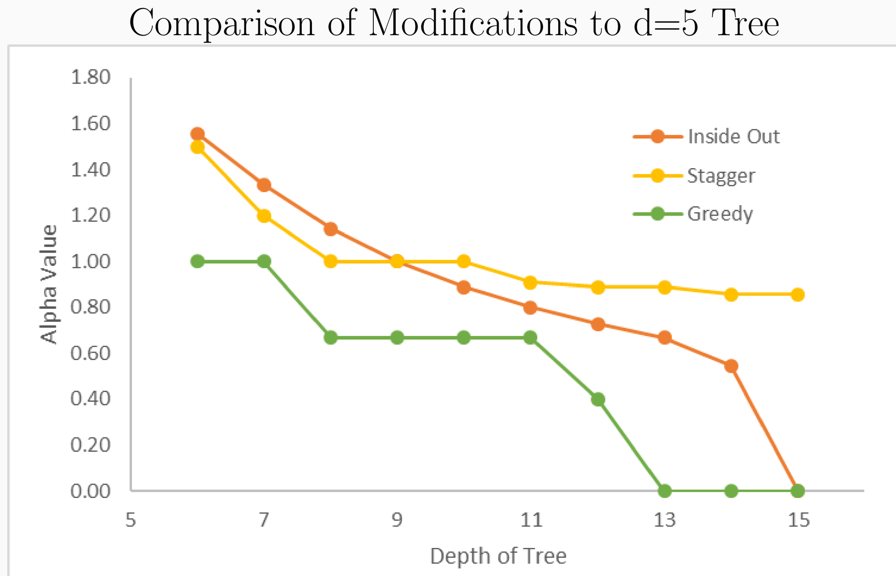


Data

Comparison of Modifications to d=3 Tree



Comparison of Modifications to d=5 Tree



Properties of Inside Out Modification

Theorem (Distinct Adjacent Symbols). Let T be a binary tree code with N distinct symbols in its d^{th} row and be extended by n additional inside out rows. Then for some row position $k \in \{x \in \mathbb{N} | x \equiv 1 \mod N\}$, the following N edges: $a_{j,k}, a_{j,k+1}, \ldots, a_{j,k+N-1}$ are pairwise distinct at any depth j.

Theorem (Distinct Parents). Let T be a binary tree code with N distinct symbols in its d^{th} row and be extended by n additional inside out rows. Then for some row position $k \in \{x \in \mathbb{N} | x \equiv 1 \mod N/2\}$, if we consider some edges $a_{j,k}, a_{j,k+1}, \ldots, a_{j,k+\frac{N}{2}-1}$, these are all distinct from their parent edges.

Conjecture (Inside Out Period) Given a sequence of symbols $a_1, a_2, \ldots, a_N$ where $N = 2^d, d \in \{x \geq 2 | x \in \mathbb{N}\}$, If we apply the Inside-Out transformation repeatedly to this sequence $d+1$ times, we will re-obtain the original sequence

Conjecture (α-threshold) Let T be a tree code with $d \geq 2$ distinct rows. Then T can be extended to a α-edit distance tree code T' with n additional rows, where

$$\alpha < \begin{cases} \dfrac{2}{1+2n} & \text{if } d = 2, 1 \leq n \leq 5 \\ \dfrac{2}{n} & \text{if } d = 3, 2 \leq n \leq 6 \\ \dfrac{2(d-1)}{n+(d-1)} & \text{if } d \geq 4, 2 \leq n \leq d+3 \end{cases}$$

Conclusions and Further Research

Results/Conclusions:

- As we increase the depth and number of times we apply a given modification, the Stagger Modification increasingly out-performs the other modifications.
- Inside Out and Stagger are both much better than the Greedy algorithm.
- Although Stagger performs better than Inside Out, its formula is much more complicated.

Further Work and Research

- Prove the Period of Inside Out and α-bounds for Stagger and Inside Out.
- Develop an efficient decoding algorithm to recover original conversation based on both Inside Out and Stagger modifications.
- Generalize Inside Out and Stagger to incorporate all available symbols in tree.

References

[1] M. Braverman, R. Gelles, J. Mao and R. Ostrovsky, "Coding for Interactive Communication Correcting Insertions and Deletions," in *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6256-6270, Oct. 2017.

[2] R. Gelles. Coding for Interactive Communication: A Survey, 2015.