On Unique Decoding from Insertions and Deletions

Kayvon Mazooji
kmazooji1@ucla.edu
UCLA, Los Angeles, CA 90095

Abstract—In this paper, we study how often unique decoding from \( t \) insertions or \( t \) deletions occurs for error correcting codes. Insertions and deletions frequently occur in synchronization problems and DNA, a medium which is beginning to be used for long term data storage.

We define natural probabilistic channels that make \( t \) insertions or \( t \) deletions, and study the probability of unique decoding. Our most substantial contribution is the derivation of tight upper bounds on the probability of unique decoding for messages passed through these channels. We also consider other aspects of the problem, and derive improved upper bounds for linear codes and VT-codes.

Index Terms—Insertions and Deletions, Codes, Combinatorics, Sequence reconstruction, Varshamov-Tenengolts codes

I. INTRODUCTION

Codes correcting insertions and deletions have historically been important to problems in synchronization [22]. Recently, such codes have been useful for DNA storage as well [14]. Researchers have had little luck in finding insertion correcting codes of optimal cardinality. However, progress has been made on finding upper and lower bounds on the optimal cardinality of such codes [10], [11], [13]. The most famous code believed to be optimal is the VT-code [1], which corrects a single insertion or deletion [2]. Other codes for insertions are found in [16], [17], [18], [19]. List decoding from insertions and deletions is considered in [20], [21].

In this paper we study the probability of uniquely decoding from insertions and deletions for broad classes of codes. This problem is particularly relevant to long term storage problems (e.g. DNA storage [15]). For example, side information may not be available if a codeword is recovered hundreds of years in the future, making a list of \( > 1 \) decoded codewords insufficient. It is therefore very desirable for a code to have a reasonable chance of being uniquely decodable beyond the error-correction guarantee, especially if no other properties are compromised.

Consider a channel which makes \( t \) insertions into a codeword \( c \) and outputs each distinct length \( n + t \) received word with equal probability. The probability of unique decoding for a particular codeword \( c \) is equal to the fraction of length \( n + t \) received words that are unique to \( c \). We refer to this channel as the uniform \( t \)-supersequence channel.

Now consider the channel where \( t \) insertions occur, one at a time. We refer to the temporal list of insertions as a \( t \)-insertion history. An insertion is represented by a tuple of the form: (position, element). For the codeword 00, an example insertion history is \([0, 1), (0, 0)]\), giving the received word 0100.

If for each insertion, the element and position are chosen uniformly at random, we refer to this channel as the uniform \( t \)-insertion channel. If a particular codeword \( c \) in a code \( C \subseteq \mathbb{F}_q^n \) is passed through this channel, then the probability that unique decoding occurs is equal to the fraction of \( t \)-insertion histories that produce a received word unique to \( c \).

We also define the analogous channels for \( t \) deletion errors in Section II.

Clearly, we can make optimizations involving these probabilities over codes of a particular cardinality, length, and error-correction capability. While it is not yet clear how to perform such optimizations in a non-exhaustive manner, upper bounds on the probabilities are useful because they help us understand how effective a code can possibly be for uniquely decoding a given number of insertions or deletions, without solving the optimization problems explicitly. The bounds can thus be used as a reference when designing codes, and can be used to prove theorems. The purpose of focusing on these channels is to provide a framework for understanding what happens when \( t \) insertions or deletions occur, regardless of whether the real-life channel always makes \( t \) insertions or deletions.

The remainder of the paper is organized as follows. In Section II, we present the necessary preliminaries. In Section III we provide tight upper bounds on the probability of unique decoding for both insertion channels, discuss the positivity of the measures, and present improved upper bounds for insertion channels that apply to VT-codes and linear codes. In Section IV, we derive tight upper bounds on the probability of unique decoding for one of the deletion channels, and give improved upper bounds with additional assumptions. In Section V, we make observations about the behavior of VT-codes, and raise open questions. We conclude the paper in Section VI.

II. PRELIMINARIES

Let \( \mathbb{F}_q = \{0, 1, \ldots, q-1\} \) be an alphabet containing \( q \in \mathbb{N} \) symbols. We only consider \( q \geq 2 \) throughout the paper. \( \mathbb{F}_q^n \) is the set of all length \( n \) words over the alphabet \( \mathbb{F}_q \). If \( a \in \mathbb{F}_q \), we denote the word

\[
\overbrace{aa\ldots a}^{n \ \text{times}}
\]

by \( a_n \). For a sequence \( c \in \mathbb{F}_q^n \), let \( c[i] \) be the \( i \)th element in \( c \) where \( i \in \{0, 1, \ldots, n-1\} \). We use the terms word and sequence interchangeably.
Let $x \in \mathbb{F}_q^n$. We define an insertion as the addition of an element from $\mathbb{F}_q$ into some position in $x$. We define a deletion as the removal of an element from $x$. If $t$ insertions occur, the resulting word is referred to as a $t$-supersequence of $x$. Similarly, if $t$ deletions occur, the resulting word is referred to as a $t$-subsequence of $x$. A substring is a contiguous subsequence of a word. Given a word, a run is a substring $a_n$, such that the potential elements on both sides of $a_n$ are not $a$. The Levenshtein distance $d_L(x,y)$ between two words $x \in \mathbb{F}_q^n$, $y \in \mathbb{F}_q^m$, is defined as the minimum number of insertions and deletions necessary to transform $x$ into $y$. Clearly, we have that $d_L(x,y) = d_L(y,x)$.

The $t$-insertion ball of $x$ is the set of all words in $\mathbb{F}_q^{n+t}$ that are formed by inserting $t$ symbols into $x$. We denote the $t$-insertion ball of $x$ by $I_t(x)$. It is known that $|I_t(x)| = \sum_{i=0}^{t} \binom{n+t}{i}(q-1)^i$ for any word $x \in \mathbb{F}_q^n$. Similarly, the $t$-deletion ball of $x$ is the set of all words in $\mathbb{F}_q^{n-t}$ formed by deleting $t$ symbols from $x$. We denote the $t$-deletion ball of $x$ by $D_t(x)$. Unfortunately, a general formula for $|D_t(x)|$ is not known. However in the $t=1$ case, we have that $|D_1(x)| = r(x)$, where $r(x)$ is the number of runs in $x$.

Because $|I_t(x)|$ is independent of the exact length $n$-sequence $x$, we will sometimes use the expression $I_t(n,q)$ to mean the number of sequences in the $q$-ary $t$-insertion ball of a length $n$-sequence.

We define a $t$-insertion correcting code to be a set of codewords $C \subseteq \mathbb{F}_q^n$ such that $I_t(c_1) \cap I_t(c_2) = \emptyset \ \forall c_1, c_2 \in C, c_1 \neq c_2$. Similarly, we define a $t$-deletion correcting code to be a set of codewords $C \subseteq \mathbb{F}_q^n$ such that $D_t(c_1) \cap D_t(c_2) = \emptyset \ \forall c_1, c_2 \in C, c_1 \neq c_2$.

It was shown that $C$ is a $t$-insertion correcting code if and only if it is a $t$-deletion correcting code. Furthermore, it was also shown that $C$ is a $t$-insertion correcting code if and only if $d_L(c_1,c_2) > 2t \ \forall c_1, c_2 \in C, c_1 \neq c_2$. We denote the minimum Levenshtein distance of an insertion/deletion correcting code by $d_{\text{min}}$, and denote the cardinality of the code by $M$.

Varshamov-Tenengolts codes introduced in [11] are commonly used single-insertion correcting codes defined as the set of all words $x = (x_1, ..., x_n) \in \mathbb{F}_2^n$ such that

$$\sum_{i=1}^{n} ix_i \equiv a \pmod{n+1}$$

for some $a$ such that $0 \leq a \leq n$.

Varshamov-Tenengolts codes are perfect, and are optimal for $n = 1, ..., 9$ when $a = 0$. It is conjectured, though unproven that Varshamov-Tenengolts codes are optimal for all values of $n$ when $a = 0$. We refer to the Varshamov-Tenengolts code of length $n$ with parameter $a$ as $V T_a(n)$. Some excellent general resources on insertions and deletions can be found in [8, 11], and [9].

We focus on two distinct insertion channels, namely the uniform $t$-supersequence channel, which we denote by USC$_t$, and the uniform $t$-insertion channel, which we denote by UIC$_t$. These channels are both defined in the introduction.

We consider two deletion channels which make $t$ deletions. The first is the uniform $t$-deletion channel, denoted by UDC$_t$. This channel makes $t$ sequential deletions, where the deletion at each step is chosen with equal probability. This is equivalent to saying each $t$-deletion history occurs with equal probability. Here, a $t$-deletion history is represented as a temporal list of deletions of length $t$, where the $i$th deletion is represented by the index of the element in the length $n-i+1$ word that is deleted at step $i \in \{1, ..., t\}$. For example, if the word 001100 is affected by the 2 deletion history [0, 1], the received word would be 0100.

The second deletion channel is the uniform $t$-subsequence channel, which outputs each distinct $t$-subsequence of a word with equal probability. We denote this channel by UBC$_t$.

We define a unique $t$-supersequence of a codeword $c \in C \subseteq \mathbb{F}_q^n$ as a $t$-supersequence of $c$ that is not a $t$-supersequence of any other codeword in $C$.

We define a unique $t$-subsequence of a codeword $c \in C \subseteq \mathbb{F}_q^n$ as a $t$-subsequence of $c$ that is not a $t$-subsequence of any other codeword in $C$.

We consider two measures of a code’s effectiveness for uniquely decoding beyond its error-correction guarantee. Let $f_K(c,C)$ be the probability of $c \in C$ being uniquely decodable after being passed through channel $K$. The first measure is $W_K(C) = \min_{c \in C} I_K(c,C)$. This measure addresses the worst case distribution on the codeword sent i.e. no matter which codeword is sent, the probability of unique decoding is at least $W_K(C)$. The second measure is $U_K(C) = \frac{1}{M} \sum_{c \in C} I_K(c,C)$. This measure gives the probability of unique decoding if each codeword is sent with equal probability.

A $W_K$ optimal code is a code that solves

$$\max_{C: \{n,q,d_{\text{min}},M\} \text{ fixed}} W_K(C).$$

A $U_K$ optimal code is a code that solves

$$\max_{C: \{n,q,d_{\text{min}},M\} \text{ fixed}} U_K(C).$$

Finally, we define $B(n,q,K)$ as a tight upper bound on $f_K(c,C)$ over all codes $C \subseteq \mathbb{F}_q^n$ such that $|C| \geq 2$, and over all codewords $c \in C$.

Similarly, we define $B_d(n,q,K)$ as a tight upper bound on $f_K(c,C)$ over all codes $C \subseteq \mathbb{F}_q^n$ such that $|C| \geq 2$, and over all codewords $c \in C$ such that there is a codeword $c' \in C$ where $d_L(c,c') = 2d$ for $1 \leq d \leq n$.

### III. Bounds for Insertions

#### A. General Bounds

In this subsection we find $B(n,q,K)$ for $K \in \{\text{USC}_t, \text{UIC}_t\}$, consider the limiting behavior of the upper bounds as $t$ or $n$ increases, and establish the positivity of $U_K$ for $K \in \{\text{USC}_t, \text{UIC}_t\}$ and all $t$. Recall that $q \in \mathbb{N}$, $q \geq 2$ throughout the paper.

We begin by recalling a recursion for the intersection cardinality of two insertion balls. This recursion was discovered...
by Levenshtein in [4], where he studied what is now known as Levenshtein’s reconstruction problem.

\textbf{Lemma 1.} Let $X' \in \mathbb{F}_q^{n+t-k}$, $Y' \in \mathbb{F}_q^n$, where $n, t, k$ are positive integers such that $k < n + t$. Write $X' = aX$ and $Y' = bY$ with $a, b \in \mathbb{F}_q$. Then, if $a = b$, 
\[
|I_k(X') \cap I_l(Y')| = |I_k(X) \cap I_l(Y)| + (q - 1)|I_{k-1}(aX) \cap I_{l-1}(aY)|.
\]

If $a \neq b$, 
\[
|I_k(X') \cap I_l(Y')| = |I_k(X) \cap I_l(bY)| + |I_{k-1}(aX) \cap I_l(Y)| + (q - 2)|I_{k-1}(aX) \cap I_l(bY)|.
\]

We define the minimum intersection cardinality as 
\[
\hat{N}_q^+(n_1, n_2, t_1, t_2) = \min_{X \in \mathbb{F}_q^{n_1}, Y \in \mathbb{F}_q^{n_2}} |I_1(X) \cap I_2(Y)|
\]
where $n_1, n_2, t_1, t_2 \in \mathbb{N}$, and $t_1 + n_1 = t_2 + n_2$. Clearly, $\hat{N}_q^+(n_1, n_2, t_1, t_2) = \hat{N}_q^+(n_2, n_1, t_2, t_1)$. In Theorem [1] we give a closed form for $\hat{N}_q^+(n_1, n_2, t_1, t_2)$ in his study of sequence reconstruction [4].

\textbf{Theorem 1.} Let $n_1, n_2, t_1, t_2 \in \mathbb{N}$ and $t_1 + n_1 = t_2 + n_2$. Then, we have 
\[
\hat{N}_q^+(n_1, n_2, t_1, t_2) = \sum_{k=n_1}^{t_2} \sum_{i=0}^{k-n_1} \binom{k}{i} (q - 2)^i \binom{n_2 + t_2}{k}.
\]

To prove Theorem [1], we found $I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})$ to equal the formula in the theorem. We then proved the formula is the minimum intersection cardinality using an inductive argument on $n_1 + t_1$. Lemma [1] was used in the inductive step. The proof is very long, and is given in the appendix. It should be noted that every pair of sequences at Levenshtein distance $n_1 + n_2$ achieves the minimum intersection cardinality, as given in Lemma [2].

\textbf{Lemma 2.} Let $n_1, n_2, t_1, t_2 \in \mathbb{N}$ and $N = t_1 + n_1 = t_2 + n_2$. Then for any length $n_1$ sequence $X$ and any length $n_2$ sequence $Y$ such that $d_L(X, Y) = n_1 + n_2$, we have that 
\[
|I_{t_1}(X) \cap I_{t_2}(Y)| = \hat{N}_q^+(n_1, n_2, t_1, t_2) = \sum_{k=n_1}^{t_2} \sum_{i=0}^{k-n_1} \binom{k}{i} (q - 2)^i \binom{n_2 + t_2}{k}.
\]

In addition, the following corollary can be proved logically, or through an application of the binomial theorem as shown below.

\textbf{Corollary 1.} For $n, t \in \mathbb{N}$, we have that 
\[
\hat{N}_q^+(0, n, n + t, t) = I_t(n, q).
\]

\textbf{Proof:}

\[
\hat{N}_q^+(0, n, n + t, t) = \sum_{k=0}^{t} \sum_{i=0}^{k} \binom{k}{i} (q - 2)^i \binom{n + t}{k}
= \sum_{k=0}^{t} (q - 1)^k \binom{n + t}{k} = I_t(n, q).
\]

where the second equality follows from the binomial theorem.

With Theorem [1] we were able to derive $B(n, q, USC_t)$. Consider any length $n$ code $C$ with cardinality $\geq 2$. The following theorem gives a tight upper bound on the fraction of unique $t$-supersequences for any $c \in C$. It is thus an upper bound on the probability of unique decoding for any codeword under the uniform $t$-supersequence channel.

\textbf{Theorem 2.} For $n, t \in \mathbb{N}$, we have that 
\[
B(n, q, USC_t) = 1 - \frac{\sum_{k=0}^{t} \sum_{j=0}^{k-n} \binom{k}{j} (q - 2)^j \binom{n+t}{k}}{\sum_{i=0}^{t} \binom{n+i}{i} (q - 1)^i}.
\]

\textbf{Proof:} In addition to $c$, there must exist another codeword $c'$ since $|C| \geq 2$. The quantity $|I_t(c) \cap I_t(c')|$ must be greater than or equal to 
\[
\hat{N}_q^+(n, n, t, t) = \sum_{k=n}^{t} \sum_{j=0}^{k-n} \binom{k}{j} (q - 2)^j \binom{n + t}{k}
\]
as proved in Theorem [1]. Thus, the number of sequences unique to $c$ in $I_t(c)$ must be less than or equal to 
\[
\sum_{i=0}^{t} \binom{n + t}{i} (q - 1)^i - \sum_{k=n}^{t} \sum_{j=0}^{k-n} \binom{k}{j} (q - 2)^j \binom{n + t}{k}
\]
The upper bound equals $B(n, q, USC_t)$ because $f_{USC_t}(0_n, \{0_n, 1_n\})$ achieves the upper bound as proved in Lemma [2].

In addition to providing an upper bound on the fraction of unique $t$-supersequences for a particular codeword, the formula above serves as a tight upper bound on $W_{USC_t}(C)$ and $U_{USC_t}(C)$.

The upper bound on the fraction of unique $t$-supersequences approaches zero as $t$ goes to infinity for fixed $n$ in the binary case as proved in Lemma [3] and exemplified in Figure [1]. The bound is clearly equal to one as $n$ goes to infinity for fixed $t$ because the minimum intersection cardinality is only positive for $t \geq n$. If $t = n + O(n^a)$ for $a < .5$, our tight upper bound approaches one in the binary case as $n$ goes to infinity. This is proved in Lemma [4] and is relevant to the questions raised in Section V.

\textbf{Lemma 3.} For $n, t \in \mathbb{N}$, we have
\[
limit_{t \to \infty} B(n, 2, USC_t) = 0.
\]
Proof:
\[
\lim_{t \to \infty} B(n, 2, \text{USC}_t) = \lim_{t \to \infty} \left( 1 - \frac{\sum_{k=n}^{t} \sum_{j=0}^{k-n} \binom{k}{j} (q-2)^j \binom{n+t}{k} (q-1)^i}{\sum_{j=0}^{n+t} \binom{n+t}{j} (q-1)^i} \right) = 1 - \lim_{t \to \infty} \frac{\sum_{k=n}^{t} \sum_{j=0}^{k-n} \binom{k}{j} (q-2)^j \binom{n+t}{k} (q-1)^i}{\sum_{j=0}^{n+t} \binom{n+t}{j} (q-1)^i} = 0
\]
The first equality follows from plugging in \( q = 2 \). The second equality follows from the fact that the \( n \) extra terms in the denominator are dominated by the other terms as \( t \) increases.

Lemma 4. For all values of \( n \in \mathbb{N} \) and \( \alpha < .5 \), we have
\[
\lim_{n \to \infty} B(n, 2, \text{USC}_{n+O(n^\alpha)}) = 1.
\]

Proof: Let \( f : \mathbb{Z} \to \mathbb{Z} \) be some function. For \( q = 2 \) and letting \( t = n + f(n) \),
\[
\lim_{n \to \infty} B(n, 2, \text{USC}_{n+f(n)}) = \lim_{n \to \infty} \left( 1 - \frac{\sum_{k=n}^{n+f(n)} \sum_{j=0}^{k-n} \binom{k}{j} (q-2)^j \binom{n+2f(n)}{k} (q-1)^i}{\sum_{j=0}^{n+2f(n)} \binom{n+2f(n)}{j} (q-1)^i} \right) = 1 - \lim_{n \to \infty} \frac{\sum_{k=n}^{n+f(n)} \sum_{j=0}^{k-n} \binom{k}{j} (q-2)^j \binom{n+2f(n)}{k} (q-1)^i}{\sum_{j=0}^{n+2f(n)} \binom{n+2f(n)}{j} (q-1)^i} \geq 1 - \lim_{n \to \infty} \frac{2(n+2f(n))^{n+2f(n)/2}}{2(n+f(n))^{n+f(n)/2}} \geq 1 - \lim_{n \to \infty} \frac{2(f(n) + 1)}{\sqrt{3(n + f(n))/2} + 1} = 1.
\]

The third line follows from plugging in \( q = 2 \). To obtain the denominator in the fourth line, we observe that
\[
\sum_{k=n}^{n+f(n)} \binom{2n+f(n)}{k} \leq (f(n) + 1)^{2n+f(n)/2} \binom{2n+f(n)/2}{n}/(2n+f(n)/2) + 1,
\]
and the fact the \( \binom{2n}{k} \) is maximized at \( k = n \). To obtain the denominator in the fourth line, we observe that
\[
\sum_{i=0}^{n+f(n)} \binom{2n+f(n)}{i} = \sum_{i=0}^{n+f(n)/2} \binom{2n+f(n)/2}{i} = \frac{1}{2} 2^{2n+f(n)} = 2f(n + 1)^{n+1}.
\]
The proof when \( 2n + f(n) \) is odd uses the same ideas after an application of Pascal’s rule to the binomial coefficients in the third line.

To improve this upper bound for a code at minimum Levenshtein distance \( d_{\text{min}} \), we could find \( B_{d_{\text{min}}}(n, q, \text{USC}_t) \). This could be approached by finding the minimum insertional ball intersection cardinality over all pairs of sequences that have Levenshtein distance \( \leq d_{\text{min}} \), and proving that this measure increases as \( d_{\text{min}} \) decreases. Such a formula could be thought of as the dual of the formula derived by Sala et al. in their study of Levenshtein’s reconstruction problem [9], which gives the maximum possible insertional ball intersection cardinality over all sequence pairs with Levenshtein distance \( \geq d_{\text{min}} \). The derivation of this formula in conjunction with [9] would result in a spectrum of insertional ball intersection cardinalities among the Levenshtein distance between the sequences in the pair. It would also be interesting to derive a formula for how many sequence pairs have a particular Levenshtein distance.

To upper bound the probability of unique decoding for the uniform \( t \)-insertion channel, we find the analogue of the minimum intersection cardinality for insertion histories in Theorem 3.

Theorem 3. For any distinct sequences \( X, Y \in \mathbb{F}_q^n \) with \( n, t \in \mathbb{N} \), the fraction of \( t \)-insertion histories for \( X \) giving \( t \)-supersequences in \( I_t(Y) \) is lower bounded by \( \sum_{i=0}^{t-n} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i) \). The lower bound is tight.

Proof: The number of insertion histories for \( X \) where \( Y \) is a subsequence of the insertion pattern is equal to \( I_t(n, q) \Pi_{i=1}^{t-n}(n+i) = \sum_{i=0}^{t-n} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i) \). This is because there are \( \Pi_{i=1}^{t-n}(n+i) \) ways to choose the positions for an insertion history. Once the insertion positions are chosen, the insertions clearly appear in some order in the resulting sequence. So for each history of insertion positions, we can assign a length \( t \) sequence to the ordered insertion positions in the resulting sequence. There are \( I_t(n, q) \Pi_{i=1}^{t-n}(n+i) \) total insertion histories, where \( Y \) is a subsequence of the insertion pattern.

So, the number of \( t \) insertion histories for \( X \) giving \( t \)-supersequences in \( I_t(Y) \) is lower bounded by \( \sum_{i=0}^{t-n} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i) \).

The pair \( X = 0_n \) and \( Y = 1_n \) achieves this upper bound because \( 0_n \) and \( 1_n \) have no elements in common, so the only elements in a \( t \)-supersequence of \( X \) that form a copy of \( Y \) must be inserted elements. By the same logic, any pair at Levenshtein distance \( 2n \) achieves the upper bound.

Theorem 4. For \( n, t \in \mathbb{N} \), we have that
\[
B(n, q, \text{UC}_t) = \sum_{i=t-n+1}^{t} \binom{i}{t}(q-1)^i \left( q^t \right)^{-1}.
\]

Proof: There must be another codeword \( c' \). The number of \( t \) insertion histories for \( c \) giving \( t \)-supersequences in \( I_t(c') \) is lower bounded by \( \sum_{i=0}^{t-n} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i) \). There are \( q^t \Pi_{i=1}^{t-n}(n+i) \) insertion histories total. Thus,
\[
f_{\text{UC}_t}(c, C) \leq 1 - \frac{\sum_{i=0}^{t-n} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i)}{q^t \Pi_{i=1}^{t-n}(n+i)} = \sum_{i=t-n+1}^{t} \binom{i}{t}(q-1)^i \Pi_{i=1}^{t-n}(n+i) \left( q^t \right)^{-1}.
\]
where the second line follows from the binomial theorem.
Lemma 6. $f_{UIC}(0_n \{0_n, 1_n\})$ achieves the upper bound as shown in the proof of Theorem 3, so the upper bound equals $B(n, q, UIC_t)$.

$B(n, q, UIC_t)$ is also an upper bound on $U_{UIC}(C)$ and $W_{UIC}(C)$. The following two lemmas give the limiting behavior of the tight upper bound as $t$ and $n$ increase. The behavior in the lemmas is exemplified in Figure 2. Interestingly the bound in Figure 2 decays much slower than the bound for USC in Figure 1.

Lemma 5. For $n, t \in \mathbb{N}$, we have $\lim_{t \to \infty} B(n, q, UIC_t) = 0$.

Proof: We have

$$\lim_{t \to \infty} B(n, q, UIC_t) = \lim_{t \to \infty} \frac{\sum_{i=0}^{t-n+1} \binom{t}{i} (q-1)^i}{q^t} = 0.$$ 

because there are a finite number of terms in the numerator, each of which grows slower than $q^t$.

Lemma 6. For $n \in \mathbb{N}$ and constant $b \in \mathbb{N}$, we have $\lim_{n \to \infty} B(n, q, UIC_{n+b}) = 1$.

Proof:

$$\lim_{n \to \infty} B(n, q, UIC_{n+b}) = \lim_{n \to \infty} \frac{\sum_{i=0}^{n+b} \binom{n+b}{i} (q-1)^i}{q^{n+b}}$$

$$= \lim_{n \to \infty} 1 - \frac{\sum_{i=0}^{n+b} \binom{n+b}{i} (q-1)^i}{q^{n+b}}$$

$$= 1 - \lim_{n \to \infty} \frac{\sum_{i=0}^{n} \binom{n}{i} (q-1)^i}{q^{n+b}} = 1$$

The second line follows from the binomial theorem, and the last equality follows because there are a constant number of terms in the numerator that each grow slower than the denominator.

In addition to showing upper bounds on the probabilities of unique decoding, we can show that for both insertion channels and every code $C$, there exists some codeword where the probability of unique decoding is non-zero for all $t$. This proves the positivity of $U_K(C)$ for all codes $C$, for both channels of interest.

Lemma 7. For every code $C \subseteq \mathbb{F}_q^n$, and $n \in \mathbb{N}$ there exists a codeword $c \in C$ such that for all $t \in \mathbb{N}$, $c$ has at least one unique $t$-supersequence.

Proof: Consider a codeword $c$ with maximal value of $r_c(c)$, where $r_c(c)$ is the length of the right-most run in $c$. Suppose that for some $t \in \mathbb{N}$, $c$ did not have at least one unique $t$-supersequence. Let $a$ be the substring of $c$ to the left of $c$’s rightmost run. Consider the supersequence $s$ formed by duplicating the right-most element $t$ times. Then one or more length $n$ subsequences of $s$ must be codewords in $C$ that are distinct from $c$. Any such codeword $c'$ must be a length $n$ subsequence of $s$. $c'$ therefore must be of the form $xy$ where $x$ is a strict subsequence of $a$, and $y$ is the rightmost element of $c$ repeated $n-|x|$ times. $x$ must be a strict subsequence of $a$, because if it was not, then $c'$ would be the same as $c$. Since $x$ is a strict subsequence, $y$ must be of length greater than $r_c(c)$. Thus, $c'$ has a rightmost run of length greater than $r_c(c)$. This is a contradiction.

Theorem 5. For every code $C \subseteq \mathbb{F}_q^n$, and $n, t \in \mathbb{N}$, we have that $U_K(C)$ is strictly positive for $K \in \{USC_t, UIC_t\}$.

B. Bounds with Additional Assumptions

In this subsection, we find upper bounds for the probability of unique decoding when additional assumptions are made about the code and the codeword. All of the results here are applicable to VT-codes of even length with $a = 0$. Recall that VT-codes with $a = 0$ have maximal cardinality over all choice of $a$, and are asymptotically optimal single insertion/deletion correcting codes.

Lemma 8. For any binary code $C \subseteq \mathbb{F}_2^n$ with $n \in \mathbb{N}$, such that $|C| \geq 3$ and $\{0_n, 1_n\} \subset C$, we have that $f_{K}(c, C) = 0$ for any $c \not\in \{0_n, 1_n\}$, for $K \in \{USC_t, UIC_t\}$ and $t \geq n - 1, t \in \mathbb{N}$.
For any codeword $c \notin \{0_n, 1_n\}$, consider $I_{n-1}(c)$. Suppose $c$ is composed of $x_1$ 1’s and $x_0$ 0’s. Observe that $1 \leq x_0, x_1 \leq n - 1$.

By proving the statement for $t = n - 1$, we prove the statement for $t \geq n - 1$. Consider a sequence $s$ in $I_{n-1}(c)$, and let $s_1$ be the number of ones in $s$, and $s_0$ be the number of zeros in $s$.

If $\geq n - x_1$ ones are inserted, then $s$ is a subsequence of $1_n$. If $n - x_1$ ones are inserted, then $n - 1 - (n - x_1) = x_1 - 1$ zeros are inserted, and thus $s_0 > x_0 + x_1 - 1$. So, $s$ is a subsequence of $0_n$.

**Corollary 2.** For any binary code $C \subseteq F_2^n$ with $n \in \mathbb{N}$, such that $|C| \geq 3$ and $\{0_n, 1_n\} \subseteq C$, we have that $W_K(C) = 0$ for $K \in \{USC_t, UIC_t\}$ and $t \geq n - 1, t \in \mathbb{N}$.

**Proof:** There are $|C| - 2$ codewords such that $f_{USC_t}(c, C) = 0$. Therefore, at least one codeword $c \notin \{0_n, 1_n\}$ has all of its $t$-supersequences in $I_t(0_n) \cup I_t(1_n)$.

Thus, the number of non-unique $t$-supersequences in $I_t(0_n) \cup I_t(1_n)$ is $\leq 2I_t(n, q) - I_t(n, q) = I_t(n, q)$.

$VT_0(n)$ codes of even $n$ always contain $0_n$ and $1_n$. So the above results apply.

With knowledge of a codeword’s weight, we are able to find upper bounds for UIC$_t$ that drop below before $t = n$. These results are also applicable to VT-codes.

**Lemma 9.** For any binary code $C \subseteq F_2^n$ with $n \in \mathbb{N}$, such that $|C| \geq 2$ and $\{0_n, 1_n\} \subseteq C$, we have that for any codeword $c \in C$ with weight $1 \leq w \leq n - 1$, it follows that $f_{UIC_t}(c, C) \leq 1 - \frac{\sum_{i=n-w}^{n-1}(i)}{2^n \sum_{i=w}^{n-1}(i)}$ for $1 \leq t \leq n - 1, t \in \mathbb{N}$.

**Proof:** There are $\sum_{i=n-w}^{n-1}(i)\Pi^t_{j=1}(n + j)$ $t$-insertion histories for $c$ that give sequences in $I_t(1_n)$. Call this set of insertion histories $H_1$.

There are $\sum_{i=w}^{n-1}(i)\Pi^t_{j=1}(n + j)$ $t$-insertion histories for $c$ that give sequences in $I_t(0_n)$. Call this set of insertion histories $H_2$.

Each history in $H_1$ consist of $\geq n - w$ one insertions, and thus $t - (n - w)$ zero insertions. For $t \leq n - 1$, this implies that $\leq w - 1$ zero insertions are made. Each history in $H_2$ consists of $\geq w$ zero insertions, and thus, $H_1 \cap H_2 = 0$.

There are $2^t\Pi^t_{j=1}(n + j)$ insertion histories total, so the result follows.

**Lemma 10.** For any code $C \subseteq F_2^n$ with $n \in \mathbb{N}$, such that $|C| \geq 2$ and $0_n \in C$, we have that for any codeword $c \in C$ with weight $w \geq 1$, it follows that $f_{UIC_t}(c, C) \leq \frac{\sum_{i=0}^{w-1}(i)(q - 1)^{t-i}}{q^t}$ for $t \in \mathbb{N}$.

**Proof:** There are $\sum_{i=0}^{w-1}(i)(q - 1)^{t-i}\Pi^t_{j=1}(n + j)$ $t$-insertion histories for $c$ that give sequences not in $I_t(0_n)$. This is because only $t$-supersequences containing $\leq w - 1$ inserted zeros are not in $I_t(0_n)$. For each number of inserted zeros $i$, there are $\Pi^t_{j=1}(n + j)$ ways to choose the insertion position history. Given the $t$ and the insertion position history, there are $\sum_{i=0}^{w-1}(i)(q - 1)^{t-i}$ ways to choose which insertions are zero insertions, and $(q - 1)^{t-i}$ ways to choose the elements for the remaining $t-i$ insertions.

There are $q^t\Pi^t_{j=1}(n + j)$ insertion histories total, so the result follows.

**Lemma 9** applies to VT$_0(n)$ of even length. Lemma 10 holds for any linear code since all linear codes contain $0_n$. Lemma 10 also holds for VT$_0(n)$ of any length. Provided there is a codeword $c \in C$ of weight $w$, the corresponding bounds in Lemmas 9 and 10 clearly serve as upper bounds on $W_{UIC_t}(C)$.

In Figure 3, we plot the bound in Lemma 9 and the probability of unique decoding for the codewords in VT$_0(6)$ for the uniform insertion channel. The plot also exemplifies the behavior at $t = n - 1$ presented in Lemma 8.

**Fig. 3.** Note: The red curves are underneath the purple curves. The blue curve is underneath the black curve.

**IV. BOUNDS FOR DELETIONS**

In this section, we consider the uniform $t$-deletion channel UDC$_t$, which is defined in Section II. Specifically, we derive $B_d(n, q, UDC_t)$, and derive improved upper bounds with additional assumptions.

Recall that UDC$_t$ incurs each $t$-deletion history with equal probability. As an example, if the word 001100 is affected by the 2-deletion history $[0, 1]$, the received word would be 0100.

Clearly, without any assumptions other than $|C| \geq 2$, the tight upper bound on the probability of unique decoding is 1 if $t < n$, and 0 if $t = n$, for both the uniform $t$-deletion channel and the uniform $t$-subsequence channel. This can be seen by taking the code $C = \{0_n, 1_n\}$.

Our main result for this section is the derivation of $B_d(n, q, UDC_t)$, a tight upper bound on the probability of unique decoding for the uniform $t$-deletion channel, for a codeword $c \in C$, when there is guaranteed to be a codeword $c' \in C$ such that $d_L(c, c') = 2d$ for $1 \leq d \leq n$.

To accomplish this, we must first prove a bijection between the set of all $t$-deletion histories to the set of all $t$-deletion patterns. Here, a $t$-deletion pattern is a temporal list of indices
Lemma 11. There exists a bijection between the set of t-deletion patterns and the set of t-deletion histories.

Proof: We will show there is a bijective mapping from the set of t-deletion histories to the set of t-deletion patterns. Given a t-deletion history, we map it to the resulting t-insertion pattern that occurs.

Any t-deletion pattern clearly has a t-deletion history that maps to it. Thus the mapping is onto.

Now consider the two distinct t-deletion histories. Consider the first position k where they differ. After the first k − 1 deletions, the two deletion histories give the same sequence (with same original codeword elements). Since the kth deletion is different, a different original element from the codeword is deleted from the word. Thus the corresponding deletion patterns differ in the kth element. Thus, the mapping is also one-to-one, and is thus a bijection.

We now give the central result for the uniform t-deletion channel.

Theorem 6. For n, t ∈ N and 1 ≤ d ≤ n, we have that

\[ B_d(n, q, UDC_t) = \begin{cases} \frac{1}{n!} & \text{for } 1 \leq t < d \\ \frac{d!t(t-d)\binom{n-d}{d}(n-t)}{n!} & \text{for } d \leq t \leq n \end{cases} \]

Proof: The total number of t-insertion histories is equal to \( \frac{n!}{(n-t)!} \). The formula works when \( t = n \) because \( 0! = 1 \).

We proceed to work with the equivalent t-deletion pattern definition to calculate the numerator in the bound.

Since the Levenshtein distance between c and c′ is 2d, we know that there exists some set S of d positions in c, such that if we delete the elements at those positions from c, the resulting word will be a d-subsequence of c′. We begin by counting the number of ways to select the d steps in the t-deletion pattern corresponding to the deletions at indices in S. There are \( \binom{d}{t} \) ways to choose these d steps in the deletion pattern. Fixing those steps in the deletion pattern, there are \( d! \) ways to order the deletions. There are then \( \binom{n-d}{t-d} \) ways to choose the remaining indices corresponding to elements in c that are deleted. Once these are chosen, there are \( (t-d)! \) ways to order these deletions. Thus, there are at least \( d!(t-d)!\binom{n-d}{t-d} \) insertion patterns that give subsequences in \( D_t(c) \cap D_t(c') \).

This bound is tight because the codeword \( 1_d0_{n-d} \) in the code \( \{1_d0_{(n-d)}, 0_n\} \) clearly achieves the bound since there is only one possible set \( S \) of size \( d \). This is \( S = \{0, 1, \ldots, d-1\} \).

Interestingly, this bound is alphabet free. The result also serves as an upper bound on \( W_{UDC_t}(C) \). Because the upper bound is tight, the asymptotic analysis of a constant number of extra insertions is of interest. This is given in the Lemma 12.

Lemma 12. For constant \( b \in \mathbb{N} \), we have that

\[ \lim_{n \to \infty} B_d(n, q, UDC_{d+b}) = 1. \]

Proof: We have that

\[ \lim_{n \to \infty} B_d(n, q, UDC_{d+b}) = \lim_{n \to \infty} 1 - \frac{d!\binom{n-b}{d}(n-d-b)!}{n!} = 1 - \frac{d!\binom{n-b}{d}(n-d-b)!}{n!} = \frac{d!\binom{n-b}{d}(n-d-b)!}{n!} = \frac{1 - \prod_{i=0}^{d-1} (n - i)}{1}. \]

One of the difficulties in finding an analogous upper bound for the uniform t-subsequence channel lies in the fact that there is no general formula for the size of a t-deletion ball.

In contrast to the result for insertions in Lemma 7 there exist codes where all codewords have 0 unique t-subsequences for \( t \geq \frac{\lceil \log_2 n \rceil}{2} + 1 \) deletions e.g. \( VT_0(6) \). Thus, the average probability of unique decoding for the uniform deletion channel and the uniform subsequence channel is not always positive. This is an example of how the unique decoding of extra deletions differs from the unique decoding of extra insertions.

Finally, we can find upper bounds on the probability of unique decoding for UDC, which make additional assumptions about the code. The following is the analogue of Lemma 9 for deletions.

Lemma 13. For any binary code \( C \subseteq \mathbb{F}_2^n \) with \( n \in \mathbb{N} \), such that \( |C| \geq 2 \) and \( \{0_n, 1_a\} \subset C \), we have that for any
codeword $c \in C$ with weight $1 \leq w \leq n - 1$, it follows that
\[ f_{UDE}(c, C) \leq 1 - A_n^w \left( t, \frac{(n-t)!w!(t-w)!(t-w)}{n!} \right) \]
\[ - A_{n-w}^w \left( t, \frac{(n-t)!(n-w)!(t-n+w)!(t-w)}{n!} \right) \]
for $1 \leq t < n$, $t \in \mathbb{N}$, and
\[ f_{UDE}(c, C) = 0 \]
for $t = n$, where
\[ A_n^w(t, e) = \begin{cases} 0 & \text{for } 1 \leq t < x \\ e & \text{for } x \leq t < n \end{cases} \]

Proof: The number of $t$-deletion histories giving sequences in $|D_t(c) \cap D_t(0_n)|$ is equal to $w!(t-w)!(t-w)$ by the logic used in the Theorem. Similarly, the number of $t$-deletion histories giving sequences in $|D_t(c) \cap D_t(1_n)|$ is equal to $(n-w)!(t-n+w)!t-w$. The first set of insertion histories results in sequences of all 0's. The second set of insertion histories results in sequences of all 1's. Thus, they are disjoint for $1 \leq t < n$, $t \in \mathbb{N}$.

V. Observations

For either insertions or deletions, the most practical value of $t$ to first consider is $\lfloor \frac{d_{min}}{2} \rfloor + 1$ i.e. one greater than the guaranteed number of correctable insertions/deletions. In this section we discuss the observed behavior of various codes for a small number of extra insertions and raise some open questions.

Recall the single insertion correcting VT-code. Given a VT-code of length $n$, it is known that $a = 0$ always gives the highest cardinality, though there are often multiple values of $a$ that give rise to VT-codes of maximum cardinality.

Among $a$ values that give maximum cardinality, we can select the one which maximizes $W_K(VT_a(n))$, or $U_K(VT_a(n))$ for $K \in \{USC_t, UIC_t, UDC_t, UBC_t\}$, for some $t > 1$ of our interest. In the case of $t = 2$ insertions for example, decoding is efficient. Given any algorithm for decoding 1 insertion that outputs a subsequence of the received word (e.g. using the general approach given in [2]), 2 insertions can be decoded by simply running that decoding algorithm on each of the 1-subsequences of the received word and checking whether the decoding is a codeword.

We observed that by varying $a$, $W_K(VT_a(n))$ and $U_K(VT_a(n))$ for $K \in \{USC_2, UIC_2, UDC_2, UBC_2\}$ can change significantly for smaller $n$, making this a practical observation for relatively small $n$. However, as $n$ increases, we have observed that $W_K(VT_a(n))$ and $U_K(VT_a(n))$ approach 0 for $K \in \{USC_2, UIC_2, UDC_2, UBC_2\}$. Through our observations, we ask the following question about the USC_2, UIC_2, UDC_2, and UBC_2.

For an optimal length $n$ code $C$ with minimum Levenshtein distance $d_{min} = \Theta(1)$, is it always true that $\lim_{n \to \infty} U_K(C) = 0$ and $\lim_{n \to \infty} W_K(C) = 0$ for $K \in \{USC_2, UIC_2, UDC_2, UBC_2\}$?

On the other hand, for the code $C = \{00 \ldots 0, 00 \ldots 11 \ldots 1, 11 \ldots 00 \ldots 0, 11 \ldots 1\}$, we proved that $\lim_{n \to \infty} U_K(C) = 1$ and $\lim_{n \to \infty} W_K(C) = 1$ for $K \in \{USC_2, UIC_2\}$. The code $C = \{0_n, 1_n\}$ is an example where for $\lim_{n \to \infty} U_K(C) = 1$ and $\lim_{n \to \infty} W_K(C) = 1$ for $K = USC_2, UIC_2$ for constant $b \in \mathbb{N}$ (see Lemmas 4 and 5). Through these observations, we ask the following question.

For a length $n$ code $C$ with minimum Levenshtein distance $d_{min} = \Theta(n)$, is it always true that $\lim_{n \to \infty} U_K(C) = 1$ and $\lim_{n \to \infty} W_K(C) = 1$ for $K \in \{USC_2, UIC_2\}$ and constant $b \in \mathbb{N}$?

We are also interested in the analogues of these questions when $d_{min}$ and $b$ are different functions of $n$.

VI. Conclusion

In this work, we examined the unique decoding capability of codes from insertions and deletions beyond the guaranteed number of correctable insertions/deletions. We defined several probabilistic channels as a framework to understand what happens when $t$ insertions or $t$ deletions occur, regardless of whether the real-life channel always makes $t$ insertions or $t$ deletions. We computed tight upper bounds on the probability of unique decoding for the channels, and studied the limiting behavior of the bounds. We then found upper bounds specific to particular classes of codes, such as linear codes and VT-codes. We also studied the positivity of the measures, and raised several open questions.

References

[1] R. R. Varshamov and G. M. Tenengolts, “Codes which correct single asymmetric errors (in russian),” Automatika i Telemkhania, vol. 161, no. 3, pp. 288-292, 1965.
[2] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals” (in Russian), Doklady Akademii Nauk SSSR, 163 (4, 1965), 845-848. English translation in Soviet Physics Dokl., 10 (No. 8, 1966), 770-711.
[3] V.I. Levenshtein, “Efficient reconstruction of sequences,” Trans. Info. Theory, vol. 47, no. 1, pp. 2-22, Jan. 2001.
[4] V.I. Levenshtein, “Efficient reconstruction of sequences from their subsequences or supersequences,” Journal of Combinatorial Theory, vol. 93, no. 2, pp. 310-332, 2001.
[5] V. Guruswami and M. Sudan, “Improved decoding of Reed-Solomon and algebraic-geometric codes,” in FOCS ’98, IEEE Computer Society (1998), pp. 28-39.
[6] N.J.A. Sloane, “On single-deletion-correcting codes,” Codes and Designs: Proceedings of a Conference Honoring Professor Dijen K. Ray-Chaudhuri on the Occasion of His 65th Birthday, Ohio State University, May 18-21, 2000, 2002.
[7] P. A. Bours, “Codes for correct insertion and deletion errors,” PhD thesis, Eindhoven University of Technology, Jun. 1994.
[8] M. Mitzenmacher, “A survey of results for deletion channels and related synchronization channels,” Probability Surveys, pp. 1-33, June 2009.
[9] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, “Exact reconstruction from insertions in synchronization codes,” IEEE Trans. Inform. Theory, vol. 63, no. 4, pp. 2428-2445, Apr. 2017.
[10] F. Sala, R. Gabrys, L. Dolecek, “Gilbert-Varshamov-like Lower Bounds for Deletion-Correcting Codes,” IEEE Information Theory Workshop (ITW) 2014.
observe that for a sequence to be in the intersection of $I_{k-n_1}(n_1, q - 1)$. Let $A_k(T)$ be the set of all sequences formed by inserting $k$ ones into a sequence $T$. Consider $T_1, T_2 \in S_{0_{n_1}}, T_1 \neq T_2$. Then clearly $|A_k(T_1) \cap A_k(T_2)| = 0$. To count the number of sequences in $A_k(T_1)$, we can view each of the $k + 1$ spaces adjacent to sequence elements in $T_1$ as a labeled bin, and the $N - k$ ones that will be inserted into the sequence as $N - k$ unlabeled balls that are being tossed into the $k + 1$ bins.

Using the formula, for unlabeled balls tossed into labeled bins, we see that $|A_k(T_1)| = (N-k+1)^{k+1} = \binom{N}{k} = \left(\frac{n_2 + t_2}{k}\right)$.

Thus, the number of supersequences in $I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})$ with $N - k$ ones is equal to $I_{k-n_1}(n_1, q - 1)$. Summing over the possible values of $k$, we obtain

$$|I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})| = \sum_{k=1}^{n_2} I_{k-n_1}(n_1, q - 1) \binom{n_2 + t_2}{k}$$

We proceed to show that

$$\min_{X \in F^{n_1}_q, Y \in F^{n_2}_q} |I_{t_1}(X) \cap I_{t_2}(Y)| = |I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})|$$

by induction on $N = n_1 + t_1 = n_2 + t_2$.

The property trivially holds for $N = 0$. For $N = 1$, $|I_0(0) \cap I_0(1)| = 0$ while in any other case, the cardinality must be $\geq 0$ since set cardinalities are always non-negative.

Suppose the property holds for all $N$ in the range $0 \leq N \leq T$. For $N = T + 1$, we consider the pair of sequences $X = 0_{n_1}$ and $Y = 1_{n_2}$ for any $n_1, n_2, t_1, t_2 \in \mathbb{N}$ such that $t_1 + n_1 = t_2 + n_2 = T + 1$. From Lemma 1 we see that

$$|I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})| = |I_{t_1}(0_{n_1-1}) \cap I_{t_2}(1_{n_2-1})| + |I_{t_1-1}(0_{n_1}) \cap I_{t_2}(1_{n_2-1})| + (q-2)|I_{t_1-1}(0_{n_1-1}) \cap I_{t_2-1}(1_{n_2-1})|.$$  

Now consider two arbitrary sequences $X = aX' \in F^{n_1}_q$ and $Y = bY' \in F^{n_2}_q$. We see from Lemma 1 that if $a \neq b$, then

$$|I_{t_1}(X) \cap I_{t_2}(Y)| = |I_{t_1}(X') \cap I_{t_2-1}(bY')| + |I_{t_1-1}(aX') \cap I_{t_2}(Y')| + (q-2)|I_{t_1-1}(aX') \cap I_{t_2-1}(bY')|.$$  

By the inductive hypothesis,

$$|I_{t_1}(0_{n_1-1}) \cap I_{t_2-1}(1_{n_2})| \leq |I_{t_1}(X') \cap I_{t_2-1}(bY')|.$$
and
\[ |I_t(0_{n_1}) \cap I_t(1_{n_2-1})| \leq |I_t(aX') \cap I_t(Y')| \]

and
\[ |I_{t-1}(0_{n_2}) \cap I_{t-1}(1_{n_2})| \leq |I_{t-1}(aX') \cap I_{t-1}(bY')| \]

Thus, if \( a \neq b \),
\[ |I_t(0_{n_1}) \cap I_t(1_{n_2})| \leq |I_t(X) \cap I_t(Y)|. \]

Now consider the case where \( a = b \). In this case, we have
\[ |I_t(X) \cap I_t(Y)| = |I_t(X') \cap I_t(Y')| + (q-1)|I_{t-1}(aX') \cap I_{t-1}(bY')|. \]

By the inductive hypothesis, we have
\[ |I_{t-1}(0_{n_1-1}) \cap I_{t-1}(1_{n_2-1})| \leq |I_{t-1}(X') \cap I_{t-1}(Y')| \]

and
\[ |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \leq |I_{t-1}(aX') \cap I_{t-1}(bY')| \]

We thus have that
\[
|I_t(X) \cap I_t(Y)| \\
\geq |I_t(0_{n_1-1}) \cap I_t(1_{n_2-1})| \\
+ (q-1)|I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})|
\]

We will complete the proof by showing that
\[
|I_t(0_{n_1-1}) \cap I_t(1_{n_2-1})| + (q-1)|I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
= |I_t(0_{n_1-1}) \cap I_t(1_{n_2-1})| + |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
+ (q-2)|I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
\geq |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| + |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
+ (q-2)|I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
= |I_t(0_{n_1}) \cap I_t(1_{n_2})|
\]

To show the above, it suffices to prove that
\[
|I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| + |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| \\
\geq |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})| + |I_{t-1}(0_{n_1}) \cap I_{t-1}(1_{n_2})|
\]

We begin by manipulating the left side of the inequality.
Thus, if \( a = b \), we have
\[
|I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})| \leq |I_{t_1}(X) \cap I_{t_2}(Y)|
\]
and we have proved the inductive step.

**B. Proof of Lemma 2**

Lemma 2 is restated below, and follows directly from the proceeding auxiliary lemmas along with Theorem 1.

**Lemma 2** Let \( n_1, n_2, t_1, t_2 \in \mathbb{N} \) and \( N = t_1 + n_1 = t_2 + n_2 \). Then for any length \( n_1 \) sequence \( X \) and any length \( n_2 \) sequence \( Y \) such that \( d_L(X, Y) = n_1 + n_2 \), we have that
\[
|I_{t_1}(X) \cap I_{t_2}(Y)| = \hat{N}^+_q(n_{1, n_2, t_1, t_2}) = \sum_{k=n_1}^{t_2} \sum_{i=0}^{n_2} \binom{k}{i} (q - 2)^i \binom{n_2 + t_2 - 1}{k}.
\]

**Lemma 14.** Let \( n_1, n_2, t_1, t_2 \in \mathbb{N} \) and \( N = t_1 + n_1 = t_2 + n_2 \). Then for any length \( n_1 \) sequence \( X \) and any length \( n_2 \) sequence \( Y \) such that \( d_L(X, Y) = n_1 + n_2 \), we have that \( Y \) only contains elements from some subset \( B \) of the alphabet \( \mathbb{F}_q \), and \( X \) only contains elements from the subset \( \mathbb{F}_q \setminus B \). Letting \( q_B = |B| \), we have that
\[
|I_{t_1}(X) \cap I_{t_2}(Y)| = \sum_{k=0}^{N-n_1-n_2} I_k(n_{1, q - q_B}) I_{N-k-n_1-n_2}(n_{2, q_B}) (N_{t_1 - k}).
\]

**Proof:** Clearly, if \( d_L(X, Y) = n_1 + n_2 \), we have that \( X \) and \( Y \) do not have any elements in common. Thus we consider the set of elements in \( Y \) as a new alphabet \( B \), which is clearly a subset of the alphabet \( \mathbb{F}_q \). We then consider the subset \( \mathbb{F}_q \setminus B \) as the alphabet for \( X \).

To count the number of sequences in \( I_{t_1}(X) \cap I_{t_2}(Y) \), we partition the sequences based on the number \( n_1+k \) of elements from \( \mathbb{F}_q \setminus B \) that the sequence contains. Given some value of \( k \), there are \( I_k(n_{1, q - q_B}) \) possibilities for the sequence’s subsequences of elements from \( \mathbb{F}_q \setminus B \).

A sequence in \( I_{t_1}(X) \cap I_{t_2}(Y) \) with \( k \) additional elements from \( \mathbb{F}_q \setminus B \) must have \( N - n_1 - k \) elements from \( B \). Given \( k \), there are \( I_{N-k-n_1-n_2}(n_{2, q_B}) \) possibilities for the sequence’s subsequences of elements from \( B \).

Given the sequence’s length \( n_1+k+1 \) subsequence \( T_X \) of elements from \( \mathbb{F}_q \setminus B \) and the sequence’s length \( N - n_1 - k \) subsequence \( T_Y \) of elements from \( B \), the number of possible sequences containing both of these subsequences is the number of ways \( T_Y \) can be interleaved with \( T_X \). This can be counted using a balls and bins approach.

We consider the \( n_1+k+1 \) possible insertion positions into \( T_X \) as labeled bins, and the \( N - k - n_1 \) positions of elements from \( T_Y \) as unlabeled balls. There are thus \( \binom{n_1+k+1}{N-k-n_1} \) ways for \( T_Y \) to be interleaved with \( T_X \).
$k$ can range from 0 to $N - n_1 - n_2$ so we have

$$|I_{t_1}(X) \cap I_{t_2}(Y)|$$

$$= \sum_{k=0}^{N-n_1-n_2} I_k(n_1, q - q_B) I_{N-k-n_1-n_2}(n_2, q_B) \binom{N}{t_1 - k}.$$ 

Lemma 15. Let $n_1, n_2, t_1, t_2 \in \mathbb{N}$ and $N = t_1 + n_1 = t_2 + n_2$. Consider any subset $B \subseteq \mathbb{F}_q$ such that $1 \leq q_B \leq q - 1$, where $q_B = |B|$. Then,

$$|I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})|$$

$$= \sum_{k=0}^{N-n_1-n_2} I_k(n_1, q - q_B) I_{N-k-n_1-n_2}(n_2, q_B) \binom{N}{t_1 - k}.$$ 

Proof: Using a similar approach as in Lemma 14 we will choose a sub-alphabet for the sequence $1_{n_2}$, and a sub-alphabet for $0_{n_1}$.

$0_{n_1}$ and $1_{n_2}$ have no elements in common. So, the alphabet $A$ for $1_{n_2}$ can be chosen to be any subset of $\mathbb{F}_q$ such that $1 \in A$, and $0 \notin A$, and $|A| = q_B$. Once the alphabet for $1_{n_2}$ is chosen, the alphabet for $0_{n_1}$ will be $\mathbb{F}_q \setminus A$. Thus, we have that

$$|I_{t_1}(0_{n_1}) \cap I_{t_2}(1_{n_2})|$$

$$= \sum_{k=0}^{N-n_1-n_2} I_k(n_1, q - q_B) I_{N-k-n_1-n_2}(n_2, q_B) \binom{N}{t_1 - k}.$$ 

by following the proof in Lemma 14. 

■