Tight Bounds on the Number of String Subsequences

DANIEL S. HIRSCHBERG, *Department of Information and Computer Science, University of California, Irvine, CA 92697-3425, Email:* dan@ics.uci.edu

MIREILLE REGNIER, *INRIA Rocquencourt, 78153 Le Chesnay Cedex, FRANCE, Email:* Mireille.Regnier@inria.fr

ABSTRACT: The problem considered is that of determining the number of subsequences obtainable by deleting t symbols from a string of length n over an alphabet of size s. Recurrences are proven and solved for the maximum and average case values, and bounds on these values are exhibited.

Keywords: subsequence, recurrence

1 Problem Definition

We prove bounds on the number of subsequences of a given length that a string on a fixed-size alphabet can have. Such bounds have been the basis for an efficient algorithm that reconstructs a binary string from knowledge of a sufficient number of its subsequences [6]. This research area is linked to applications of Levenshtein distance whose usage "plays the central role" in "the study of block codes capable of correcting substitution and synchronization errors" [2].

A Σ-*string* is a string over Σ, where Σ is an s-alphabet (that is, $|\Sigma| = s$); Σ^n denotes the set of all Σ-strings of length n. A *series* is a maximal run of identical symbols and $\tau(X)$ denotes the number of series in string X. A *subsequence* Y of string X is a string obtained by deleting zero or more symbols from X, and X is said to be a *supersequence* of Y. A t-*subsequence* is a subsequence obtained by deleting exactly t symbols. The set of t-subsequences of X is denoted by $D_t(X)$. For example, $D_2(aab) = \{a, b\}$ and $D_t(X) = \emptyset$ for $t > |X|$.

Calabi ([1], as cited in [2]) proved that a particular string form attains the maximum value of $|D_t(X)|$ and found an expression for the generating function of that maximum value. We present a direct alternative proof of the upper bound and prove a simple underlying recurrence, thereby enabling its efficient evaluation.

Levenshtein [3] proved that, for any binary string X, $\binom{\tau(X)-t+1}{t} \leq |D_t(X)| \leq \binom{\tau(X)+t-1}{t}$. (These bounds can be generalized to Σ-strings [5].) However, while the upper bound is tight, the lower bound is not. We prove a tight lower bound.

Assuming that X is equally likely any string in Σ^n, we derive and solve a recurrence

on the average value of $|D_t(X)|$.

2 Upper Bounds for $|D_t(X)|$

We determine an upper bound on the number of subsequences obtainable by deleting t symbols from a string of length n over an alphabet of size s.

Let $\Sigma = \{\sigma_1, \ldots, \sigma_s\}$, where the $\{\sigma_i\}$ are listed in some order. Let $C_n = c_1 \cdots c_n$ be a string in Σ^n, where $c_i = \sigma_{1+(i-1 \bmod s)}$. Thus, C_n has the symbols of Σ in circular order, cycling as many times as needed.

Let $d_s(t,n)$ denote $|D_t(C_n)|$, the number of subsequences obtainable by deleting t symbols from C_n, where Σ has cardinality s.

Calabi ([1], as cited in [2]) proved that, for all X in $\Sigma^n, |D_t(X)| \leq d_s(t,n)$, and that $d_s(t,n)$ is the coefficient of x^n in the generating function $\phi(x) = (\sum_{j=1}^{s} x^j)^{n-t}(\sum_{j=0}^{\infty} x^j)$. Apparently, "the proof is rather involved" ([2], p.118) and was not published. In Theorem 2.4, we present a direct alternative proof of the upper bound. We prove Calabi's Theorem, that $\phi(x)$ is the generating function for $d_s(t,n)$, and use that proof as a basis to prove Theorem 2.6, a simple recurrence on $d_s(t,n)$. This recurrence enables the efficient evaluation of $d_s(t,n)$ by the use of dynamic programming.

We use $Q^{(a)}$ to denote the subset of strings in set Q that begin with symbol a. For example, $D_t^{(b)}(X)$ denotes the set of t-subsequences of X that start with symbol b. If Q and R are sets then we use $Q + R$ to denote $Q \cup R$ with the assertion that Q and R are disjoint and, thus, $|Q + R| = |Q| + |R|$.

LEMMA 2.1
For any Σ-string X, $D_t(X) = \sum_{a \in \Sigma} D_t^{(a)}(X)$.

PROOF. The set of strings $D_t(X)$ is partitioned into subsets organized by each string's first symbol. ∎

LEMMA 2.2
For $s \geq 1, d_s(t-1, n-1) \leq d_s(t,n)$.

PROOF. $d_s(t-1, n-1)$ counts subsequences of length $n-t$ as does $d_s(t,n)$, but of a smaller string. ∎

If Q is a set of Σ-strings and $\sigma \in \Sigma$ is a symbol then σQ denotes the set of Σ-strings $\{\sigma q \mid q \in Q\}$.

LEMMA 2.3
For $s \geq 1, d_s(t,n) = \sum_{i=1}^{s} d_s(t+1-i, n-i)$.

PROOF. For $1 \leq j \leq s$, let $C_n^{(\sigma_j)} = c_1 \cdots c_n$ be a string in Σ^n, where $c_i = \sigma_{1+(j+i-2 \bmod s)}$. Thus, $C_n^{(\sigma_j)}$ has the symbols of Σ in circular order, beginning with σ_j. Using Lemma 2.1, we see that $D_t(C_n) = \sum_{i=1}^{s} D_t^{(\sigma_i)}(C_n^{(\sigma_1)}) = \sum_{i=1}^{s} \sigma_i D_{t+1-i}(C_{n-i}^{(\sigma_{i+1})})$. The set of subsequences is partitioned by their beginning symbols. The subsequences that begin with σ_i are obtained by deleting the first $i-1$ symbols of C_n and $t+1-i$ symbols from among the symbols occurring after σ_i. The statement of the lemma follows directly. ∎

THEOREM 2.4
For $s \geq 1$ and for any $X \in \Sigma^n, |D_t(X)| \leq d_s(t,n)$.

PROOF. By induction on n and $n-t$. The theorem is trivially true for $n \leq 1$ and $n-t \leq 1$. Let $X = x_1 \cdots x_n \in \Sigma^n$. For each symbol σ_i, let f_i be the smallest index j such that $x_j = \sigma_i$ (and f_i is $n+1$ if σ_i does not appear in X), where the elements of Σ, $\{\sigma_1, \ldots, \sigma_n\}$, are ordered by their first appearance in X, thereby ordering f_i smallest to largest. Consequently, $f_i \geq i$. We use $X[i:j]$ to denote the substring $x_i \cdots x_j$ of X.

Using Lemma 2.1, we have

$$D_t(X) = \sum_{i=1}^{s} D_t^{(\sigma_i)}(X) = \sum_{i=1}^{s} \sigma_i D_{t+1-f_i}(X[f_i+1:n]) \ .$$

Therefore,

$$\begin{aligned}
|D_t(X)| &= \sum_{i=1}^{s} |D_{t+1-f_i}(X[f_i+1:n])| \\
&\leq \sum_{i=1}^{s} |D_{t+1-f_i}(C_{n-f_i})|, \text{ using the inductive hypothesis,} \\
&\leq \sum_{i=1}^{s} |D_{t+1-i}(C_{n-i})|, \text{ because } f_i \geq i \text{ and by applying Lemma 2.2,} \\
&= d_s(t,n), \text{ by applying Lemma 2.3.}
\end{aligned}$$

∎

In the following, $[z^n]\text{poly}(z)$ denotes the coefficient of z^n in a polynomial of z.

THEOREM 2.5
[Calabi] $d_s(t,n)$ satisfies

$$d_s(t,n) = [z^n](\sum_{j=1}^{s} z^j)^{n-t} \frac{1}{1-z} \ .$$

PROOF. The following process exhibits a coding of any t-subsequence X of C_n as an $(n-t)$-tuple $\alpha = (\alpha_1, \ldots, \alpha_{n-t})$, where α_i are nonnegative integers, $0 \leq \alpha_i \leq s-1$, and $\sum_{i=1}^{n-t} \alpha_i \leq t$.

The elements of X can be associated with their counterparts in C_n. Each α_i corresponds with the i-th element of X, $X[i]$, and denotes the shift in C_n, *i.e.*, the minimum number of symbols of C_n that are to be deleted before encountering $X[i]$. For example, for $C_n = ABCABC$ and 2-subsequence $X = ABBC$, the first two symbols of X match with C_n and are encoded with zeros but the third symbol of X requires a skip of two symbols in C_n and is encoded with a 2. Because of the cyclic nature of the

symbols in C_n, a sequence of s contiguous deleted symbols do not cause a mismatch and hence the shift will be an integer in $\{0, \ldots, s-1\}$.

We note that $\sum_{i=1}^{n-t} \alpha_i \leq t$ and that the inequality is strict when deletions occur at the end of the sequence. We also note that the number, r, of deletions occurring at the end of the sequence may be greater than $s-1$. This number need not be coded, as it is determined by the equality $\sum_{i=1}^{n-t} \alpha_i + r = t$.

Example. Let $C_{12} = ABCABCABCABC$ and $t = 3$. The 3-subsequence $AABCABCAB$ is coded on a 3-alphabet by 020000000. Reciprocally, from coding 001002000, one recovers 3-subsequence $ABABCCABC$.

Enumeration of $d_s(t,n)$ reduces to an enumeration of $\{\alpha_1, \ldots, \alpha_{n-t}\}$. Each value α_i is associated with a subword of length $\alpha_i + 1$ of the initial sequence. There are $n-t$ choices of subword lengths from $\{1, \ldots, s\}$. Using a generating function, this is enumerated by $(\sum_{j=1}^{s} z^j)^{n-t}$. A choice where $\sum_{i=1}^{n-t} \alpha_i = t - r$ is the same as a choice where the sum of the subword lengths is $n-r$, and leads to an index $n-r$ for the last chosen element. In this case, the r last symbols of C_n are deleted, where $0 \leq r \leq t$. The coefficient of z^k in the generating function denotes the number of choices of subword lengths (equivalently, the number of choices of α) that result in the last symbol being at index k. Hence,

$$\begin{aligned} d_s(t,n) &= \sum_{r=0}^{t} [z^{n-r}] (\sum_{j=1}^{s} z^j)^{n-t} \\ &= \sum_{r=0}^{t} [z^n] (\sum_{j=1}^{s} z^j)^{n-t} z^r \\ &= [z^n] (\sum_{j=1}^{s} z^j)^{n-t} \sum_{r=0}^{t} z^r \ . \end{aligned}$$

As $[z^n](\sum_{j=1}^{s} z^j)^{n-t} \sum_{r=t+1}^{\infty} z^r = 0$, we obtain our result.

Alternatively, count the t-subsequences that have total shift $t - r \leq t$. Hence,

$$\begin{aligned} d_s(t,n) &= \sum_{r=0}^{t} \text{(the number of choices of } \alpha \text{ that cause } \sum_{i=1}^{n-t} \alpha_i = r) \\ &= \sum_{r=0}^{t} [z^{t-r}] (1 + z + \cdots + z^{s-1})^{n-t} \\ &= \sum_{r=0}^{t} [z^t] (1 + z + \cdots + z^{s-1})^{n-t} z^r \\ &= [z^t] (1 + z + \cdots + z^{s-1})^{n-t} \sum_{r=0}^{t} z^r \\ &= [z^t] (1 + z + \cdots + z^{s-1})^{n-t} \sum_{r=0}^{\infty} z^r . \end{aligned}$$

This last equality comes from the fact that $[z^t](1+z+\cdots+z^{s-1})^{n-t}\sum_{k\geq t+1} z^k$ is 0. Finally, we rewrite the last expression as $= [z^n]z^{n-t}(1+z+\cdots+z^{s-1})^{n-t}\sum_{r=0}^{\infty} z^r$. ∎

THEOREM 2.6
For $0 \leq t \leq n$ and $s \geq 2, d_s(t,n) = \sum_{i=0}^{t} \binom{n-t}{i} d_{s-1}(t-i,t)$.

PROOF. We exhibit a mapping between t-subsequences of C_n on an s-alphabet and $(t-i)$-subsequences of C_t on an $(s-1)$-alphabet, for $i \in \{0,\ldots,t\}$.

The process used in the proof of Theorem 2.5 can be used to code a $(t-i)$-subsequence of string C_t on an $(s-1)$-alphabet (the subsequence having length i), as an i-tuple over the alphabet $\{0,\ldots,s-2\}$. The sum of the entries in the coding will be $t-i-r$, for some $0 \leq r \leq t-i$. Alternatively, by increasing each entry value by one, the coding will be an i-tuple over the alphabet $\{1,\ldots,s-1\}$ and the sum of the entries will be $t-r$, for some $0 \leq r \leq t-i$. We note that, similarly, a t-sequence of string C_n (having length $n-t$) over an s-alphabet could be encoded by choosing the i non-zero positions among $n-t$, and encoding the resulting i-tuple.

The set T of t-subsequences of string C_n on an s-alphabet can be coded as $(n-t)$-tuples over the alphabet $\{0,\ldots,s-1\}$, each with entry sum $\leq t$. The number of non-zero positions i in any one of these codings is therefore at most t. For each possible value of i, extract the i non-zero positions, decrease each value by 1, and obtain a coding for the sequence of length i over alphabet $\{0,\ldots,s-2\}$ with sum $\leq t-i$. Performing this for all members of set T results in obtaining each $(t-i)$-subsequence $\binom{n-t}{i}$ times. ∎

COROLLARY 2.7
For $0 \leq t \leq n, d_2(t,n) = \sum_{i=0}^{t} \binom{n-t}{i}$; for $0 \leq t \leq n, d_3(t,n) = \sum_{i=0}^{t} \binom{n-t}{i} \sum_{j=0}^{t-i} \binom{i}{j}$.

PROOF. This follows immediately from Theorem 2.6 and the fact that $d_1(t,n) = 1$ for $0 \leq t \leq n$. ∎

COROLLARY 2.8
For $0 \leq t \leq n$ and $s \geq 2$, the value of $d_s(t,n)$ can be computed in time $O(t + (s-1)n^2)$.

PROOF. The set of values $C = \{\binom{j}{i} \mid 0 \leq i \leq j \leq n\}$ can be computed in time $O(n^2)$ using dynamic programming. Let $d_s = \{d_s(i,n) \mid 0 \leq i \leq n\}$. From Corollary 2.7, the set d_2 can be computed in time $O(n^2)$ from C. For k iteratively valued $3,\ldots,s-1$, using the recurrence of Theorem 2.6, the set d_k can be computed in time $O(n^2)$ from C and d_{k-1}. Finally, again using the recurrence of Theorem 2.6, $d_s(t,n)$ can be computed in time $O(t)$ from d_{s-1} and C. ∎

OBSERVATION 2.9
It may be possible to improve the complexity of computing $d_s(t,n)$ by using memoized recursion.

By evaluating $d_s(t,n)$ and expressing its difference from $\binom{n}{t}$ as a power series, one can see that, for $t \geq s, d_s(t,n) = \binom{n}{t} - n^{t+1-s}/(t-s)! + O(n^{t-s})$.

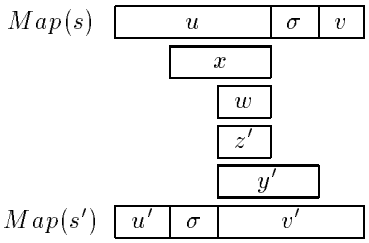

FIG. 1. If $Map(s) = Map(s')$

We note that the problem of calculating the number, $i_s(t,n)$, of supersequences obtainable by inserting t symbols in a length n string X on an alphabet of size s is much simpler, and is invariant over X. It is known [4] that, using a binary alphabet, $i_2(t,n) = \sum_{j=0}^{t} \binom{n+t}{j}$. This can be generalized to $s \geq 2$ [5]. It is easy to see that $i_s(t,n) = i_s(t,n-1) + (s-1)i_s(t-1,n)$, with boundary conditions $i_s(0,n) = 1$ and $i_s(t,0) = s^t$. Let $X \in \Sigma^{n-1}, Y \in \Sigma^{n+t-1}$, and $a, b \in \Sigma$. Then bY is a supersequence of aX if and only if either (1) $a = b$ and Y is a supersequence of X, or (2) $a \neq b$ and Y is a supersequence of aX. This recurrence is solved by $i_s(t,n) = \sum_{j=0}^{t}(s-1)^j \binom{n+t}{j}$.

3 A Lower Bound for $|D_t(X)|$

It was stated [6, 3] that, for any binary string X, $|D_t(X)| \geq \binom{\tau(X)-t+1}{t}$. Note that this bound is the same as $\binom{\tau(X)-t}{t} + \binom{\tau(X)-t}{t-1}$. We will improve and generalize this bound. We first need a few lemmas.

LEMMA 3.1
For any Σ-strings U, V and any $\sigma \in \Sigma$, $|D_t(UV)| \leq |D_t(U\sigma V)|$.

PROOF. We show that $s \in D_t(UV)$ can be mapped distinctly to $Map(s) \in D_t(U\sigma V)$.

Let $s = uv$, with u the maximum length prefix of s that is a subsequence of U. We define $Map(s)$ to be $u\sigma v$. It remains to be shown that if $s \neq s'$ then $Map(s) \neq Map(s')$.

Let $s = uv$ and $s' = u'v'$ be elements in $D_t(UV)$, where u and u' are the maximum length prefixes of s and s' that are subsequences of U. If $u = u'$ then it must be that $v \neq v'$ and thus $u\sigma v \neq u\sigma v'$. We now consider the case when $u \neq u'$.

Assume that $Map(s) = Map(s')$, *i.e.*, $u\sigma v = u'\sigma v'$. Without loss of generality, $|u| > |u'|$, and we let $u = u'x$ and $v' = y'v$, with x non-null. Refer to Figure 1. Thus $u'x\sigma v = u'\sigma y'v$ and therefore $x\sigma = \sigma y'$. Then we can express x as $x = \sigma w$ and y' as $y' = z'\sigma$. Accordingly, $\sigma w\sigma = \sigma z'\sigma$. We note that z' is non-null, else s would equal s'. Now we see that $u = u'x = u'\sigma w$ and that $s' = u'v' = u'y'v = u'z'\sigma v = u'w\sigma v$. However, this means that $u'z'$ is a prefix of s' and, since $u'z'$ is a subsequence of u, it follows that $u'z'$ is a subsequence of U, contradicting the maximality of u'. ∎

LEMMA 3.2
If X is a Σ-string such that $\tau(X) = n$ then there exists a string $Y \in \Sigma^n$, with $\tau(Y) = n$, such that $|D_t(Y)| \leq |D_t(X)|$.

PROOF. Let Y be Σ-string of length n consisting of one symbol from each of the series in X. String X can be obtained from Y by a sequence of symbol insertions. The statement of the lemma then follows from repeated applications of Lemma 3.1. ∎

LEMMA 3.3
If X is a string in Σ^n such that $\tau(X) = n$ then $|D_t(X)| \geq d_2(t, n)$.

PROOF. By induction on n and $n - t$. The lemma is trivially true for the base cases, when $n \leq 2$ or $n - t \leq 2$. For the induction step, let $X = abY$, where $a \neq b$ because each series in X has length 1. Then,

$$\begin{aligned} D_t(X) &= D_t^{(a)}(X) + D_t^{(b)}(X) + \sum_{\sigma \neq a,b} D_t^{(\sigma)}(X) \\ &\supseteq D_t^{(a)}(X) + D_t^{(b)}(X) = aD_t(bY) + bD_{t-1}(Y) \ . \end{aligned}$$

Using the inductive hypothesis and Lemma 2.3, we obtain

$$\begin{aligned} |D_t(X)| &\geq |D_t(bY)| + |D_{t-1}(Y)| \\ &\geq d_2(t, n-1) + d_2(t-1, n-2) = d_2(t, n) \ . \end{aligned}$$

∎

THEOREM 3.4
For any Σ-string $X, |D_t(X)| \geq \sum_{i=0}^{t} \binom{\tau(X)-t}{i}$ and for all values of τ and t, where $t \leq \tau$, there exists a string X such that $\tau(X) = \tau$ and $|D_t(X)| = \sum_{i=0}^{t} \binom{\tau(X)-t}{i}$.

PROOF. Follows directly from Corollary 2.7 and Lemmas 3.2 and 3.3. ∎

4 The Average Number of Subsequences

Under the assumption that Σ-strings of length n are equiprobable, the average number of subsequences obtainable by deleting one symbol has been shown to be $(n(s-1)+1)/s$ [2]. We develop and solve a recurrence on the average number of subsequences obtainable by deleting t symbols.

Let $G_t(n) = \sum_{X \in \Sigma^n} |D_t(X)|$ be the sum, over all strings in Σ^n, of the number of subsequences obtainable by deleting t symbols. Similarly, $G_t^{(a)}(n) = \sum_{X \in \Sigma^n} |D_t^{(a)}(X)|$ is the sum when the subsequences are restricted to begin with symbol a.

We see that, for $0 < t < n$,

$$G_t^{(a)}(n) = \sum_{X \in \Sigma^n} |D_t^{(a)}(X)| = \sum_{b \in \Sigma} \sum_{X \in (\Sigma^n)^{(b)}} |D_t^{(a)}(X)| \ . \tag{4.1}$$

If $b = a$ then $(\Sigma^n)^{(b)} = \{aY \mid Y \in \Sigma^{n-1}\}$. We note that the count of subsequences of aY that start with a and have length $n - t$ is the same as the count of

subsequences of Y that have length $n-1-t$ because of a simple bijection between those two sets of subsequences. As a result, we see that $\sum_{X\in(\Sigma^n)^{(a)}}|D_t^{(a)}(X)| = \sum_{Y\in\Sigma^{n-1}}|D_t^{(a)}(aY)| = \sum_{Y\in\Sigma^{n-1}}|D_t(Y)|$.

If $b\neq a$ then $(\Sigma^n)^{(b)} = \{bY \mid Y\in\Sigma^{n-1}\}$. We note that the count of subsequences of bY that start with a and have length $n-t$ is the same as the count of subsequences of Y that start with a and have length $n-t$ because the leading b of bY can just be discarded. As a result, we see that $\sum_{X\in(\Sigma^n)^{(b)}}|D_t^{(a)}(X)| = \sum_{Y\in\Sigma^{n-1}}|D_t^{(a)}(bY)| = \sum_{Y\in\Sigma^{n-1}}|D_{t-1}(Y)|$.

Therefore,

$$G_t^{(a)}(n) = \sum_{Y\in\Sigma^{n-1}}|D_t(Y)| + (s-1)\sum_{Y\in\Sigma^{n-1}}|D_{t-1}^{(a)}(Y)| \ . \tag{4.2}$$

We then see that

$$G_t^{(a)}(n) = G_t(n-1) + (s-1)G_{t-1}^{(a)}(n-1) \tag{4.3}$$

follows immediately from (4.1), (4.2) and the definitions.

From the fact that $G_t(n) = \sum_{a\in\Sigma}G_t^{(a)}(n)$, using (4.3) s times, once for each symbol in Σ,

$$\begin{aligned}
G_t^{(\sigma_1)}(n) &= G_t(n-1) + (s-1)G_{t-1}^{(\sigma_1)}(n-1)\\
G_t^{(\sigma_2)}(n) &= G_t(n-1) + (s-1)G_{t-1}^{(\sigma_2)}(n-1)\\
&\cdots\\
G_t^{(\sigma_s)}(n) &= G_t(n-1) + (s-1)G_{t-1}^{(\sigma_s)}(n-1)
\end{aligned}$$

obtains

$$G_t(n) = sG_t(n-1) + (s-1)G_{t-1}(n-1) \ . \tag{4.4}$$

Boundary conditions, $G_0(n) = G_n(n) = s^n$, hold because there is only one string obtainable by deleting none or all of the symbols in each of the s^n strings in Σ^n.

Let $E_t(n)$ be the average (or expected) value of $|D_t(X)|$, where X can equally likely be any string in Σ^n, and let $\lambda = 1-1/s$.

THEOREM 4.1
For $0<t<n$, $E_t(n) = E_t(n-1) + \lambda E_{t-1}(n-1)$, and $E_0(n) = E_t(t) = 1$.

PROOF. This follows from the recurrence (4.4) and boundary conditions on G and the fact that $E_t(n) = G_t(n)/s^n$. ∎

THEOREM 4.2
Let $E_t(n)$ be the expected value of $|D_t(X)|$. Then the bivariate generating function $E(z,u) = \sum_{n\geq 1}\sum_{1\leq t\leq n}E_t(n)u^tz^n$ has closed form

$$E(z,u) = \frac{zu}{1-zu}\left(1+\frac{z}{1-z-\lambda zu}\right) \ .$$

It follows that

$$E_t(n) = \sum_{i=0}^{t} \binom{n-t-1+i}{i} \lambda^i .$$

PROOF. From Theorem 4.1, sequence $\{E_t(n)\}_{n \geq t \geq 1}$ satisfies

$$E_t(n) = E_t(n-1) + \lambda E_{t-1}(n-1) + (1-\lambda) 1_{n=t \geq 1} , \tag{4.5}$$

with $E_n(n-1) = 0$ and $E_n(n) = 1$. Multiply both sides of (4.5) by $z^n u^t$ and sum over all n and t such that $n \geq t \geq 1$. Noting that $E_t(n-1) z^n u^t = z E_t(n-1) z^{n-1} u^t$, we obtain

$$E(z,u) = zE(z,u) + \lambda zu[E(z,u) + E_0(0)] + (1-\lambda) \sum_{n \geq 1} (uz)^n ,$$

which can be rewritten as

$$E(z,u)(1 - z - \lambda zu) = \lambda zu + \frac{(1-\lambda)zu}{1-zu} = \frac{zu}{1-zu}(1-\lambda zu)$$

and simplification yields the claimed closed form.

It follows that $E_t(n) = [u^t][z^n]E(z,u)$. If $n > t$, one observes that $[u^t][z^n]\frac{zu}{1-zu} = 0$ and using the factorization $z^n = z^{n-t} \cdot z^t$ yields

$$\begin{aligned} E_t(n) &= [(uz)^t][z^{n-t}] \frac{zu}{1-zu} \cdot \frac{z}{1-z-\lambda zu} \\ &= [(uz)^t][z^{n-t-1}] \frac{zu}{1-zu} \cdot \frac{1}{1-z-\lambda zu} . \end{aligned}$$

Expanding the last expression yields

$$E_t(n) = \sum_{i=0}^{t} [z^{n-t-1}](z + \lambda zu)^{n-t-1+i} = \sum_{i=0}^{t} \binom{n-t-1+i}{i} \lambda^i .$$

∎

5 Conclusions

We have shown bounds on the number of sequences obtainable by deleting t symbols from a length n string over an alphabet of size s. That number varies with the string and it would be of interest to be able to efficiently compute that number for any given string.

We noted that the number of sequences obtainable by inserting t symbols in a length n string X over an alphabet of size s is invariant over X. A more general problem, combining these operations, is to compute the number of sequences obtainable by applying a sequence of t *insert symbol* and *delete symbol* operations to a given string.

Acknowledgements

We thank an anonymous referee whose detailed comments significantly helped improve the presentation of our results.

References

[1] L. Calabi, "On the computation of Levenshtein's distances," TM-9-0030, Parke Math. Labs., Inc., Carlisle, Mass., 1967.

[2] 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 15*,3 (Sept 1969) 235-249. Reprinted in: W. E. Hartnett (ed), *Foundations of Coding Theory* (1974) Chapter 7, pp.107-121.

[3] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Phys. Dokl. 10* (1966) 707-710.

[4] V. I. Levenshtein, "Elements of the coding theory," in: *Discrete Math. and Math. Probl. of Cybern.*, Nauka, Moscow (1974) 207-235 (in Russian).

[5] V. I. Levenshtein, "On perfect codes in deletion and insertion metric," *Discrete Math. Appl. 2*,3 (1992) 241-258. Originally published in *Diskretnaya Matematika 3*,1 (1991) 3-20 (in Russian).

[6] V. I. Levenshtein, "Reconstructing binary sequences by the minimum number of their subsequences or supersequences of a given length," *Proc. 5th Int. Wkshp on Alg. & Comb. Coding Theory*, Sozopol, Bulgaria (1996) 176-183.

Received November 15, 1999.