ISIT'2013 1569717621

Counting Sequences Obtained From the Synchronization Channel

Frederic Sala
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, California 90095
fredsala@ucla.edu

Lara Dolecek
Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, California 90095
dolecek@ee.ucla.edu

***Abstract*—Synchronization channels, which can remove codeword symbols or introduce extraneous symbols, pose additional difficulties when compared to the commonly-studied substitution channel. A traditional problem in this area is to count the number of sequences formed when deleting a fixed number of symbols from a sequence. This work contains our first effort towards solving a similar, yet previously unexplored, problem: deriving bounds on the number of sequences obtained by deleting *and* inserting a fixed number of symbols.**

I. Introduction

Synchronization errors occur when codeword symbols are lost or extraneous symbols are added during transmission. A natural question regarding the synchronization channel is how many unique sequences are produced when deleting (or inserting) some number of symbols from an original sequence. Existing works, starting with those of Levenshtein, [4], produce bounds for sequences formed by deletion of symbols.

In this paper, we seek to generalize previous results by counting the number of sequences formed from deleting *and* inserting a constant number of symbols into a sequence. In particular, we establish an expression for the number of sequences formed from one deletion and one insertion. We then proceed to construct bounds for the general case. Our approach integrates previous results counting sequences formed by deletions *or* insertions.

In [4], it was shown that codes that correct k deletions also correct any combination of j insertions and $k-j$ deletions (for $0 \leq j \leq k$). When estimating the rate of a k-synchronization error-correction code, the use of k-deletion spheres yields better bounds than the use of k-insertion spheres [2]. It is interesting to investigate whether using j insertion and $(k-j)$ deletion spheres provides bounds of intermediate strength. This idea motivates our first effort to count the number of sequences formed by j insertions and $k-j$ deletions.

The rest of this paper is organized as follows. In Section II, we present some useful notation. In Section III, we quote results from previous works on counting subsequences, which we use throughout the paper. In Section IV, we introduce our main technique and use it to derive an exact expression for the number of sequences formed from a single insertion and deletion. In Section V, we find bounds in the general case. We summarize our contributions in Section VI.

II. Notation

We begin with some notation. Let F_q be the set $\{0, 1, \ldots, q-1\}$ for $q \geq 2$. Consider a sequence (or a string) $X = (x_1, x_2, \ldots, x_n) \in F_q^n$. Define $[n] = \{1, 2, \ldots, n\}$. We call $Y \in F_q^m$ a **subsequence** of X if there exist $i_1 \leq i_2 \leq \ldots \leq i_m \in [n]$ such that $Y = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ for some $m \leq n$. Equivalently, $Y \in F_q^m$ is a subsequence of X if it can be formed by deleting $n-m$ symbols from X.

Similarly, we can define a **supersequence**: $Z \in F_q^r$ is a supersequence of X if $n \leq r$ and X is a subsequence of Z. Again, we may view Z as being formed by inserting $r-n$ symbols from F_q into X.

Next, we define $D_t(X)$ as the set of subsequences of X of length $n-t$, that is, the set of sequences formed by deleting t symbols from X. Similarly, we let $I_v(X)$ be the set of sequences resulting from the insertion of v symbols into X. More generally, we let $E_{t,v}(X)$ be the set of sequences formed by deleting t symbols from and inserting v symbols into X.

The number of (maximal) runs of identical symbols in a sequence X is written $\tau(X)$, with $1 \leq \tau(X) \leq n$. If $X_1 = (1011)$ and $X_2 = (0120)$, $\tau(X_1) = 3$ and $\tau(X_2) = 4$.

We define the **cyclic string** C_n, a sequence of length n with n runs, as follows: the ith component $(C_n)_i = a \in F_q$ if $i - 1 - q \equiv a \pmod q$. For example, if $q = 3$ and $n = 7$, $C_7 = (0120120)$.

Lastly, we introduce the notion of an **"ancestor" distance** $d_A(X, Y)$. We say $d_A(X, Y) = j$ if there exists a sequence Z such that $X, Y \in D_j(Z)$, but there is no sequence W and no $i < j$ such that $X, Y \in D_i(W)$. In other words, $d_A(X, Y)$ measures the number of insertions required to form the shortest common supersequence, or ancestor of X and Y. Note that d_A is exactly half of the traditional "edit" distance, which is defined as the number of insertions and deletions required to transform string X into string Y.

As an example, if $X = (10120)$ and $Y = (21112)$, then $d_A(X, Y) = 2$, since $X, Y \in D_2(2101120)$, while there is no string W such that $X, Y \in D_1(W)$.

III. Prior work

First, we note that the number of subsequences $|D_t(X)|$ depends on the structure of X. For example, $|D_1(X)| = \tau(X)$. Using a simple combinatorial argument in [4], Levenshtein

found the earliest bounds on the number of sequences formed by deleting t symbols from X:

$$\binom{\tau(X)-t+1}{t} \leq |D_t(X)| \leq \binom{\tau(X)+t-1}{t}. \quad (1)$$

Calabi and Hartnett, [1], found a better upper bound by examining the sequence which maximizes $|D_t(X)|$. They found that $C_n = \arg\max_{X \in F_q^n} |D_t(X)|$, and gave a generating function for $|D_t(C_n)|$. Building on this work, Hirschberg and Regnier discovered a recursive expression for $|D_t(C_n)|$ in terms of q, [3]. Let $d_q(t,n)$ represent $|D_t(C_n)|$, where $C_n \in F_q^n$. Then, for $0 \leq t \leq n$ and $q \geq 2$,

$$d_q(t,n) = \sum_{i=0}^{t} \binom{n-t}{i} d_{q-1}(t-i,t). \quad (2)$$

This formula yields the simple expression $\sum_{i=0}^{t} \binom{n-t}{i}$ in the binary case $q = 2$. Hirschberg and Regnier also improved the lower bound, producing the tight binary bound,

$$\sum_{i=0}^{t} \binom{\tau(X)-t}{i} \leq |D_t(X)| \leq \sum_{i=0}^{t} \binom{n-t}{i}, \quad (3)$$

which was improved by Liron and Langberg in [5].

It is difficult to provide an exact expression for $|D_t(X)|$. If we are limited to knowledge of $\tau(X)$, the bounds above are the best known. If we know the length of each of the runs $r_1, r_2, \ldots, r_{\tau(X)}$, it is possible to develop a combinatorial expression for $|D_t(X)|$.

On the other hand, $|I_t(X)|$ does not depend on X. Inserting t symbols always yields the same number of supersequences. For this reason, we will refer to $|I_t(X)|$ as $I_q(n,t)$. In [4], Levenshtein showed that

$$I_q(n,t) = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i. \quad (4)$$

The preceding results will form a rough starting point for our computation of $|E_{t,v}(X)|$, which we will progressively refine.

IV. Computing sequences formed by one insertion and one deletion

We begin by examining the simplest case: $t = v = 1$. This case is particularly tractable, since we have the exact expressions $|D_1(X)| = \tau(X)$ and

$$|I_1(X)| = I_q(n,1) = 1 + (n+1)(q-1).$$

If we assume that no two sequences (of length $n-1$) in $D_1(X)$ yield any common supersequences after an insertion, the size of $E_{1,1}(X)$ is $|D_1(X)| I_q(n-1,1)$, so that

$$|E_{1,1}(X)| \leq \tau(X)(1 + n(q-1)).$$

It remains to check how many times we double counted, that is, how many sequences $Y \in E_{1,1}(X)$ and $X_1, X_2, \ldots, X_k \in D_1(X)$ are such that $Y \in I_1(X_1) \cap \ldots \cap I_1(X_k)$ for $k \geq 2$.

We introduce some additional notation before we state our result. Let a (maximal) **alternating segment** in X be a sequence of elements $x_i x_{i+1} \ldots x_{j-1} x_j = \alpha\beta\alpha\beta\ldots\beta\alpha$ or $\alpha\beta\alpha\beta\ldots\alpha\beta$, which alternate between two symbols $\alpha, \beta \in F_q$ with $\alpha \neq \beta$. We require that $x_{i-1} \neq \beta$ and $x_{j+1} \neq x_{j-1}$ to ensure that the alternating segment is maximal. Note that each element (except possibly the first and the last) forms a run of length 1 in X. We have the result,

Theorem 1. *The number of sequences formed by one deletion and one insertion from $X \in F_q^n$ with k alternating segments of lengths $s_1, s_2, \ldots, s_k$, $|E_{1,1}(X)|$, satisfies*

$$|E_{1,1}(X)| = \tau(X)(n(q-1)-1) + 2 - \sum_{i=1}^{k} \frac{(s_i-1)(s_i-2)}{2}.$$

Proof: We provide a roadmap for the proof. We begin by estimating $|E_{1,1}(X)|$ as $\tau(X)(1 + n(q-1))$. Here, certain sequences have been double counted, and thus must be subtracted. We will see that such sequences Y result exclusively from an insertion into each of a pair of sequences $X_1, X_2 \in D_1(X)$. We proceed to find each case for such pairs and subtract the appropriate number from the estimate.

Now, every sequence in $D_1(X)$ can give rise to the original sequence X after one symbol insertion. That is, for all $X' \in D_1(X)$, $X \in I_1(X')$. For this reason, X has been counted $|D_1(X)| = \tau(X)$ times, so we must subtract $\tau(X) - 1$ sequences from the estimate, yielding $|E_{1,1}(X)| = \tau(X)(n(q-1)) + 1$.

Next we seek strings other than X that have been counted more than once. Any such sequence must be formed by inserting a symbol into multiple distinct sequences in $D_1(X)$. However, inserting a symbol into three or more strings in $D_1(X)$ results in only one common sequence, X. That is, for $3 \leq j \leq \tau(X)$, $I_1(X_1) \cap I_1(X_2) \cap \ldots \cap I_1(X_j) = \{X\}$. Therefore, we need only search for sequences $Y \neq X$ such that $Y \in I_1(X_1) \cap I_1(X_2)$ for some $X_1, X_2 \in D_1(X)$[1]. If we view X as a sequence of runs $\tau_1, \tau_2, \ldots, \tau_j$, where $j = \tau(X)$, any subsequence in $D_1(X)$ can be formed by removing one symbol from one of the $\tau(X)$ runs.

Consider the case when two subsequences are formed by removing elements found in adjacent runs in X. If α is the symbol in run i and β the symbol in run $i+1$, $1 \leq i \leq \tau(X) - 1$, then, the supersequences $X = (\ldots\alpha\beta\ldots)$ and $(\ldots\beta\alpha\ldots)$ are in $I_1(X_1) \cap I_1(X_2)$. X has been accounted for, but there is now one additional common supersequence. There are $\tau(X) - 1$ adjacent runs, so we must remove an additional $\tau(X) - 1$ overcounted sequences from the estimate, resulting in the expression

$$|E_{1,1}(X)| = \tau(X)(n(q-1)-1) + 2.$$

Finally, we examine pairs of strings X_1, X_2 formed by deletions from non-adjacent runs in X. Write X as $(\ldots\alpha\tau_{i+1}\tau_{i+2}\ldots\tau_{j-1}\beta\ldots)$. Here, α is the element in run i, β is the element in run j, $i+1 \neq j$, the element in run τ_{i+1} is not α, and the element in τ_{j-1}

[1]Levenshtein established that $\max_{X_1,X_2} |I_1(X_1) \cap I_1(X_2)| = 2$ in [6]. However, in our problem, we need to count the number of such pairs (X_1, X_2) with $X_1, X_2 \in D_1(X)$. This is an entirely different problem.

is different from β. Then, removing α and removing β give subsequences $X_1 = (\ldots \tau_{i+1}\tau_{i+2}\ldots\tau_{j-1}\beta\ldots)$ and $X_2 = (\ldots\alpha\tau_{i+1}\tau_{i+2}\ldots\tau_{j-1}\ldots)$. If we insert α into X_1 before τ_{i+1}, we will recover X. If we insert α elsewhere in X_1, we must insert β before α in X_2 and have that $\tau_{i+1} = \beta$. Then, $\tau_{i+2} = \alpha$, and we proceed similarly to find that $(\ldots\alpha\tau_{i+1}\tau_{i+1}\ldots\tau_{j-1}\beta\ldots) = (\ldots\alpha\beta\alpha\beta\ldots\alpha\beta\ldots)$, which is (part of) an alternating segment as previously defined.

X contains k alternating segments with lengths $s_1, s_2, \ldots, s_k$. Adjacent run deletions have been subtracted previously, so we need only count $\binom{s_i}{2} - (s_i - 1) = \frac{1}{2}(s_i - 1)(s_i - 2)$ pairs of non-adjacent choices of runs for a deletion in the alternating segment s_i. Therefore, we must subtract an additional factor of $\sum_{i=1}^{k} \frac{1}{2}(s_i - 1)(s_i - 2)$ from our estimate. All cases having been examined, the proof is concluded. ∎

We comment on the above result. First, note that $|E_{1,1}(X)|$ depends on the structure of X, unlike $|I_t(X)|$. However, like the $t = 1$ and $v = 1$ cases for $|D_t(X)|$ and $|I_v(X)|$, an exact expression is still possible. Unfortunately, some knowledge of the structure of X is required beyond just the number of runs $\tau(X)$. If we know only $\tau(X)$, the best upper bound is

$$|E_{1,1}(X)| \leq \tau(X)(n(q-1)-1)+2.$$

V. Bounds on the Synchronization Channel

In this section, we introduce a framework for computing general bounds on $|E_{t,v}(X)|$. The framework is an application of the principle of inclusion-exclusion. In brief, we first delete t symbols, yielding $|D_t(X)|$ sequences, then insert v symbols into each of these sequences, for a total of $|D_t(X)||I_q(n-t,v)|$. Through inclusion-exclusion, we remove all overcounted sequences from this estimate. We illustrate the technique by deriving a lower bound when $t = v = 2$ and then generalizing the bound to arbitrary $t = v$.

We use the initial estimate $|D_t(X)|I_q(n-t,v)$, although there are other potential choices, such as the expression

$$|E_{t,v}(X)| \leq |D_j(X)||E_{t-j,v-k}(X')|I_q(n-t+v-k,k),$$

where $1 \leq j < t$, $1 \leq k < v$, and $X' \in D_j(X)$. To form the expression, we first delete j symbols, then delete $t-j$ and insert $v-k$ symbols, then finally insert the remaining k symbols. This allows us to take advantage of an expression for $|E_{t-j,v-k}(X)|$. However, since $|D_t(X)| \leq |D_j(X)||D_{t-j}(X)|$ and $I_q(n,v) \leq I_q(n,k)I_q(n+k,v-k)$, the use of $|D_t(X)|I_q(n-t,v)$ will yield tighter bounds. Now we may apply the principle of inclusion-exclusion to write

$$\begin{aligned}
|E_{t,v}(X)| &= |D_t(X)|I_q(n-t,v) \\
&\quad - \sum_{\substack{X_1,X_2 \\ \in D_t(X)}} |I_v(X_1)\cap I_v(X_2)| \\
&\quad + \sum_{\substack{X_1,X_2,X_3 \\ \in D_t(X)}} |I_v(X_1)\cap I_v(X_2)\cap I_v(X_3)| \\
&\quad - \ldots \\
&\quad \pm \sum_{\substack{X_1,\ldots,X_{|D_t(X)|} \\ \in D_t(X)}} |I_v(X_1)\cap\ldots\cap I_v(X_{|D_t(X)|})|,
\end{aligned}$$

that is,

$$\begin{aligned}
|E_{t,v}(X)| &= |D_t(X)|I_q(n-t,v) \\
&\quad + \sum_{k=2}^{|D_t(X)|} (-1)^{k+1} \sum_{\substack{X_1,\ldots,X_k \\ \in D_t(X)}} |I_v(X_1)\cap\ldots\cap I_v(X_k)|.
\end{aligned}$$

Naturally, we seek to compute the sizes of intersections of the type $|I_v(X_1)\cap I_v(X_2)\cap\ldots\cap I_v(X_k)|$, where $X_1, X_2, \ldots, X_k \in D_t(X)$. In the case where $v \geq t$, that is, there at least as many insertions as deletions, every such intersection will trivially contain the sequence X, or one of its supersequences. Therefore, we define

$$F_{t,v}(X,k) = \sum_{\substack{X_1,\ldots,X_k \\ \in D_t(X)}} (|I_v(X_1)\cap\ldots\cap I_v(X_k)|-1),$$

which represents over- (under-) counted sequences which are not X (or one of its supersequences). Now, for $v \geq t$, we may express $|E_{t,v}(X)|$ in terms of the $F(X,k)$'s:

$$\begin{aligned}
|E_{t,v}(X)| &= |D_t(X)|I_q(n-t,v) + \sum_{k=2}^{|D_t(X)|} (-1)^{k+1} F_{t,v}(X,k) \\
&\quad - \binom{|D_t(X)|}{2} + \binom{|D_t(X)|}{3} - \ldots \pm \binom{|D_t(X)|}{|D_t(X)|} \\
&= |D_t(X)|I_q(n-t,v) \\
&\quad + \sum_{k=2}^{|D_t(X)|} (-1)^{k+1} F_{t,v}(X,k) + 1 - |D_t(X)|. \qquad (5)
\end{aligned}$$

We are particularly interested in the case $t = v$ where the same number of symbols are inserted and deleted. We will deal with this scenario throughout the remainder of the paper.

It is difficult to compute the $F(X,k)$'s exactly. Instead, we focus on providing bounds. We note that by cutting off the alternating sum (5) formed by inclusion/exclusion, we can form lower and upper bounds of increasing tightness. We have

Lemma 1. *For $l \in \{1,2,\ldots,\lfloor\frac{|D_t(X)|}{2}\rfloor\}$, let $L_{t,t}(X,l)$ be*

$$|D_t(X)|I_q(n-t,t) + \sum_{k=2}^{2l}(-1)^{k+1}F_{t,t}(X,k) + 1 - |D_t(X)|,$$

and for $i \in \{1,2,\ldots,\lfloor\frac{|D_t(X)|-1}{2}\rfloor\}$, let $U_{t,t}(X,i)$ be defined as

$$|D_t(X)|I_q(n-t,t) + \sum_{k=2}^{2i+1}(-1)^{k+1}F_{t,t}(X,k) + 1 - |D_t(X)|.$$

Then,

$$\begin{aligned}
L_{t,t}(X,1) \leq \ldots \leq L_{t,t}\left(X, \left\lfloor\frac{|D_t(X)|}{2}\right\rfloor\right) \leq |E_{t,t}(X)| \\
\leq U_{t,t}\left(X, \left\lfloor\frac{|D_t(X)|-1}{2}\right\rfloor\right) \leq \ldots \leq U_{t,t}(X,1).
\end{aligned}$$

Next, we demonstrate how to estimate the $F(X,k)$ terms. We have found an expression for $t=v=1$ in Theorem 1, so we proceed with the next smallest case $t=v=2$. We will upper bound $F_{2,2}(X,2)$ to derive the lower bound $L_{2,2}(X,1)$ for $|E_{2,2}(X)|$ in Lemma 1. In the following analysis, we will not consider alternating segments. Although it is possible to generalize the alternating segments analysis to yield more precise bounds, our lower bound "gives away" enough overlaps to make up for the contributions from the alternating segment case. We omit the proof of this fact due to space limitations.

Various possibilities for the intersections $I_2(X_1)\cap I_2(X_2)$ (with $X_1,X_2\in D_2(X)$) inside the sum in $F(X,2)$ are shown in Fig. 1. In each graph, the top level node represents the sequence X, while the next levels show sequences after one deletion, two deletions, one insertion, and two insertions, respectively. We draw an edge between two sequences if one can be formed from the other through one insertion or one deletion. The middle level shows the two subsequences X_1 and X_2, while the bottom level depicts $I_2(X_1)\cap I_2(X_2)$.

For the pair of sequences (X_1,X_2), either $d_A(X_1,X_2)=1$ or $d_A(X_1,X_2)=2$. In the first case, $d_A(X_1,X_2)=1$, we have that $I_1(X_1)\cap I_1(X_2)$ is not empty, so we let Y be in this intersection. But then, $I_1(Y)\subset I_2(X_1)\cap I_2(X_2)$. This case is shown in the first graph in Fig. 1. After eliminating the element X, which is in $I_1(Y)$, X_1 and X_2 still contribute $|I_1(Y)|-1=I_q(n-1,1)-1$ to $F_{2,2}(X,2)$.

We can upper bound the number of pairs (X_1,X_2) with $d_A(X_1,X_2)=1$ by $\tau(X)\binom{\tau(X)}{2}$. This is due to there being $\tau(X)$ parent sequences in $D_1(X)$ for X_1 and X_2, followed by at most $\binom{\tau(X)}{2}$ pairs of distance-1 sequences for each parent sequence. We also note that for some pairs (X_1,X_2) with $d_A(X_1,X_2)=1$, $|I_1(X_1)\cap I_1(X_2)|=2$. This is shown in the second graph in Fig. 1. In this scenario, we gain an additional contribution of $I_q(n-1,1)-1$ per pair. The case occurs when the distinct deletions forming X_1 and X_2 are in adjacent runs in X, so we may upper bound the number of such pairs by $\tau(X)(\tau(X)-1)$. Thus the total contribution to $F(X,2)$ from pairs with $d_A(X_1,X_2)=1$ is

$$\tau(X)\left(\binom{\tau(X)}{2}+\tau(X)-1\right)(I_q(n-1,1)-1)$$

It remains to examine the cases where $d_A(X_1,X_2)=2$. Here we may apply the same principles from the proof of Theorem 1. We again view X as the sequence of runs $\tau_1,\tau_2,\ldots,\tau_j$ where $j=\tau(X)=|D_1(X)|$. Now, if $d_A(X_1,X_2)=2$, the runs deleted symbols come from may not be the same for X_1 and X_2. However, if none of the runs are adjacent, $I_2(X_1)\cap I_2(X_2)=\{X\}$, so there is no contribution.

The only remaining cases are where either one or both of the deleted symbols are in adjacent runs. These cases are shown in the last two graphs of Fig. 1. There is a contribution of 1 in the first case and of 3 in the second case. To see why this is so, consider symbols α,β and γ,ω, each in adjacent runs. By this, we mean that X_1 is formed by deleting α and γ from X, while X_2 is formed by deleting β and ω from

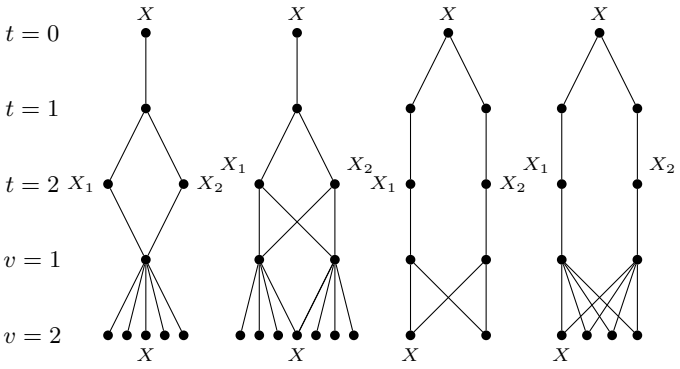

Fig. 1. Contributions to $I_2(X_1)\cap I_2(X_2)$ in the two-deletion, two-insertion case. The levels above represent the original sequence X, sequences after a single deletion ($t=1$), the sequences X_1 and X_2 after 2 deletions ($t=2$), sequences after 2 deletions and 1 insertion ($v=1$), and sequences after 2 deletions and 2 insertions ($v=2$). The bottom level depicts the strings in $I_2(X_1)\cap I_2(X_2)$. In the first two graphs, $d_A(X_1,X_2)=1$, while in the second two graphs, $d_A(X_1,X_2)=2$.

X, where α and β are the symbols in runs p and $p+1$, respectively, and γ and ω are the symbols in runs s and $s+1$, respectively. Here, $1\le p<s\le\tau(X)-1$ and $s\neq p+1$. Then, $I_2(X_1)\cap I_2(X_2)$ contains $X=(\ldots\alpha\beta\ldots\gamma\omega\ldots)$, along with $(\ldots\beta\alpha\ldots\gamma\omega\ldots)$, $(\ldots\beta\alpha\ldots\omega\gamma\ldots)$, and $(\ldots\alpha\beta\ldots\omega\gamma\ldots)$.

The number of pairs with one adjacent run is at most $\binom{\tau(X)-1}{3}$, while the number of pairs with two adjacent runs is at most $\binom{\tau(X)-2}{2}$. Thus, the total contribution of pairs with $d_A(X_1,X_2)=2$ can be upper bounded by

$$\binom{\tau(X)-1}{3}+3\binom{\tau(X)-2}{2}.$$

This concludes our computation of an upper bound for $F_{2,2}(X,2)$. From Lemma 1, $L_{2,2}(X,1)\le|E_{2,2}(X)|$, where

$$L_{2,2}(X,1)=|D_2(X)|I_q(n-2,2)-F_{2,2}(X,2)+1-|D_2(X)|.$$

Substituting our upper bound for $F_{2,2}(X,2)$, we have that

$$\begin{aligned}|E_{2,2}(X)|\ge{}&|D_2(X)|(I_q(n-2,2)-1)-\\&\left(\binom{\tau(X)-1}{3}+3\binom{\tau(X)-2}{2}\right)-\tau(X)\times\\&\left(\binom{\tau(X)}{2}+\tau(X)-1\right)(I_q(n-1,1)-1)+1.\end{aligned}$$

Using the lower bound for $|D_2(X)|$ in (3), we replace $|D_2(X)|$ by $\sum_{i=0}^{2}\binom{\tau(X)-2}{i}$. Similarly, we may expand each I_q term. We have the result:

Theorem 2. *The number of sequences $|E_{2,2}(X)|$ formed by two deletions and two insertions is lower bounded as*

$$\begin{aligned}|E_{2,2}(X)|\ge{}&\left(\sum_{i=0}^{2}\binom{\tau(X)-2}{i}\right)\left(n(q-1)+\binom{n}{2}(q-1)^2\right)\\&-\left(\binom{\tau(X)-1}{3}+3\binom{\tau(X)-2}{2}\right)\\&-\tau(X)\left(\binom{\tau(X)}{2}+\tau(X)-1\right)n(q-1)+1.\end{aligned}$$

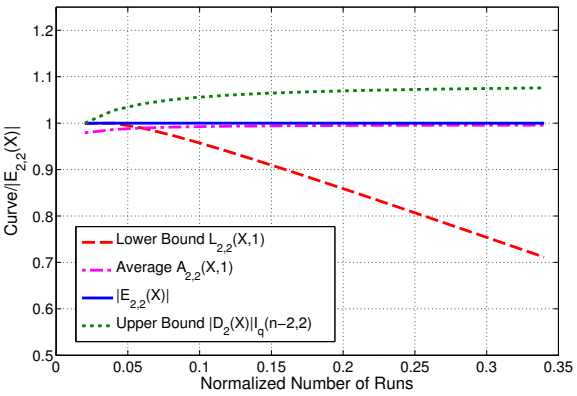

Fig. 2. Curves $L_{2,2}(X,1)$, $A_{2,2}(X,1)$, $|E_{2,2}(X)|$, and $|D_2(X)|I_q(n-2,2)$ normalized by $|E_{2,2}(X)|$. All strings X have $n=50$ and $q=3$. The x axis denotes normalized number of runs in string X, $\frac{\tau(X)}{n}$

The bound in our theorem was derived by examining the number of overlaps in the worst-case scenario. In the average case, the factor $\binom{\tau(X)}{2}+\tau(X)-1$ is replaced by $2\tau(X)-1$:

Theorem 3. *The size of the set $E_{2,2}(X)$ is approximated by*

$$A_{2,2}(X,1)=|D_2(X)|(I_q(n-2,2)-1) - \left(\binom{\tau(X)-1}{3}+3\binom{\tau(X)-2}{2}\right) -\tau(X)\,(2\tau(X)-1)\,(I_q(n-1,1)-1)+1.$$

We omit the proof due to space constraints. The excellent accuracy of the approximation is illustrated in Fig. 2.

Next, we proceed to generalize the above to bound $F_{t,t}(X,2)$ (where $t>2$). In this case, the distances $d_A(X_1,X_2)\in\{1,2,\ldots,t\}$. Consider a pair (X_1,X_2) at distance j. Then, there is at least one sequence $Y\in I_j(X_1)\cap I_j(X_2)$. Y is of length $n-t+j$. Since $I_{t-j}(Y)\in I_t(X_1)\cap I_t(X_2)$, the pair (X_1,X_2) contributes $I_q(n-t+j,t-j)-1$ to $F_{t,t}(X,2)$. The number of pairs of subsequences at distance j can be upper bounded as follows: there are $|D_{t-j}(X)|$ sequences to use as an ancestor for X_1 and X_2. For each of these, there are at most $\binom{\tau(X)}{2j}$ pairs (X_1,X_2).

However, we must also consider pairs (X_1,X_2) where $|I_j(X_1)\cap I_j(X_2)|>1$. As before, this takes place when some of the deletions are from adjacent runs. Let us say that among the j deletions that are not common to X_1 and X_2, exactly k are in adjacent runs. Then, for each such pair, we have an additional (2^k-1) common subsequences in $|I_j(X_1)\cap I_j(X_2)|$. This is because we may produce any sequence where the adjacent run elements are exchanged. We subtract one because we already counted X. The number of such pairs (X_1,X_2) with k adjacent run deletions is upper bounded by $\binom{\tau(X)-k}{2j-k}$. Thus we have,

$$F_{t,t}(X,2) \le \sum_{j=1}^{t}|D_{t-j}(X)|\left[\binom{\tau(X)}{2j}+\sum_{k=1}^{j}(2^k-1)\binom{\tau(X)-k}{2j-k}\right](I_q(n-t+j,t-j)-1). \quad (6)$$

Now we will write the bound $L_{t,t}(X,1)\le|E_{t,t}(X)|\le|D_t(X)|I_q(n-t,t)$. We replace $F_{t,t}(X,2)$ in $L_{t,t}(X,1)$ with our upper bound (6).

Theorem 4. *$|E_{t,v}(X)|$, the number of sequences formed by t deletions and v insertions from $X\in F_q^n$, satisfies*

$$\left(\sum_{i=0}^{t}\binom{\tau(X)-t}{i}\right)\left(\sum_{i=0}^{t}\binom{n}{i}(q-1)^i-1\right)- \sum_{j=1}^{t}|D_{t-j}(X)|\left[\binom{\tau(X)}{2j}+\sum_{k=1}^{j}(2^k-1)\binom{\tau(X)-k}{2j-k}\right]\times \left(\sum_{i=0}^{t-j}\binom{n}{i}(q-1)^i-1\right)+1\le|E_{t,t}(X)|\le \binom{\tau(X)+t-1}{t}\left(\sum_{i=0}^{v}\binom{n}{i}(q-1)^i\right).$$

Of course, it is possible to estimate $F_{t,v}(X,k)$ for $k\ge3$. However, the complexity and length of such expressions makes finding and writing them increasingly challenging. Nevertheless, even initial bounds such as $L_{t,t}(X,1)$ and $D_t(X)|I_q(n-t,t)|$ are quite good. We plot these bounds, along with the approximation $A_{2,2}(X,1)$ against $|E_{t,v}(X)|$ for the $t=v=2$ case in Fig. 2. Here, we examined sequences of length $n=50$ with increasing number of runs ($2\le\tau(X)\le17$), and pre-computed $|D_2(X)|$ and $I_q(n-2,2)$. In the $\tau(X)=2$ case, $L_{2,2}(X,1)=|E_{2,2}(X)|$, so our lower bound is tight. With a large number of runs, the upper bound is better than the lower bound, though the approximation $A_{2,2}(X,1)$ derived from $L_{2,2}(X,1)$ is increasingly accurate and beats both lower and upper bounds.

VI. Conclusion

In our work, we have explored the general problem of counting sequences formed from a constant number of insertions *and* deletions. We provided exact expressions in the tractable case where $t=v=1$. We introduced a framework for determining bounds in the general case $t=v$ and gave some particular bounds for $t=v=2$.

Acknowledgment

Research supported in part by NSF grants CCF-1029030 and CCF-1150212, and the NSF GRFP.

References

[1] L. Calabi and W.E. Hartnett, "Some general results of coding theory with applications to the study of codes for the correction of synchronization errors," *Information and Control*, vol. 15, no. 3, 1969.

[2] A. Kulkarni and N. Kiyavash, "Non-asymptotic upper bounds for deletion correcting codes," *IEEE Trans. on Info. Theory*, accepted, 2012.

[3] D.S. Hirschberg and M. Regnier, "Tight bounds on the number of string subsequences," *Journal of Disc. Algorithms*, vol. 1 no. 1, 2000.

[4] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, 1966.

[5] Y. Liron and M. Lanbgerg, "A characterization of the number of subsequences obtained via the deletion channel," *ISIT*, Cambridge, MA, 2012.

[6] V.I. Levenshtein, "Efficient reconstruction of seqeunces from their subsequences or supersequences," *Journal of Combinatorial Theory*, vol. 93, no. 2, pp. 310-332, 2001.