# Sequence Reconstruction Over the Deletion Channel

Ryan Gabrys, *Member, IEEE*, and Eitan Yaakobi, *Senior Member, IEEE*

*Dedicated to the memory of Solomon W. Golomb (1932–2016)*

***Abstract*— The *sequence reconstruction problem*, first proposed by Levenshtein, models the setup in which a sequence from some set is transmitted over several channels, and the decoder receives the outputs from every channel. The channels are almost independent as it is only required that all outputs are different from each other. The main problem of interest is to determine the minimum number of channels required to reconstruct the transmitted sequence. In the combinatorial context, the problem is equivalent to finding the maximum intersection between two balls of radius $t$, where the distance between their centers is at least $d$. The setup of this problem was studied before for several error metrics such as the Hamming metric, the Kendall-tau metric, and the Johnson metric. In this paper, we extend the study initiated by Levenshtein for reconstructing sequences over the deletion channel. While he solved the case where the transmitted sequence can be arbitrary, we study the setup, where the transmitted sequence belongs to a single-deletion-correcting code and there are $t$ deletions in every channel. Under this paradigm, we study the minimum number of different channel outputs in order to construct a successful decoder.**

***Index Terms*— Reconstruction of sequences, deletion and insertion correcting codes.**

## I. Introduction

THE *sequence reconstruction problem* was first introduced by Levenshtein in [13] and [14]. Under this paradigm, he studied the minimum number of different (noisy) channels that are required in order to reconstruct a transmitted sequence, given that the same sequence is transmitted through every channel, each channel output is distinct, and a decoder receives all the outputs. For a sequence $\boldsymbol{x}$, the *error ball* of $\boldsymbol{x}$ is the set of possible sequences given that $\boldsymbol{x}$ is transmitted through some noisy channel. In [13], Levenshtein showed that the number of channels required to recover such a sequence has to be greater than the maximum intersection between the error balls of any two transmitted sequences. This problem was first motivated by the fields of biology and chemistry, however it is also relevant for applications in wireless sensor networks where a collection of nodes, each with partial information about the operating environment, are trying to form a *common operational picture* (*COP*). In this case, each of the received subsequences could represent the information available to that node [18].

Mathematically speaking, let $\mathcal{C}$ be a code over a space $V$ with a distance metric $\rho : V \times V \rightarrow \mathbb{N}$. Assume that its minimum distance is $d$ and there are at most $t$ errors in every channel, where $t > (d-1)/2$. Then the problem of calculating the value of

$$\max_{\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{C},\rho(\boldsymbol{x}_1,\boldsymbol{x}_2)\geqslant d} \{|B_t(\boldsymbol{x}_1) \cap B_t(\boldsymbol{x}_2)|\}, \tag{1}$$

where $B_t(\boldsymbol{x}) = \{\boldsymbol{y} \in V \mid \rho(\boldsymbol{x}, \boldsymbol{y}) \leqslant t\}$ is the ball of radius $t$ surrounding $\boldsymbol{x}$, is referred to as the *sequence reconstruction problem*.

Solving the reconstruction problem stated in (1) was studied in [13] with respect to several channels such as the Hamming distance, Johnson graphs and other metric distances. In [9]–[11], it was analyzed for permutations, and in [15], [16] for other general error graphs. The problem was studied in [24] for permutations with the Kendall's $\tau$ distance and the Grassmann graph, and in [21] for insertions. Recently, this problem was extended in the context of associative memories [23]. Under this setup, the largest intersection of multiple balls was studied, where the distance between the centers of every two of them is at least some prescribed value. In the reconstruction model, this problem is equivalent to the required number of sequences in order to output a list of some $L$ sequences which contains the transmitted sequence. This problem was also studied in [7] for the purpose of asymptotically improving the Gilbert-Varshamov bound.

Solving the reconstruction problem for the deletion channel has received a significant attention in the literature. In fact, one of the first models Levenshtein studied in [14] was for insertions and deletions along with reconstruction algorithms. The design of such algorithms for the probabilistic model of the reconstruction problem was also studied in [1], [8], and [22] and only for deletions in [5] and [6]. In In [19] and [20] an information- an information-theoretic study was carried for a special case of deletions in the context of DNA sequences.

In this work, we consider the combinatorial reconstruction problem for the deletion channel. While Levenshtein assumed in [13] and [14] that the transmitted sequences are arbitrary, we assume here that their *Levenshtein distance* is at least two and study the minimum number channel outputs needed in order to correct $t$ deletions in every channel. Here we refer to the deletion ball of a sequence $\boldsymbol{x}$ with radius $t$ to be the set of all sequences obtainable from $t$ deletions in $\boldsymbol{x}$. Our main result is showing that for $t < \frac{n}{2}$ and $n \geqslant 8$, this value is

Manuscript received May 31, 2017; revised January 4, 2018; accepted January 18, 2018. Date of publication January 31, 2018; date of current version March 15, 2018. R. Gabrys was supported by the NISE Program at SSC Pacific. E. Yaakobi was supported by the Israel Science Foundation under Grant 1624/14. This paper was presented at the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain [3].

R. Gabrys is with the Spawar Systems Center, San Diego, CA 92115 USA (e-mail: ryan.gabrys@navy.mil).

E. Yaakobi is with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: yaakobi@cs.technion.ac.il).

Communicated by P. V. Kumar, Guest Editor.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2800044

given by

$$N(n,t) \triangleq 2D(n-4,t-2)+2D(n-5,t-2) \\ +2D(n-7,t-2)+D(n-6,t-3)+D(n-7,t-3),$$

where $D(m,s)$ is the largest size of a deletion ball for sequences of length $m$ with $s$ deletions. For example, we get that $N(n,2)=6$. That is, if the transmitted sequence belongs to a single-deletion-correcting code and two deletions occurred in every channel, then 7 channels are sufficient to construct a successful decoder.

We believe the case where the transmitted sequences belong to a single-deletion-correcting code is interesting for several reasons. First, we believe this result may shed light on the more general setup where the transmitted sequence belongs to a multiple-deletion-correcting code. This is discussed later in Section V. In addition, this work contains results which pertain to computing the number of possible subsequences of a given sequence. Such results could potentially be used to derive bounds on the cardinalities of deletion-correcting codes. Lastly, we note that the Varshamov-Tenengolts code can be used to construct a single-deletion-correcting codebook and the resulting code is currently the only known code construction which is asymptotically optimal [12].

The rest of this paper is organized as follows. In Section II, we introduce our notation and establish some preliminary results. In Section III, we show that $N(n,t)$ is a lower bound; that is, we find two sequences of Levenshtein distance two such that the intersection size of their deletion balls is $N(n,t)$. In Section IV, we show that $N(n,t)$ is also an upper bound, thereby establishing equality. Section V concludes the paper.

## II. Definitions and Preliminaries

We denote by $\mathbb{F}_2$ the set $\{0,1\}$. Let $\boldsymbol{x}$ be a length-$n$ binary sequence in $\mathbb{F}_2^n$. A sequence $\boldsymbol{y} \in \mathbb{F}_2^{n-t}$ is the outcome of $t$ deletions from $\boldsymbol{x}$ if $\boldsymbol{y}$ is a subsequence of $\boldsymbol{x}$. The *deletion ball* of radius $t$ centered at $\boldsymbol{x} \in \mathbb{F}_2^n$ is defined to be

$$\mathcal{D}_t(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{F}_2^{n-t} \mid \boldsymbol{y} \text{ is a subsequence of } \boldsymbol{x}\}.$$

For two sequences $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{F}_2^n$, we say that their *Levenshtein distance* is $t$, and denote $d_L(\boldsymbol{x}_1,\boldsymbol{x}_2)=t$ if $\mathcal{D}_t(\boldsymbol{x}_1)\cap\mathcal{D}_t(\boldsymbol{x}_2)\neq\emptyset$ and $\mathcal{D}_{t-1}(\boldsymbol{x}_1)\cap\mathcal{D}_{t-1}(\boldsymbol{x}_2)=\emptyset$. Equivalently, the Levenshtein distance is one half the minimum number of insertions and deletions required to convert $\boldsymbol{x}$ to $\boldsymbol{y}$. The *minimum Levenshtein distance* of a code $\mathcal{C} \subseteq \mathbb{F}_2^n$ is the minimum Levenshtein distance between any two different sequences in $\mathcal{C}$. The code is called a *$t$-deletion-correcting code* if its minimum Levenshtein distance is at least $t+1$. For notational convenience, we denote $\mathcal{D}_t(\boldsymbol{x})=\emptyset$ when $t<0$ and thus $|\mathcal{D}_t(\boldsymbol{x})|=0$. A *run* of $\boldsymbol{x}$ is a maximal interval which consists of the same symbol.

Let $\boldsymbol{a}_n \in \mathbb{F}_2^n$ be the alternating sequence where its first bit is 1. For $0<t<n$, we denote by $D(n,t)$ the maximum size of a deletion ball of radius $t$, i.e.,

$$D(n,t) = \max_{\boldsymbol{x}\in\mathbb{F}_2^n}\{|\mathcal{D}_t(\boldsymbol{x})|\}.$$

It is known from [2] that

$$D(n,t) = |\mathcal{D}_t(\boldsymbol{a}_n)| = \sum_{i=0}^{t}\binom{n-t}{i}, \tag{2}$$

and from [2] the following recursion holds

$$D(n,t) = D(n-1,t)+D(n-2,t-1). \tag{3}$$

We assume here and afterwards that $D(n,t)=0$ if $t>n$, $n<0$, or $t<0$. The main goal of this work is to study the combinatorial value

$$N(n,t_1,t_2) = \max_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{F}_2^n, d_L(\boldsymbol{x},\boldsymbol{y})\geqslant t_1}\{|\mathcal{D}_{t_2}(\boldsymbol{x})\cap\mathcal{D}_{t_2}(\boldsymbol{y})|\}, \tag{4}$$

where $0<t_1<t_2<n$. In [14], Levenshtein studied the case $t_1=1$ (i.e. $\boldsymbol{x}$ and $\boldsymbol{y}$ only need to be different from each other) and showed that for $1<t_2\leqslant n-1$

$$N(n,1,t_2) = 2D(n-2,t_2-1). \tag{5}$$

This value is achieved for example as the intersection of the balls centered at the two sequences

$$\boldsymbol{x} = (0,1,\boldsymbol{a}_{n-2}), \quad \boldsymbol{y} = (1,0,\boldsymbol{a}_{n-2}).$$

In this paper we will focus on the combinatorial problem stated in (4) for $t_1=2$. Our main result in the paper is showing that for $t_1=2$ and $n\geqslant 8$,

$$N(n,2,t) = N(n,t),$$

where $N(n,t)$ is given by

$$N(n,t) = 2D(n-4,t-2)+2D(n-5,t-2) \\ +2D(n-7,t-2) \\ +D(n-6,t-3)+D(n-7,t-3), \tag{6}$$

when $t<n/2$ and $N(n,t)=2^{n-t}$ when $t\geqslant n/2$. In fact, it is not hard to verify that if $t\geqslant n/2$ then $N(n,t)=2^{n-t}$ since one can consider the intersection of the sequences $\boldsymbol{x}=1001\boldsymbol{a}_{n-4}$ and $\boldsymbol{y}=0110\boldsymbol{a}_{n-4}$, which are of Levenshtein distance two from each other since $\mathcal{D}_2(0110)=\mathcal{D}_2(1001)=\{00,01,10,11\}$. Furthermore, for $t\geqslant\frac{n}{2}$, $\mathcal{D}_t(\boldsymbol{x})=\mathcal{D}_t(\boldsymbol{y})=\{0,1\}^{n-t}$ since both of these sequences are concatenations of pairs of 01 or 10. We used a computer search to determine the values of $N(n,2,t)$ for $n\leqslant 9$, and the relation $N(n,2,t)=N(n,t)$ holds for $n=8$ and $n=9$. Thus, unless stated otherwise, we assume in this paper that the values of $n$ and $t$ satisfy $n\geqslant 10$ and $t<\frac{n}{2}$.

Let $\mathcal{X}\subseteq\mathbb{F}_2^n$ be a set and $\boldsymbol{v}$ a sequence of length at most $n$. We denote by $\mathcal{X}^{\boldsymbol{v}}$ the set of all sequences in $\mathcal{X}$ that start with the sequence $\boldsymbol{v}$, that is,

$$\mathcal{X}^{\boldsymbol{v}} = \{\boldsymbol{x}\in\mathcal{X} \mid \boldsymbol{v} \text{ is a prefix of } \boldsymbol{x}\}.$$

Similarly, $\mathcal{X}_{\boldsymbol{v}}$ is the set of all sequences in $\mathcal{X}$ that end with the sequence $\boldsymbol{v}$. We can apply these notations simultaneously so for two sequences $\boldsymbol{v}_1,\boldsymbol{v}_2$ each of length at most $n$, $\mathcal{X}_{\boldsymbol{v}_2}^{\boldsymbol{v}_1}$ is the set of all sequences starting with $\boldsymbol{v}_1$ and ending with $\boldsymbol{v}_2$. For a sequence $\boldsymbol{v}\in\mathbb{F}_2^m$ and a set $\mathcal{X}\subseteq\mathbb{F}_2^n$, the set $\boldsymbol{v}\circ\mathcal{X}$ is the result of prepending the sequence $\boldsymbol{v}$ before every sequence in $\mathcal{X}$,

$$\boldsymbol{v}\circ\mathcal{X} = \{(v_1,\ldots,v_m,x_1,\ldots,x_n) \mid (x_1,\ldots,x_n)\in\mathcal{X}\}.$$

Similarly, the set $\mathcal{X}\circ\boldsymbol{v}$ consists of the concatenation of the sequence $\boldsymbol{v}$ at the end of every sequence in $\mathcal{X}$.

We finish this section with the following two lemmas which follow from the same ideas as in [4] and [17], and therefore whose proofs have been omitted. These lemmas will be used repeatedly in the next two sections. The first one claims that finding the cardinality of a set can be done by splitting it into mutually disjoint sets according to the prefixes and suffixes of its sequences.

*Lemma 1: Let $n, m_1, m_2$, be positive integers such that $m_1 + m_2 \leqslant n$, and $\mathcal{X} \subseteq \mathbb{F}_2^n$. Then,*

$$|\mathcal{X}| = \sum_{\boldsymbol{x}_1 \in \mathbb{F}_2^{m_1}} \sum_{\boldsymbol{x}_2 \in \mathbb{F}_2^{m_2}} \left|\mathcal{X}_{\boldsymbol{x}_2}^{\boldsymbol{x}_1}\right|.$$

As an immediate result of Lemma 1, we conclude that for every $\boldsymbol{x} \in \mathbb{F}_2^n$ and two positive integers $m_1, m_2$ such that $m_1 + m_2 \leqslant n - t$, the following equality holds

$$|\mathcal{D}_t(\boldsymbol{x})| = \sum_{\boldsymbol{x}_1 \in \mathbb{F}_2^{m_1}} \sum_{\boldsymbol{x}_2 \in \mathbb{F}_2^{m_2}} \left|\mathcal{D}_t(\boldsymbol{x})_{\boldsymbol{x}_2}^{\boldsymbol{x}_1}\right|. \tag{7}$$

The second lemma claims that finding all sequences in a deletion ball $\mathcal{D}_t(\boldsymbol{x})$ which start with $\boldsymbol{x}_1$ and end $\boldsymbol{x}_2$ can be done by first finding the smallest prefix, suffix that contains $\boldsymbol{x}_1, \boldsymbol{x}_2$ as a subsequence, respectively, and then calculating the deletion ball in the remainder of the sequence $\boldsymbol{x}$.

*Lemma 2: Let $n, m_1, m_2, t$ be positive integers, and $\boldsymbol{x} \in \mathbb{F}_2^n, \boldsymbol{x}_1 \in \mathbb{F}_2^{m_1}, \boldsymbol{x}_2 \in \mathbb{F}_2^{m_2}$. Assume that $k_1$ is the smallest integer such that $\boldsymbol{x}_1$ is a subsequence of $(x_1, \ldots, x_{k_1})$ and $k_2$ is the largest integer where $\boldsymbol{x}_2$ is a subsequence of $(x_{k_2}, \ldots, x_n)$. If $k_1 < k_2$ then*

$$\mathcal{D}_t(\boldsymbol{x})_{\boldsymbol{x}_2}^{\boldsymbol{x}_1} = \boldsymbol{x}_1 \circ \mathcal{D}_{t^*}(x_{k_1+1}, \ldots, x_{k_2-1}) \circ \boldsymbol{x}_2,$$

*where $t^* = t - (k_1 - m_1) - (n - k_2 + 1 - m_2)$. In particular,*

$$\left|\mathcal{D}_t(\boldsymbol{x})_{\boldsymbol{x}_2}^{\boldsymbol{x}_1}\right| = \left|\mathcal{D}_{t^*}(x_{k_1+1}, \ldots, x_{k_2-1})\right|.$$

Lastly, we state the following claim, whose proof is straightforward. For a binary sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n-1}, x_n)$, let $R(\boldsymbol{x}) = (x_n, x_{n-1}, \ldots, x_2, x_1)$, and $\bar{\boldsymbol{x}} = (1 - x_1, \ldots, 1 - x_n)$.

*Claim 1: Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$. Then,*

$$\begin{aligned} |\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| &= |\mathcal{D}_t(R(\boldsymbol{x})) \cap \mathcal{D}_t(R(\boldsymbol{y}))| \\ &= |\mathcal{D}_t(\bar{\boldsymbol{x}}) \cap \mathcal{D}_t(\bar{\boldsymbol{y}})|. \end{aligned}$$

The following example illustrates our notations.

*Example 1: Let $\mathcal{X} \subseteq \mathbb{F}_2^4$ be the following set*

$$\begin{aligned} \mathcal{X} = \{&(0,0,0,1), (0,0,1,1), (0,1,0,0), (0,1,0,1), \\ &(0,1,1,0), (1,0,0,0), (1,0,1,1), (1,1,0,1)\}. \end{aligned}$$

*Then, $\mathcal{X}_0^0 = \{(0,1,0,0), (0,1,1,0)\}$, $\mathcal{X}_1^0 = \{(0,0,0,1), (0,0,1,1), (0,1,0,1)\}$, $\mathcal{X}_0^1 = \{(1,0,0,0)\}$, $\mathcal{X}_1^1 = \{(1,0,1,1), (1,1,0,1)\}$, and note that $|\mathcal{X}| = |\mathcal{X}_0^0| + |\mathcal{X}_1^0| + |\mathcal{X}_0^1| + |\mathcal{X}_1^1|$. Furthermore*

$$\begin{aligned} &(0,1) \circ \mathcal{X} \\ &\quad = \{(0,1,0,0,0,1), (0,1,0,0,1,1), (0,1,0,1,0,0), \\ &\qquad (0,1,0,1,0,1), (0,1,0,1,1,0), (0,1,1,0,0,0), \\ &\qquad (0,1,1,0,1,1), (0,1,1,1,0,1)\}. \end{aligned}$$

*If $\boldsymbol{x} = (0,1,1,0,0,1)$ then $\mathcal{D}_3(\boldsymbol{x})_0^1 = 1 \circ \mathcal{D}_1((1,0)) \circ 0 = 1 \circ \{0,1\} \circ 0$, and $|\mathcal{D}_3(\boldsymbol{x})_0^1| = 2$.*

*For $\boldsymbol{x} = (0,1,1,0)$, $\boldsymbol{y} = (0,1,0,1)$, we have that*

$$\begin{aligned} R(\boldsymbol{x}) &= (0,1,1,0), \ R(\boldsymbol{y}) = (1,0,1,0), \\ \bar{\boldsymbol{x}} &= (1,0,0,1), \ \bar{\boldsymbol{y}} = (1,0,1,0), \end{aligned}$$

*and*

$$\begin{aligned} \mathcal{D}_1(\boldsymbol{x}) \cap \mathcal{D}_1(\boldsymbol{y}) &= \{(0,1,0), (0,1,1)\} \\ \mathcal{D}_1(R(\boldsymbol{x})) \cap \mathcal{D}_1(R(\boldsymbol{y})) &= \{(0,1,0), (1,1,0)\} \\ \mathcal{D}_1(\bar{\boldsymbol{x}}) \cap \mathcal{D}_1(\bar{\boldsymbol{y}}) &= \{(1,0,1), (1,0,0)\}. \end{aligned}$$

*In particular, $|\mathcal{D}_1(\boldsymbol{x}) \cap \mathcal{D}_1(\boldsymbol{y})| = |\mathcal{D}_1(R(\boldsymbol{x})) \cap \mathcal{D}_1(R(\boldsymbol{y}))| = |\mathcal{D}_1(\bar{\boldsymbol{x}}) \cap \mathcal{D}_1(\bar{\boldsymbol{y}})|$.*

## III. The Lower Bound

The main goal of this section is to establish the value of $N(n,t)$ as a lower bound on $N(n,2,t)$. We accomplish this task by showing in Theorem 3 that the sequences $\boldsymbol{x} = (1,0,\boldsymbol{a}_{n-4},0,1)$, $\boldsymbol{y} = (0,1,\boldsymbol{a}_{n-4},1,0)$ of Levenshtein distance two satisfy

$$|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| = N(n,t).$$

We first begin with the following simple, yet useful, claim which is a restatement of a result from [17].

*Claim 2: (c.f. [17], Lemma II.5) Let $\ell < n$, $\boldsymbol{x} \in \mathbb{F}_2^n$, $\boldsymbol{y} \in \mathbb{F}_2^{n-\ell}$, where $\boldsymbol{y} \in \mathcal{D}_\ell(\boldsymbol{x})$ and $\ell < t$. Then, $\mathcal{D}_{t-\ell}(\boldsymbol{y}) \subseteq \mathcal{D}_t(\boldsymbol{x})$.*

The following three claims are the results we need in order to prove Theorem 3. The next claim is a result of counting all sequences in the set $\mathcal{D}_t(\boldsymbol{a}_n) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_n)$ that start with a 1 symbol and all sequences in the set $\mathcal{D}_t(\boldsymbol{a}_n) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_n)$ that start with the 0 symbol.

*Claim 3: For all $n$ and $t$,*

$$|\mathcal{D}_t(\boldsymbol{a}_n) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_n)| = 2D(n-2, t-1). \tag{8}$$

The next claim follows by considering all sequences in the set $\mathcal{D}_t(\boldsymbol{a}_{n-1}, 0) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_{n-1}, 0)$ that end with a 0 symbol and all sequences in the set $\mathcal{D}_t(\boldsymbol{a}_{n-1}, 0) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_{n-1}, 0)$ that end with a 1 symbol.

*Claim 4: For all $n$ and $t$,*

$$\begin{aligned} &|\mathcal{D}_t(\boldsymbol{a}_{n-1}, 0) \cap \mathcal{D}_t(\bar{\boldsymbol{a}}_{n-1}, 0)| \\ &\qquad = 2D(n-3, t-1) + D(n-3, t-2). \end{aligned}$$

The next claim can be proven by first considering all sequences in the intersection that start with the 1 symbol and applying Claim 4, and then considering all sequences in the intersection that start with the 0 symbol and applying Claim 2.

*Claim 5: For even $n$ and $t$,*

$$\begin{aligned} |\mathcal{D}_t(\boldsymbol{a}_n, 0) \cap \mathcal{D}_t(1, \boldsymbol{a}_n)| &= 2D(n-3, t-1) \\ &\quad + D(n-3, t-2) + D(n-2, t-2). \end{aligned}$$

The next theorem constitutes the lower bound.

*Theorem 3: For all $n$ and $t$, $N(n,2,t) \geqslant N(n,t)$.*

*Proof:* We show here the proof for even values of $n$, while the proof for odd values follows from similar ideas. Let

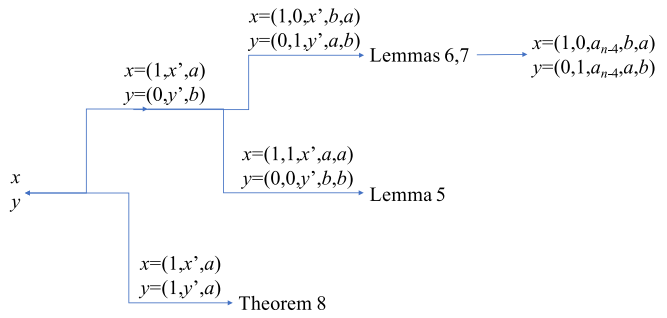


Fig. 1. Outline of proof of upper bound.

$\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, 0, 1)$, $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, 1, 0)$, and we denote $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$. We first consider the set $\mathcal{X}_0^0$.

$$\begin{aligned} |\mathcal{X}_0^0| &\overset{(a)}{=} |0 \circ (\mathcal{D}_{t-2}(\boldsymbol{a}_{n-4}) \cap \mathcal{D}_t(1, \boldsymbol{a}_{n-4}, 1)) \circ 0| \\ &= |\mathcal{D}_{t-2}(\boldsymbol{a}_{n-4}) \cap \mathcal{D}_t(1, \boldsymbol{a}_{n-4}, 1)| \\ &\overset{(b)}{=} |\mathcal{D}_{t-2}(\boldsymbol{a}_{n-4})| = D(n-4, t-2), \end{aligned} \tag{9}$$

where (a) follows from Lemma 2 and (b) follows by applying Claim 2 since $\boldsymbol{a}_{n-4} \in \mathcal{D}_2(1, \boldsymbol{a}_{n-4}, 1)$. By repeating the same steps of $\mathcal{X}_0^0$, we get that

$$|\mathcal{X}_1^1| = D(n-4, t-2). \tag{10}$$

For the set $\mathcal{X}_0^1$, we have from Lemma 2

$$\begin{aligned} |\mathcal{X}_0^1| &= |1 \circ (\mathcal{D}_{t-1}(\bar{\boldsymbol{a}}_{n-3}) \cap \mathcal{D}_{t-1}(\boldsymbol{a}_{n-3})) \circ 0| \\ &= |\mathcal{D}_{t-1}(\bar{\boldsymbol{a}}_{n-3}) \cap \mathcal{D}_{t-1}(\boldsymbol{a}_{n-3})| \overset{(a)}{=} 2D(n-5, t-2), \end{aligned} \tag{11}$$

where (a) follows from Claim 3. Similarly, for the set $\mathcal{X}_1^0$, we have

$$\begin{aligned} |\mathcal{X}_1^0| &= |0 \circ (\mathcal{D}_{t-1}(\boldsymbol{a}_{n-4}, 0) \cap \mathcal{D}_{t-1}(1, \boldsymbol{a}_{n-4})) \circ 1| \\ &= |\mathcal{D}_{t-1}(\boldsymbol{a}_{n-4}, 0) \cap \mathcal{D}_{t-1}(1, \boldsymbol{a}_{n-4})| \\ &\overset{(a)}{=} 2D(n-7, t-2) + D(n-6, t-3) \\ &\quad + D(n-7, t-3), \end{aligned} \tag{12}$$

where (a) is a result of Claim 5. Finally, by applying Lemma 1 and summing (9), (10), (11), and (12) we get the result in the statement of the theorem. ∎

## IV. The Upper Bound

We now prove that the lower bound from the previous section is also an upper bound. That is, we show that for any pair of sequences, $\boldsymbol{x}, \boldsymbol{y}$ where $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$, the intersection of their deletion balls is at most $N(n, t)$. We proceed by considering different possibilities for $\boldsymbol{x}, \boldsymbol{y}$ and showing that this result holds in all cases.

At a high level, the logical flow of the section is illustrated in Figure 1. We first assume $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$ and that $\boldsymbol{x}, \boldsymbol{y}$ differ on the first and last bits.

1) Lemma 5 shows that if $\boldsymbol{x}, \boldsymbol{y}$ have more than $N(n, t)$ common subsequences, then $\boldsymbol{x}$ cannot have the same value in positions one and two.
2) As a consequence of Lemma 5, if $\boldsymbol{x}, \boldsymbol{y}$ have more than $N(n, t)$ common subsequences then the sequences can be written as $\boldsymbol{x} = (1, 0, \boldsymbol{x}', \bar{a}, a) \in \mathbb{F}_2^n$, $\boldsymbol{y} = (0, 1, \boldsymbol{y}', a, \bar{a}) \in \mathbb{F}_2^n$ for some $a \in \mathbb{F}_2$.
3) Lemma 7 then shows that the intersection of the deletion balls for $\boldsymbol{x}, \boldsymbol{y}$ (as described in the previous sentence) is at most $N(n, t)$ and that we can write $\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, \bar{a}, a)$ and $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, a, \bar{a})$.
4) Finally, in Theorem 8, we remove the restriction that $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the first bit, and show that when $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$ for any $\boldsymbol{x}, \boldsymbol{y}$, the intersection of their deletion balls is at most $N(n, t)$.

We first introduce a few identities that will be used throughout the section. From [14], we have

$$D(n, t) \leqslant D(n+1, t+1), \tag{13}$$

and

$$D(n, t) \leqslant D(n+1, t). \tag{14}$$

The following corollary follows from (3) and (6).

*Corollary 4: For all n and t, we have*

$$N(n, t) = N(n-1, t) + N(n-2, t-1). \tag{15}$$

We begin by restricting our attention to sequences $\boldsymbol{x}, \boldsymbol{y}$ that differ on the first bit and the last bit.

*Lemma 5: For all n and t, let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n$ be such that $x_1 = 1$, $y_1 = 0$ and*

*1) $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$,*
*2) $x_n \neq y_n$, and*
*3) $x_1 = x_2$.*

*Then, $|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant N(n, t)$.*

*Proof:* Assume first that $x_n = 0$, so $y_n = 1$ and we can write

$$\boldsymbol{x} = (1, 1, \boldsymbol{x}', 0), \quad \boldsymbol{y} = (0, \boldsymbol{y}', 1),$$

where $\boldsymbol{x}' = (x_1', \ldots, x_{n-3}') \in \mathbb{F}_2^{n-3}$, $\boldsymbol{y}' = (y_1', \ldots, y_{n-2}') \in \mathbb{F}_2^{n-2}$. Let $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$. We first consider the set $|\mathcal{X}^0|$, where we have

$$|\mathcal{X}^0| \leqslant |\mathcal{D}_t(\boldsymbol{x})^0| \leqslant |\mathcal{D}_{t-2}(\boldsymbol{x}', 0)^0| \leqslant D(n-3, t-2).$$

Similarly,

$$|\mathcal{X}_0^1| \leqslant |\mathcal{D}_t(\boldsymbol{y})_0^1| \leqslant |\mathcal{D}_{t-2}(\boldsymbol{y}')_0^1| \leqslant D(n-4, t-2).$$

As for the set $|\mathcal{X}_1^1|$, assume first that $x_{n-3}' = 1$ and $y_1' = 1$. Then we get

$$\begin{aligned} |\mathcal{X}_1^1| &= \left|1 \circ \left(\mathcal{D}_{t-1}(1, x_1', \ldots, x_{n-4}') \cap \mathcal{D}_{t-1}(y_2', \ldots, y_{n-2}')\right) \circ 1\right| \\ &= \left|\left(\mathcal{D}_{t-1}(1, x_1', \ldots, x_{n-4}') \cap \mathcal{D}_{t-1}(y_2', \ldots, y_{n-2}')\right)\right| \\ &\leqslant 2D(n-5, t-2), \end{aligned}$$

where the last inequality holds since $(1, x_1', \ldots, x_{n-4}') \neq (y_2', \ldots, y_{n-2}')$ as otherwise we won't have $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$. To complete this part assume that $x_{n-3}' = 0$, then we get

$$\begin{aligned} |\mathcal{X}_1^1| &\leqslant \left|\mathcal{D}_{t-1}(1, \boldsymbol{x}')_1\right| \leqslant \left|\mathcal{D}_{t-2}(1, x_1', \ldots, x_{n-4}')_1\right| \\ &\leqslant D(n-4, t-2). \end{aligned}$$

Using (3) and (6), it can be shown that $D(n-3,t-2)+2D(n-4,t-2) \leqslant N(n,t)$ for the case where $x'_{n-3}=0$. For the case where $x'_{n-3}=1$, we show at the end of the proof that $D(n-3,t-2)+D(n-4,t-2)+2D(n-5,t-2) \leqslant N(n,t)$ so the result holds in either case.

We now consider the case $x_n=1$, so $y_n=0$ and

$$\boldsymbol{x}=(1,1,\boldsymbol{x}',1),$$
$$\boldsymbol{y}=(0,\boldsymbol{y}',0),$$

and let $\mathcal{X}=\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})$. As before, we get

$$|\mathcal{X}^0| \leqslant D(n-3,t-2), |\mathcal{X}_1^1| \leqslant D(n-4,t-2).$$

and $|\mathcal{X}_0^1| \leqslant 2D(n-5,t-2)$.

Finally, we conclude that in both cases

$$\begin{aligned}|\mathcal{X}| &\leqslant D(n-3,t-2)+D(n-4,t-2)+2D(n-5,t-2)\\ &\leqslant N(n,t),\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}&N(n,t)-(D(n-3,t-2)\\ &\quad+D(n-4,t-2)+2D(n-5,t-2))\\ &=N(n,t)-(2D(n-4,t-2)+D(n-5,t-3)\\ &\quad+2D(n-5,t-2))\\ &=2D(n-7,t-2)+D(n-6,t-3)+D(n-7,t-3)\\ &\quad-(D(n-6,t-3)+D(n-7,t-4))\\ &=2D(n-7,t-2)+D(n-7,t-3)\\ &\quad-D(n-7,t-4)\geqslant 0.\end{aligned}$$

To prove this part we used the relation on $D(n,t)$ from (3) and the expression for $N(n,t)$ from (6). ∎

As a result of Lemma 5, if $\boldsymbol{x},\boldsymbol{y}\in\mathbb{F}_2^n$ are such that $d_L(\boldsymbol{x},\boldsymbol{y})\geqslant 2$, $x_1=1$, $\boldsymbol{x},\boldsymbol{y}$ disagree in the first and last bits, and $|\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})|>N(n,t)$, then $\boldsymbol{x}=(1,0,\boldsymbol{x}',1,0)$, $\boldsymbol{y}=(0,1,\boldsymbol{y}',0,1)$ or $\boldsymbol{x}=(1,0,\boldsymbol{x}',0,1)$, $\boldsymbol{y}=(0,1,\boldsymbol{y}',1,0)$ where $\boldsymbol{x}',\boldsymbol{y}'\in\mathbb{F}_2^{n-4}$. The purpose of the next lemma is to show that in either case $\boldsymbol{x}',\boldsymbol{y}'\in\{\boldsymbol{a}_{n-4},\bar{\boldsymbol{a}}_{n-4}\}$. In order to establish this result, we first state the following three useful claims. The proof of Claim 6 appears in Appendix A. Appendices B and C contain proof sketches for Claims 7 and 8, respectively.

*Claim 6: For all $n$ and $t$, suppose the number of runs in $\boldsymbol{x}\in\mathbb{F}_2^n$ is at most $n-1$. Then,*

$$|\mathcal{D}_t(\boldsymbol{x})| \leqslant D(n-2,t)+D(n-2,t-1)+D(n-4,t-2).$$

*Claim 7: For all $n$ and $t$, suppose $\boldsymbol{x}\in\mathbb{F}_2^n$, $\boldsymbol{y}\in\mathbb{F}_2^n$ are such that $\boldsymbol{x}\neq\boldsymbol{y}$, the number of runs in $\boldsymbol{x}=(x_1,\ldots,x_n)$ is $n-1$, $\boldsymbol{y}$ has at least $n-1$ runs, and $x_2\neq x_3, x_{n-1}\neq x_{n-2}$. Then,*

$$\begin{aligned}|\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})| &\leqslant D(n-2,t-1)+D(n-4,t-1)\\ &\quad+D(n-4,t-2)+D(n-6,t-3).\end{aligned}$$

*Claim 8: For all $n$ and $t$, suppose $\boldsymbol{x}\in\mathbb{F}_2^n$, $\boldsymbol{y}\in\mathbb{F}_2^n$ are such that $\boldsymbol{x}\neq\boldsymbol{y}$ and the number of runs in $\boldsymbol{x}$ is at most $n-2$. Then,*

$$\begin{aligned}|\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})| &\leqslant D(n-2,t-1)+D(n-4,t-1)\\ &\quad+D(n-4,t-2)+D(n-6,t-3).\end{aligned}$$

Using Claims 6, 7, and 8, we prove the following lemma.

*Lemma 6: For all $n$ and $t$, let $a\in\mathbb{F}_2$. Let $\boldsymbol{x}=(1,0,\boldsymbol{x}',\bar{a},a)\in\mathbb{F}_2^n$, $\boldsymbol{y}=(0,1,\boldsymbol{y}',a,\bar{a})\in\mathbb{F}_2^n$ be such that $d_L(\boldsymbol{x},\boldsymbol{y})\geqslant 2$. If*

$$|\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})| \geqslant N(n,t),$$

*then $\boldsymbol{x}',\boldsymbol{y}'\in\{\boldsymbol{a}_{n-4},\bar{\boldsymbol{a}}_{n-4}\}$.*

*Proof:* Similar to before, let $\mathcal{X}=\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})$. Assume in the contrary that $\boldsymbol{x}'\notin\{\boldsymbol{a}_{n-4},\bar{\boldsymbol{a}}_{n-4}\}$, then we can apply Claim 6 to get

$$\begin{aligned}|\mathcal{X}_{\bar{a}}^0| &= |0\circ\left(\mathcal{D}_{t-2}(\boldsymbol{x}')\cap\mathcal{D}_t(1,\boldsymbol{y}',a)\right)\circ\bar{a}|\\ &\leqslant |\mathcal{D}_{t-2}(\boldsymbol{x}')| \leqslant D(n-6,t-2)\\ &\quad+D(n-6,t-3)+D(n-8,t-4).\end{aligned}$$

We also have

$$|\mathcal{X}_a^1| \leqslant |1\circ\mathcal{D}_{t-2}(\boldsymbol{y}')\circ a| \leqslant D(n-4,t-2).$$

Furthermore, notice

$$|\mathcal{X}_{\bar{a}}^1| = |1\circ\left(\mathcal{D}_{t-1}(0,\boldsymbol{x}')\cap\mathcal{D}_{t-1}(\boldsymbol{y}',a)\right)\circ\bar{a}|,$$

where $(\boldsymbol{y}',a)\neq(0,\boldsymbol{x}')$, since otherwise we will get that $d_L(\boldsymbol{x},\boldsymbol{y})=1$. Therefore, we can apply (5) to get that

$$|\mathcal{X}_{\bar{a}}^1| \leqslant 2D(n-5,t-2).$$

Next we consider $|\mathcal{X}_a^0|$ where

$$|\mathcal{X}_a^0| = |0\circ\left(\mathcal{D}_{t-1}(\boldsymbol{x}',\bar{a})\cap\mathcal{D}_{t-1}(1,\boldsymbol{y}')\right)\circ a|$$

If $(\boldsymbol{x}',\bar{a}),(1,\boldsymbol{y}')$ satisfy the conditions of Claim 7 or Claim 8, then we can write

$$\begin{aligned}|\mathcal{X}_a^0| &\leqslant D(n-5,t-2)+D(n-7,t-2)\\ &\quad+D(n-7,t-3)+D(n-9,t-4).\end{aligned}$$

We now consider the case where $(\boldsymbol{x}',\bar{a}),(1,\boldsymbol{y}')$ do not satisfy the conditions in Claim 7 and Claim 8. If $(\boldsymbol{x}',\bar{a})$ does not satisfy the conditions for $\boldsymbol{x}$ in Claim 8, then $(\boldsymbol{x}',\bar{a})$ has at least $(n-3)-1=n-4$ runs and, in particular, $\boldsymbol{x}'\in\mathbb{F}_2^{n-4}$ has $(n-4)-1=n-5$ runs by assumption since $\boldsymbol{x}'\notin\{\boldsymbol{a}_{n-4},\bar{\boldsymbol{a}}_{n-4}\}$. If $(1,\boldsymbol{y}')$ does not satisfy the conditions for $\boldsymbol{x}$ in Claim 8, then using the same logic as before we conclude that $\boldsymbol{y}'\in\mathbb{F}_2^{n-4}$ has $n-5$ runs. If $(\boldsymbol{x}',\bar{a}),(1,\boldsymbol{y}')$ do not satisfy the conditions in Claim 7 then $x'_2=x'_3$ or $x'_{n-4}=x'_{n-5}$. Suppose that $x'_2=x'_3$ (the case where $x'_{n-4}=x'_{n-5}$ follows from the same ideas). Since $\boldsymbol{x}'$ has $n-5$ runs, then $x'_1\neq x'_2$ and $x'_{n-6}\neq x'_{n-5}$. Thus, $(0,\boldsymbol{x}'),(\boldsymbol{y}',a)$ satisfy the conditions in Claim 7. Therefore, using the same logic as before

$$|\mathcal{X}_a^0| \leqslant 2D(n-5,t-2),$$

and

$$\begin{aligned}|\mathcal{X}_{\bar{a}}^1| &\leqslant D(n-5,t-2)+D(n-7,t-2)\\ &\quad+D(n-7,t-3)+D(n-9,t-4).\end{aligned}$$

Thus, applying Lemma 1 along with (3), it is possible to show that if $\boldsymbol{x},\boldsymbol{y}$ satisfy the conditions in this lemma, then

$$\begin{aligned}|\mathcal{X}| &\leqslant D(n-4,t-2)+3D(n-5,t-2)+D(n-6,t-2)\\ &\quad+D(n-7,t-2)+D(n-6,t-3)+D(n-7,t-3)\\ &\quad+D(n-8,t-4)+D(n-9,t-4)\leqslant N(n,t).\end{aligned}$$

∎

The next lemma will be directly used in the derivation of the upper bound. The proof relies on simply enumerating the possible choices for $\boldsymbol{x}, \boldsymbol{y}$ and applying previous results such as Lemma 6.

*Lemma 7: For all $n$ and $t$, let $\boldsymbol{x} = (1, 0, \boldsymbol{x}', \bar{a}, a) \in \mathbb{F}_2^n$, $\boldsymbol{y} = (0, 1, \boldsymbol{y}', a, \bar{a}) \in \mathbb{F}_2^n$ be such that $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$. Then,*

$$|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant N(n, t).$$

*Proof Sketch:* Let $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$. The result can be proven by considering the following eight cases:

(a) $\boldsymbol{x} = (1, 0, \bar{\boldsymbol{a}}_{n-4}, 1, 0)$, $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, 0, 1)$,
(b) $\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, 1, 0)$, $\boldsymbol{y} = (0, 1, \bar{\boldsymbol{a}}_{n-4}, 0, 1)$,
(c) $\boldsymbol{x} = (1, 0, \bar{\boldsymbol{a}}_{n-4}, 0, 1)$, $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, 1, 0)$,
(d) $\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, 0, 1)$, $\boldsymbol{y} = (0, 1, \bar{\boldsymbol{a}}_{n-4}, 1, 0)$,
(e) $\boldsymbol{x} = (1, 0, \bar{\boldsymbol{a}}_{n-4}, 1, 0)$, $\boldsymbol{y} = (0, 1, \bar{\boldsymbol{a}}_{n-4}, 0, 1)$,
(f) $\boldsymbol{x} = (1, 0, \bar{\boldsymbol{a}}_{n-4}, 0, 1)$, $\boldsymbol{y} = (0, 1, \bar{\boldsymbol{a}}_{n-4}, 1, 0)$,
(g) $\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, 1, 0)$, $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, 0, 1)$,
(h) $\boldsymbol{x} = (1, 0, \boldsymbol{a}_{n-4}, 0, 1)$, $\boldsymbol{y} = (0, 1, \boldsymbol{a}_{n-4}, 1, 0)$.

■

We now have the main result of this paper.

*Theorem 8: For all $n$ and $t$, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_2^n$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{F}_2^n$ be such that $d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant 2$. Then,*

$$|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant N(n, t).$$

*Proof:* The proof will be by induction on the lengths of the sequences $\boldsymbol{x}, \boldsymbol{y}$. The base cases for $n = 10, 11$ were verified using a computerized search.

Suppose the result holds for all $m < n$ where $n$ is the length of $\boldsymbol{x}, \boldsymbol{y}$. Let us write $\boldsymbol{x} = (x_1, \boldsymbol{x}')$, $\boldsymbol{y} = (y_1, \boldsymbol{y}')$ (for $x_1, y_1 \in \mathbb{F}_2$) and denote $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$. There are two cases:

1) $x_1 = y_1$, or
2) $x_1 \neq y_1$.

We first consider case 1). Then, $d_L(\boldsymbol{x}', \boldsymbol{y}') \geqslant 2$ and from the inductive hypothesis we get

$$\begin{aligned}\left|\mathcal{X}^{x_1}\right| &= \left|x_1 \circ \left(\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')\right)\right| = |\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')| \\ &\leqslant \max_{\boldsymbol{x}^*, \boldsymbol{y}^* \in \mathbb{F}_2^{n-1}, d_L(\boldsymbol{x}^*, \boldsymbol{y}^*) \geqslant 2} \{|\mathcal{D}_t(\boldsymbol{x}^*) \cap \mathcal{D}_t(\boldsymbol{y}^*)|\} \\ &\leqslant N(n-1, t).\end{aligned}$$

Suppose $x_k$ is the first occurrence of the symbol $\bar{x}_1$ in $\boldsymbol{x}$ and the symbol $\bar{x}_1$ appears in $\boldsymbol{x}$ not after it appears in $\boldsymbol{y}$. Notice that under this setup, we have $(x_1, x_2, \ldots, x_{k-1}) = (y_1, y_2, \ldots, y_{k-1}) = (x_1, x_1, \ldots, x_1)$ and so $d_L(\boldsymbol{x}'', \boldsymbol{y}'') \geqslant 2$, where $\boldsymbol{x}'' = (x_k, \ldots, x_n)$ and $\boldsymbol{y}'' = (y_k, \ldots, y_n)$. Since $k \geqslant 2$, applying the inductive hypothesis along with Lemma 2 gives

$$\begin{aligned}\left|\mathcal{X}^{\bar{x}_1}\right| &\leqslant |\bar{x}_1 \circ (\mathcal{D}_{t-(k-1)}(\boldsymbol{x}'') \cap \mathcal{D}_{t-(k-1)}(\boldsymbol{y}''))| \\ &= |\mathcal{D}_{t-(k-1)}(\boldsymbol{x}'') \cap \mathcal{D}_{t-(k-1)}(\boldsymbol{y}'')| \\ &\leqslant \max_{\boldsymbol{x}^*, \boldsymbol{y}^* \in \mathbb{F}_2^{n-k}, d_L(\boldsymbol{x}^*, \boldsymbol{y}^*) \geqslant 2} \{|\mathcal{D}_{t-(k-1)}(\boldsymbol{x}^*) \cap \mathcal{D}_{t-(k-1)}(\boldsymbol{y}^*)|\} \\ &\leqslant N(n-k, t-k+1) \\ &\leqslant N(n-2, t-1).\end{aligned}$$

According to Corollary 4, $N(n, t) = N(n-1, t) + N(n-2, t-1)$ which completes the proof for the case where $x_1 = y_1$.

In order to prove case 2), suppose that $x_1 \neq y_1$. Then, we also assume $x_n \neq y_n$ since otherwise if $x_n = y_n$ we can reverse the sequences and apply the same logic as above. However, if $x_1 \neq y_1$ and $x_n \neq y_n$ then from Lemmas 5 and 7, we have $|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant N(n, t)$ in this case as well and so the result holds. ■

## V. Conclusion and Open Problems

In this work, we solved the problem of computing the maximum intersection of the deletion balls for two sequences such that the two sequences belong to a single deletion-correcting code. Recall that our motivation for studying this problem was part of a larger effort to determine the quantity:

$$N(n, t_1, t_2) = \max_{\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_2^n, d_L(\boldsymbol{x}, \boldsymbol{y}) \geqslant t_1} \{|\mathcal{D}_{t_2}(\boldsymbol{x}) \cap \mathcal{D}_{t_2}(\boldsymbol{y})|\},$$

for arbitrary values of $t_1$ and so the problem of extending these results to the case of $t_1 \geqslant 3$ remains open.

There are a few possible directions for future work based upon the ideas introduced here. Recall from Section I that the maximum size of the set $\mathcal{D}_t(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{F}_2^n$ is achieved when $\boldsymbol{x} = \boldsymbol{a}_n$ and in this case we have $|\mathcal{D}_t(\boldsymbol{a}_n)| = D(n, t)$. For $t_1 = 1$, $N(n, t_1, t_2)$ is achieved for the two sequences: $\boldsymbol{x} = (1, 0, 1, 0, \ldots, 1, 0)$, $\boldsymbol{y} = (0, 1, 1, 0, \ldots, 1, 0)$ so that $\boldsymbol{x} = \boldsymbol{a}_n$ and $\boldsymbol{y}$ can be obtained from $\boldsymbol{a}_n$ by transposing the first two bits of $\boldsymbol{x}$. For $t_2 = 2$, we showed in this work that $N(n, t_1, t_2)$ is achieved when $\boldsymbol{x} = (1, 0, 1, 0, \ldots, 0, 1)$ and $\boldsymbol{y} = (0, 1, 1, 0, \ldots, 1, 0)$, so that $\boldsymbol{y}$ is as before and, in this case, $\boldsymbol{x}$ can be obtained from $\boldsymbol{a}_n$ by transposing the last two bits of $\boldsymbol{a}_n$. It remains an open question whether a similar pattern exists for sequences which obtain $N(n, t_1, t_2)$ when $t_1 \geqslant 3$.

## Appendix A
## Proof of Claim 6

*Proof of Claim 6:* We make use of the following lemma which can be found in [17].

*Lemma 9: (c.f. [17], Lemma 6) For any binary strings $\boldsymbol{u}, \boldsymbol{v}$, and for any bit $b$, $|\mathcal{D}_t((\boldsymbol{u}, b, b, \boldsymbol{v}))| \leqslant |\mathcal{D}_t((\boldsymbol{u}, b, \bar{b}, \bar{\boldsymbol{v}}))|$.*

According to Lemma 9, we only need to consider the case in which $\boldsymbol{x}$ has exactly $n-1$ runs, and assume without loss of generality that $x_1 = 1$. In this case, $\boldsymbol{x}$ is the alternating sequence $\boldsymbol{a}_{n-1}$, where one of its bits is repeated. For $1 \leqslant i \leqslant n-1$, let us denote by $\boldsymbol{a}_{n-1,i}$ the sequence $\boldsymbol{a}_{n-1}$ where its $i$-th bit is repeated, and the same for $\bar{\boldsymbol{a}}_{n-1,i}$.

Let us denote $g_1(n, t) = D(n-1, t) + D(n-3, t-2)$ and $g_2(n, t) = D(n-2, t) + D(n-4, t-2) + D(n-2, t-1)$ and for $3 \leqslant i \leqslant n-1$ we recursively define

$$g_i(n, t) = g_{i-2}(n-2, t-1) + g_{i-1}(n-1, t).$$

We will prove by induction that for $1 \leqslant i \leqslant n-1$,

$$|\mathcal{D}_t(\boldsymbol{a}_{n-1,i})| = g_i(n, t).$$

In the base case, we first verify that $g_1(n, t) = |\mathcal{D}_t(\boldsymbol{a}_{n-1,1})|$ and $g_2(n, t) = |\mathcal{D}_t(\boldsymbol{a}_{n-1,2})|$. For $i = 1$, notice that $|\mathcal{D}_t(\boldsymbol{a}_{n-1,1})| = D(n-1, t) + D(n-3, t-2) = g_1(n, t)$. In addition for $i = 2$, $|\mathcal{D}_t(\boldsymbol{a}_{n-1,2})| = g_1(n-1, t) + D(n-2,$

$t-1) = g_2(n,t)$. Let us assume that the claim holds for all $j < i$ and consider the sequence $\boldsymbol{a}_{n-1,i}$. The proof follows from the observation that

$$\begin{aligned}|\mathcal{D}_t(\boldsymbol{a}_{n-1,i})| &= |\mathcal{D}_t(\boldsymbol{a}_{n-1,i})^0| + |\mathcal{D}_t(\boldsymbol{a}_{n-1,i})^1| \\ &= |\mathcal{D}_{t-1}(\boldsymbol{a}_{n-3,i-2})| + |\mathcal{D}_t(\bar{\boldsymbol{a}}_{n-2,i-1})| \\ &= |\mathcal{D}_{t-1}(\boldsymbol{a}_{n-3,i-2})| + |\mathcal{D}_t(\boldsymbol{a}_{n-2,i-1})| \\ &= g_{i-2}(n-2,t-1) + g_{i-1}(n-1,t) \\ &= g_i(n,t).\end{aligned}$$

We now prove that $1 \leqslant i \leqslant n-1$, $g_i(n,t) \leqslant D(n-2,t) + D(n-4,t-2) + D(n-2,t-1)$ by induction on $i$. For the base case, for $g_2(n,t)$ we have equality and this inequality holds for $g_1(n,t)$ since

$$\begin{aligned}&D(n-2,t) + D(n-4,t-2) + D(n-2,t-1) - g_1(n,t) \\ &\qquad = D(n-4,t-2) - D(n-5,t-3) \geqslant 0,\end{aligned}$$

from (13). Now assume that $g_i(n,t) \leqslant D(n-2,t) + D(n-4, t-2) + D(n-2,t-1)$ for all $i \leqslant i^{(*)}$ and consider the case where $i = i^{(*)} + 1$. Then,

$$\begin{aligned}g_{i^{(*)}}(n,t) &= g_{i^{(*)}-2}(n-2,t-1) + g_{i^{(*)}-1}(n-1,t) \\ &\leqslant D(n-4,t-1) + D(n-6,t-3) \\ &\quad + D(n-4,t-2) + D(n-3,t) \\ &\quad + D(n-5,t-2) + D(n-3,t-1) \\ &= D(n-2,t) + D(n-4,t-2) + D(n-2,t-1),\end{aligned}$$

which follows by applying (3) three times. ∎

## Appendix B
## Proof Sketch of Claim 7

*Proof Sketch of Claim 7:* Let us first denote $H(n,t) = D(n-2,t-1) + D(n-4,t-1) + D(n-4,t-2) + D(n-6,t-3)$ for $t < n/2$ and $H(n,t) = 2^{n-t}$ for $t \geqslant n/2$. The proof is by induction. The cases where $n = 5, 6$ were verified by a computerized search. Suppose the result holds for all $n < m$. We consider two cases:

1) $\boldsymbol{x}, \boldsymbol{y}$ agree in the first and last bits, and
2) $\boldsymbol{x}, \boldsymbol{y}$ disagree in the first bit or in the last bit.

Suppose that 1) holds so that $x_1 = y_1 = a \in \mathbb{F}_2$. Let $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$. Then, from Lemma 2, we have

$$\left|\mathcal{X}^a\right| = \left|a \circ \left(\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')\right)\right| = \left|\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')\right| \quad (16)$$

where $\boldsymbol{x}'$, $\boldsymbol{y}'$ are the length-$(n-1)$ sequences that satisfy $\boldsymbol{x} = (a, \boldsymbol{x}')$ and $\boldsymbol{y} = (a, \boldsymbol{y}')$. Now if $\boldsymbol{x}'$, $\boldsymbol{y}'$ satisfy the conditions in the claim, we can apply the inductive hypothesis and conclude

$$\left|\mathcal{X}^a\right| \leqslant H(n-1,t) \quad (17)$$

Suppose $x_k$ is the first occurrence of the symbol $\bar{a}$ in $\boldsymbol{x}$ and the symbol $\bar{a}$ appears in $\boldsymbol{x}$ not after it appears in $\boldsymbol{y}$. If $\boldsymbol{x}'' = (x_{k+1}, x_{k+2}, \ldots, x_n)$, $\boldsymbol{y}'' = (y_{k+1}, y_{k+2}, \ldots, y_n)$ satisfy the conditions in the claim, we can apply the inductive hypothesis along with Lemma 2 so that

$$\left|\mathcal{X}^{\bar{a}}\right| \leqslant H(n-2,t-1).$$

Thus, if $x_1 = y_1$ and $\boldsymbol{x}'$, $\boldsymbol{y}'$ and $\boldsymbol{x}''$, $\boldsymbol{y}''$ satisfy the conditions in the claim, then $|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant H(n-1,t) + H(n-2, t-1) = H(n,t)$.

Suppose $x_1 = y_1$ and $\boldsymbol{x}'$, $\boldsymbol{y}'$ do not satisfy the conditions in the claim. Then we have that one of the following holds since $\boldsymbol{x} \neq \boldsymbol{y}$:

1) $x'_2 = x'_3$, or
2) $x'_{n-2} = x'_{n-1}$,
3) $x_1 = x_2$.

If $x_1 = x_2$, then $|\mathcal{X}| \leqslant 2D(n-3,t-1) + D(n-3,t-2)$, which, using the recurrence from (3), can be shown is less than $H(n,t)$. Thus, we need to only consider cases 1) and 2). Without loss of generality, we assume $x'_2 = x'_3$. Then, $x_3 = x_4$ and since $\boldsymbol{x}$ has $n-1$ runs we have $x_2 \neq x_3$, $x_{n-2} \neq x_{n-3}, x_{n-3} \neq x_{n-4}$. Let $R(\mathcal{X}) = \mathcal{D}_t((R(\boldsymbol{x})) \cap \mathcal{D}_t(R(\boldsymbol{y}))$. Then, since $x_1 = a = y_1$ we can apply the inductive hypothesis to conclude

$$\begin{aligned}|\mathcal{X}^a| &= |R(\mathcal{X}_a)| = |\left(\mathcal{D}_t(R(\boldsymbol{x}')) \cap \mathcal{D}_t(R(\boldsymbol{y}'))\right) \circ a| \\ &\leqslant H(n-1,t),\end{aligned}$$

where $R(\boldsymbol{x}')$, $R(\boldsymbol{y}')$ satisfy the inductive hypothesis since $x_2 \neq x_3, x_{n-2} \neq x_{n-3}$, and $x_{n-3} \neq x_{n-4}$. If $\boldsymbol{x}''$, $\boldsymbol{y}''$ satisfy the conditions in the lemma, then $|\mathcal{X}^{\bar{a}}| \leqslant H(n-2,t-1)$ and the result follows.

We still need to consider the case where $\boldsymbol{x}'$, $\boldsymbol{y}'$ from (16) satisfy the conditions in the lemma but $\boldsymbol{x}''$, $\boldsymbol{y}''$ from (17) do not. Then, we can handle this case similar to the one where $\boldsymbol{x}' \neq \boldsymbol{y}'$, and we can conclude that $|\mathcal{X}| \leqslant H(n,t)$ as desired. We have just shown that if $x_1 = y_1$, then the statement in the claim holds.

Suppose then that $x_1 \neq y_1 = a \in \mathbb{F}_2$; this case can be handled by considering all sequences in the intersection that start with the symbol $a$ followed by all sequences that start with the symbol $\bar{a}$. The case where $x_n \neq y_n$ can be handled by considering the reverse of the sequences $\boldsymbol{x}$, $\boldsymbol{y}$. ∎

## Appendix C
## Sketch of Proof of Claim 8

A slightly stronger result can be proven. Let $H(n,t)$ be as defined in Appendix B.

*Claim 9: Suppose $n, t$ are integers where $n \geqslant 5$. Suppose $\boldsymbol{x} \in \mathbb{F}_2^n$, $\boldsymbol{y} \in \mathbb{F}_2^n$ are such that $\boldsymbol{x} \neq \boldsymbol{y}$ and the number of runs in $\boldsymbol{x}$ is at most $n-2$. Then,*

$$|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant H(n,t).$$

*Proof Sketch:* The proof can be done by induction. The case where $n = 5, 6$ was verified by a computerized search. Suppose the results holds for all $n < m$. Then the result can be proven by considering the cases where 1) $\boldsymbol{x}, \boldsymbol{y}$ either agree in the first and last bits or 2) $\boldsymbol{x}, \boldsymbol{y}$ disagree in the first or the last bits. Let $\mathcal{X} = \mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})$.

Suppose 1) holds so that $x_1 = y_1$ and $x_n = y_n$. Then, similar to the proof of Claim 7

$$\left|\mathcal{X}^a\right| = \left|a \circ \left(\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')\right)\right| = \left|\mathcal{D}_t(\boldsymbol{x}') \cap \mathcal{D}_t(\boldsymbol{y}')\right|.$$

If $\boldsymbol{x}'$, $\boldsymbol{y}'$ satisfy the conditions either in the inductive assumption or the conditions from Claim 7 or Claim 8, we can write

$$\begin{aligned}\left|\mathcal{X}^a\right| &\leqslant \max_{\boldsymbol{x}',\boldsymbol{y}'\in\mathbb{F}_2^{n-1}}\{|\mathcal{D}_t(\boldsymbol{x}')\cap\mathcal{D}_t(\boldsymbol{y}')|\}\\ &\leqslant H(n-1,t).\end{aligned}$$

If $\boldsymbol{x}'' = (x_{k+1}, x_{k+2}, \ldots, x_n)$, and $\boldsymbol{y}'' = (y_{k+1}, y_{k+2}, \ldots, y_n)$ satisfy the conditions in this claim or Claim 7 or Claim 8, we can again use the inductive hypothesis to write $\left|\mathcal{X}^{\bar{a}}\right| \leqslant H(n-2,t-1)$. Notice that if the following conditions are true, then Claim 8 holds:

1) $x_1 = y_1, x_n = y_n$,
2) $\boldsymbol{x}'$, $\boldsymbol{y}'$ satisfy the conditions in Claim 7 or Claim 8, and
3) $\boldsymbol{x}''$, $\boldsymbol{y}''$ satisfy the conditions in Claim 7 or Claim 8.

For the case where $x_1 = y_1, x_n = y_n$ and $\boldsymbol{x}'$, $\boldsymbol{y}'$ do not satisfy the conditions in Claim 7 or Claim 8, it can be shown that $|\mathcal{X}| = 2D(n-3,t-1) + D(n-3,t-2)$, which is less than $H(n,t)$.

Suppose now that $x_1 = y_1, x_n = y_n$ and $\boldsymbol{x}''$, $\boldsymbol{y}''$ do not satisfy the conditions in Claim 7 or Claim 8, but $\boldsymbol{x}'$, $\boldsymbol{y}'$ satisfy the inductive assumption. Then, it can be shown from induction that $|\mathcal{X}^a| \leqslant H(n-1,t)$, and $\left|\mathcal{X}^{\bar{a}}\right| \leqslant 2D(n-5, t-2) + D(n-5,t-3) \leqslant H(n-2,t-1)$. This implies that when $x_1 = y_1, x_n = y_n$, then $|\mathcal{D}_t(\boldsymbol{x}) \cap \mathcal{D}_t(\boldsymbol{y})| \leqslant H(n,t)$.

The last case to consider is where $x_1 \neq y_1 = a \in \mathbb{F}_2$. It can be shown that the result holds in this case as well and so the statement holds. ∎

## ACKNOWLEDGEMENTS

The authors would like to thank the associate editor along with the numerous anonymous reviewers for their help in improving the quality of the paper.

## REFERENCES

[1] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proc. 15th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2004, pp. 903–911.
[2] L. Calabi, "On the computation of Levenshtein's distances," Parke Math. Labs. Inc., Carlisle, MA, USA, Tech. Rep., 1967.
[3] R. Gabrys and E. Yaakobi, "Sequence reconstruction over the deletion channel," in *Proc. Int. Symp. Inf. Theory*, Jul. 2016, pp. 1596–1600.
[4] D. S. Hirschberg and M. Regnier, "Tight bounds on the number of string subsequences," *J. Discrete Algorithms*, vol. 1, no. 1, pp. 123–132, 2000.
[5] B. Haeupler and M. Mitzenmacher, "Repeated deletion channels," in *Proc. IEEE Inform. Theory Workshop*, Hobart, TAS, Australia, Nov. 2014, pp. 152–156.
[6] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proc. 19th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2008, pp. 389–398.
[7] T. Jiang and A. Vardy, "Asymptotic improvement of the Gilbert–Varshamov bound on the size of binary codes," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1655–1664, Aug. 2004.
[8] S. Kannan and A. McGregor, "More on reconstructing strings from random traces: Insertions and deletions," in *Proc. IEEE Int. Sym. Inform. Theory*, Adelaide, SA, Australia, Sep. 2005, pp. 297–301.
[9] E. Konstantinova, "Reconstruction of permutations distorted by single reversal errors," *Discrete Appl. Math.*, vol. 155, no. 18, pp. 2426–2434, 2007.
[10] E. Konstantinova, V. Levenshtein, and J. Siemons. (Feb. 2007). "Reconstruction of permutations distorted by single transposition errors." [Online]. Available: https://arxiv.org/abs/math/0702191v1
[11] E. Konstantinova, "On reconstruction of signed permutations distorted by reversal errors," *Discrete Math.*, vol. 308, pp. 974–984, Mar. 2008.
[12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.
[13] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 2–22, Jan. 2001.
[14] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *J. Combinat. Theory A*, vol. 93, no. 2, pp. 310–332, 2001.
[15] V. I. Levenshtein, E. Konstantinova, E. Konstantinov, and S. Molodtsov, "Reconstruction of a graph from 2-vicinities of its vertices," *Discrete Appl. Math.*, vol. 156, pp. 1399–1406, May 2008.
[16] V. I. Levenshtein and J. Siemons, "Error graphs and the reconstruction of elements in groups," *J. Combinat. Theory, A*, vol. 116, no. 4, pp. 795–815, 2009.
[17] Y. Liron and M. Langberg, "A characterization of the number of subsequences obtained via the deletion channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2300–2312, May 2015.
[18] R. Mittu and F. Segaria, "Common operational picture (COP) and common tactical picture (CTP) management via a consistent networked information stream," in *Proc. Command Control Res. Technol. Symp.*, Monterey, CA, USA, 2000, pp. 3–7.
[19] A. S. Motahari, G. Bresler, and D. Tse. (Mar. 2012). "Information theory of DNA shotgun sequencing." [Online]. Available: https://arxiv.org/abs/1203.6233
[20] S. Motahari, G. Bresler, and D. Tse, "Information theory for DNA sequencing: Part I: A basic model," in *Proc. IEEE Int. Symp. Inform. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 2741–2745.
[21] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017.
[22] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Proc. 19th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2008, pp. 399–408.
[23] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," in *Proc. Int. Symp. Inform. Theory*, Jul. 2012, pp. 106–110.
[24] E. Yaakobi, M. Schwartz, M. Langberg, and J. Bruck, "Sequence reconstruction for Grassmann graphs and permutations," in *Proc. Int. Symp. Inform. Theory*, Jul. 2013, pp. 874–878.

**Ryan Gabrys** received his Ph.D. degree in electrical engi- neering from the University of California, Los Angeles. Since 2014, he has been a postdoctoral researcher at the University of Illinois, Urbana Champaign. Currently, he works at SPAWAR Systems Center San Diego. His research interests include cod- ing theory with applications to storage and synchronization.

**Eitan Yaakobi** (S'07–M'12–SM'17) is an Assistant Professor at the Computer Science Department at the Technion Israel Institute of Technology. He received the B.A. degrees in computer science and mathematics, and the M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California, San Diego, in 2011. Between 2011-2013, he was a postdoctoral researcher in the department of Electrical Engineering at the California Institute of Technology. His research interests include information and coding theory with applications to non-volatile memories, associative memories, data storage and retrieval, and voting theory. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship in 2010-2011.