On Subsets of Binary Strings Immune to Multiple Repetition Errors

Lara Dolecek
EECS Department
University of California
Berkeley, CA 94720, USA
Email: dolecek@eecs.berkeley.edu

Venkat Anantharam
EECS Department
University of California
Berkeley, CA 94720, USA
Email: ananth@eecs.berkeley.edu

***Abstract*— In this paper we revisit previously proposed techniques for constructing some families of subsets of binary strings (codes) that are immune to multiple repetition errors. In particular, we discuss a technique to construct single repetition error correcting codes and use number theoretic methods to give an explicit formula for the cardinalities of these codes. This approach results in codes the ratio of whose cardinality to the best upper bounds approaches unity in the increasing codelength limit (asymptotic optimality). We also discuss a somewhat different technique to construct multiple repetition error correcting codes. Here the cardinalities are asymptotically within a fixed constant of the best known upper bounds. Our constructions are asymptotically better by a constant factor than the best previously known such constructions, due to Levenshtein.**

I. Introduction

Substitution error correcting codes are traditionally used in communication systems for encoding of a binary input message $\mathbf{x}$ into a coded sequence $\mathbf{c} = C(\mathbf{x})$. The modulated version of this sequence is usually corrupted by additive noise, and is seen at the receiver as a waveform $s(t)$,

$$s(t) = \sum_i c_i h(t - iT) + n(t), \tag{1}$$

where c_i is the ith bit of $\mathbf{c}$, $h(t)$ is the modulating pulse, and $n(t)$ is the noise introduced in the channel. The received waveform $s(t)$ is sampled at certain sampling points determined by the timing recovery process, and the resulting sampled sequence is passed to the decoder which then produces the estimate of $\mathbf{c}$ (or $\mathbf{x}$). In the analysis of substitution error correcting codes and their decoding algorithms it is traditionally assumed that the decoder receives a sequence which is a properly sampled version of the waveform $s(t)$.

The timing recovery process involves a substantial overhead in the design of communication chips, both in terms of occupying area on the chip and in terms of power consumption. To avoid some of this cost, particularly in high speed systems, chip designers could attempt to make do with poorer timing recovery, while oversampling the received waveform to attempt to ensure that no information is lost. Thus the waveform $s(t)$ instead of being sampled at instances $kT_s + \tau_k$ might be sampled at instances roughly T apart, for $T < T_s$. In the idealized infinite SNR limit of a PAM system, this appears as if some symbols are sampled more than once. As a result, instead of creating n samples from $s(t)$, $n + r$ samples are produced, where $r \geq 0$. As a consequence, when $r > 0$, the decoder is presented with a sampled sequence whose length exceeds the length of a codeword.

Motivated by this scenario, in this paper we study the problem of finding maximally sized subsets of binary strings (codes) that are immune to a given number r of repetitions, in the sense that no two strings in the code can give rise to the same string after r repetitions.

A closely related problem of studying codes capable of overcoming a certain number of insertions and deletions was first studied by Levenshtein [8] where it was shown that the so-called Varshamov-Tenengolts codes [13] originally proposed for the correction of asymmetric errors are capable of overcoming one deletion or one insertion. They were also shown to be asymptotically optimal. They have been further studied in [5] and [2]. In [11] further results on their cardinalities were obtained. Extensions to constructions for overcoming multiple insertions and deletions have so far found limited success, [6], [12].

In Section II we first introduce an auxiliary transformation that converts our problem into that of creating subsets of binary strings immune to the insertions of 0's. In Section III we focus on subsets of binary strings immune to single repetitions. We present explicit constructions of such subsets and use number theoretic techniques to give explicit formulas for their cardinalities. Our constructions here are asymptotically optimal. In Section IV we discuss subsets of binary strings immune to multiple repetitions. Our constructions here are asymptotically within a constant factor of the best known upper bounds and asymptotically better, by a constant factor than the best previously known such constructions, due to Levenshtein [7].

II. Auxiliary transformation

To construct a binary, r repetitions correcting code C of length n we first construct an auxiliary code $\tilde{C}$ of length $m = n - 1$ which is an r '0'-insertions correcting code. These two codes are related through the following transformation.

Suppose $\mathbf{c} \in C$. We let $\tilde{\mathbf{c}} = \mathbf{c} \times T_n \bmod 2$, where T_n is $n \times n - 1$ matrix, satisfying

$$T_n(i,j) = \begin{cases} 1, & \text{if } i = j, j+1 \\ 0, & \text{else.} \end{cases} \quad (2)$$

Now, the repetition in $\mathbf{c}$ in position p corresponds to the insertion of '0' in position $p-1$ in $\tilde{\mathbf{c}}$, and weight($\tilde{\mathbf{c}}$) = number of runs in $\mathbf{c}-1$. We let $\tilde{C}$ be the collection of strings of length $n-1$ obtained by applying T_n to all strings C. Note that $\mathbf{c}$ and its complement both map into the same string in $\tilde{C}$.

It is thus sufficient to construct a code of length $n-1$ capable of overcoming r '0'-insertions and apply the inverse T_n transformation to obtain r repetitions correcting code of length n.

III. One repetition case

Following the analysis of Sloane [11] and Levenshtein [8] of the related Varshamov-Tenengolts codes [13] known to be capable of overcoming one deletion or one insertion, let A_w^m be the set of all binary strings of length m and w ones, for $0 \leq w \leq m$. Partition A_w^m based on the value of the first moment of each string. More specifically, let $S_{w,k}^m$ be the subset of A_w^m such that

$$S_{w,k}^m = \{(s_1, s_2, ..., s_m) | \sum_{i=1}^{m} i \times s_i \equiv k \bmod (w+1)\}. \quad (3)$$

Lemma 1: Each subset $S_{w,k}^m$ is a single '0'-insertion correcting code.

Proof: Suppose the string $\mathbf{s}'$ is received. We want to uniquely determine the codeword $\mathbf{s} = (s_1, s_2, ..., s_m) \in S_{w,k}^m$ such that $\mathbf{s}'$ is the result of inserting at most one zero in $\mathbf{s}$. If the length of $\mathbf{s}'$ is m, conclude that no insertion occurred, and that $\mathbf{s} = \mathbf{s}'$.

If the length of $\mathbf{s}'$ is $m+1$, a zero has been inserted. For $\mathbf{s}' = (s_1', s_2', ..., s_m', s_{m+1}')$, compute $\sum_{i=1}^{m+1} i \times s_i' \bmod (w+1)$. Due to the insertion, $\sum_{i=1}^{m+1} i \times s_i' = \sum_{i=1}^{m} i \times s_i + R_1$ where R_1 denotes the number of 1's to the right of the insertion. Note that R_1 is always between 0 and w.

Let k' be equal to $\sum_{i=1}^{m+1} i \times s_1' \bmod (w+1)$. If $k' = k$ the insertion occurred after the rightmost one, so we declare $\mathbf{s}$ to be the m leftmost bits in $\mathbf{s}'$. If $k' > k$ we declare $\mathbf{s}$ to be the string obtained by deleting the zero immediately preceding the rightmost $k' - k$ ones. Finally, if $k' < k$, we declare $\mathbf{s}$ to be the string obtained by deleting the zero immediately preceding the rightmost $w + 1 - k + k'$ ones. ■

Since $|A_w^m| = \binom{m}{w}$ there exists k such that

$$|S_{w,k}^m| \geq \frac{1}{w+1} \binom{m}{w}.$$

Since two codewords of different weights cannot result in the same string when at most one zero is inserted we let $\tilde{C}$ be the union of largest sets $S_{w,k_w^*}^m$ over different weights w, i.e.

$$\tilde{C} = \bigcup_{w=0}^{m} S_{w,k_w^*}^m,$$

where $S_{w,k_w^*}^m$ is the set of largest cardinality among all sets $S_{w,k}^m$ for $0 \leq k \leq w$, and the all-zeros string represents $S_{0,k_0^*}^m$.

Thus, the cardinality of $\tilde{C}$ is at least

$$\sum_{w=0}^{m} \binom{m}{w} \frac{1}{w+1} = \frac{1}{m+1} \left(2^{m+1} - 1\right).$$

The upper bound $U_1(m)$ on any set of strings each of length m capable of overcoming one insertion of a zero is derived in [7] to be

$$U_1(m) = \frac{2^{m+1}}{m}. \quad (4)$$

Hence the proposed construction is asymptotically optimal.

By applying the inverse T_n transformation for $n = m + 1$ to $\tilde{C}$ we obtain a code of length n and of size at least $\frac{1}{n}\left(2^{n+1} - 2\right)$.

The cardinalities of the sets $S_{w,k}^m$ may be computed explicitly as we now show.

Recall that the Möbius function $\mu(x)$ of a positive integer $x = p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k}$ for distinct primes $p_1, p_2, \ldots, p_k$ is defined as [1],

$$\mu(x) = \begin{cases} 1 & \text{for } x = 1 \\ (-1)^k & \text{if } a_1 = \cdots = a_k = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

and that the Euler function $\phi(x)$ denotes the number of integers y, $1 \leq y \leq x - 1$ that are relatively prime with x. By convention $\phi(1) = 1$.

Lemma 2: Let $g = gcd(m+1, w+1)$. The cardinality of $S_{w,k}^m$ is

$$|S_{w,k}^m| = \frac{1}{m+1} \sum_{d|g} \binom{\frac{m+1}{d}}{\frac{w+1}{d}} (-1)^{(w+1)(1+\frac{1}{d})} \phi(d) \frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} \quad (6)$$

where $gcd(d,k)$ is the greatest common divisor of d and k, interpreted as d if $k = 0$.

Proof: Motivated by the analysis of Sloane [11] of the Varshamov-Tenengolts codes, let us introduce the function $f_{b,n}(U,V)$ in which the coefficient of $U^s V^k$, call it $g_{k,s}^b(n)$, represents the number of strings of length n, weight s and the first moment equal to $k \mod b$

$$f_{b,n}(U,V) = \sum_{k=0}^{b-1} \sum_{s=0}^{n} g_{k,s}^b(n) U^s V^k. \quad (7)$$

Observe that $f_{b,n}(U,V)$ can be written as a generating function

$$f_{b,n}(U,V) = \prod_{z=1}^{n} (1 + UV^z) \mod (V^b - 1). \quad (8)$$

Let $a = e^{i\frac{2\pi}{b}}$ so that for $V = a^j$

$$f_{b,n}(U, e^{i\frac{2\pi j}{b}}) = \sum_{k=0}^{b-1} \sum_{s=0}^{n} g_{k,s}^b(n) U^s e^{i\frac{2\pi j k}{b}}. \quad (9)$$

With this substitution, we apply the inverse discrete Fourier transform and write

$$\begin{aligned} & \textstyle\sum_{s=0}^{n} g_{k,s}^{b}(n)U^s \\ = \; & \tfrac{1}{b}\textstyle\sum_{j=0}^{b-1} f_{b,n}(U, e^{i\frac{2\pi j}{b}})e^{-i\frac{2\pi jk}{b}} \qquad (10) \\ = \; & \tfrac{1}{b}\textstyle\sum_{j=0}^{b-1}\prod_{z=1}^{n}(1+Ue^{i\frac{2\pi jz}{b}})e^{-i\frac{2\pi jk}{b}} \; . \end{aligned}$$

Our next goal is to evaluate the coefficient U^b on the right hand side. To do so we first evaluate the following expression

$$\prod_{z=1}^{b}(1+Ue^{i\frac{2\pi jz}{b}}) \; . \qquad (11)$$

Let $d_j = b/gcd(b,j)$ and $s_j = j/gcd(b,j)$, and write

$$\begin{aligned} & \textstyle\prod_{z=1}^{b}(1+Ue^{i\frac{2\pi jz}{b}}) \\ = \; & \textstyle\prod_{z=1}^{d_j}(1+Ue^{i\frac{2\pi s_j z}{d_j}})\cdots\prod_{z=(d_j-1)\cdot gcd(b,j)+1}^{d_j\cdot gcd(b,j)}(1+Ue^{i\frac{2\pi s_j z}{d_j}}) \\ = \; & \left(\textstyle\prod_{z=1}^{d_j}(1+Ue^{i\frac{2\pi s_j z}{d_j}})\right)^{gcd(b,j)} \\ = \; & \Bigg(1+U\textstyle\sum_{z_1=1}^{d_j} e^{i\frac{2\pi s_j z_1}{d_j}} + \\ & \quad U^2\textstyle\sum_{z_1=1}^{d_j-1}\sum_{z_2=z_1+1}^{d_j} e^{i\frac{2\pi s_j(z_1+z_2)}{d_j}} + \\ & \quad + \cdots + U^{d_j} e^{i\frac{2\pi s_j(1+2+\cdots+d_j)}{d_j}}\Bigg)^{gcd(b,j)} \; . \end{aligned} \qquad (12)$$

Since $gcd(d_j, s_j) = 1$, the set

$$V = \{e^{i\frac{2\pi s_j 1}{d_j}}, e^{i\frac{2\pi s_j 2}{d_j}} \ldots e^{i\frac{2\pi s_j d_j}{d_j}}\}$$

represents all distinct solutions of the equation

$$x^{d_j} - 1 = 0 \; . \qquad (13)$$

For a polynomial equation $P(x)$ of degree d, the coefficient multiplying x^k is a scaled symmetric function of $d-k$ roots. Hence, symmetric functions involving at most $d_j - 1$ elements of V evaluate to zero. The symmetric function involving all elements of V, which is their product, evaluates to $(-1)^{d_j+1}$.

Therefore,

$$\prod_{z=1}^{b}(1+Ue^{i\frac{2\pi jz}{b}}) = \left(1+(-1)^{1+d_j}U^{d_j}\right)^{gcd(b,j)} . \qquad (14)$$

Returning to the inner product in (10), let us first suppose that $b|n$. Then

$$\begin{aligned} & \textstyle\prod_{z=1}^{n}\left(1+Ue^{i\frac{2\pi jz}{b}}\right) \\ = \; & \left(\textstyle\prod_{z=1}^{b}\left(1+Ue^{i\frac{2\pi jz}{b}}\right)\right)^{n/b} \\ = \; & \left(1+(-1)^{1+d_j}U^{d_j}\right)^{gcd(b,j)n/b} \qquad (15) \\ = \; & \textstyle\sum_{l=0}^{\frac{n}{d_j}}\binom{\frac{n}{d_j}}{l}(-1)^{l(1+d_j)}U^{ld_j} \; . \end{aligned}$$

Thus (10) becomes

$$\begin{aligned} & \sum_{s=0}^{n} g_{k,s}^{b}(n)U^s \\ = \; & \frac{1}{b}\sum_{j=0}^{b-1}\sum_{l=0}^{\frac{n}{d_j}}\binom{\frac{n}{d_j}}{l}(-1)^{l(1+d_j)}U^{ld_j}e^{-i\frac{2\pi jk}{b}} \; . \end{aligned}$$

We now regroup the terms whose j's yield the same d_j's

$$\begin{aligned} \sum_{s=0}^{n} g_{k,s}^{b}(n)U^s = \; & \frac{1}{b}\sum_{d|b}\sum_{l=0}^{\frac{n}{d}}\binom{\frac{n}{d}}{l}(-1)^{l(1+d)}U^{dl} \\ & \times \sum_{j:gcd(j,b)=b/d, 0\le j\le b-1} e^{-i\frac{2\pi jk}{b}} . \end{aligned}$$

The rightmost sum can also be written as

$$\sum_{j:gcd(j,b)=b/d, 0\le j\le b-1} e^{-i\frac{2\pi jk}{b}} = \sum_{s:0\le s\le d-1, gcd(s,d)=1} e^{-i\frac{2\pi sk}{d}} . \qquad (16)$$

This last expression is known as the Ramanujan sum [1] and simplifies to

$$\sum_{s:0\le s\le d-1, gcd(s,d)=1} e^{-i\frac{2\pi sk}{d}} = \phi(d)\frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} \; . \qquad (17)$$

Now the coefficient of U^b in (10) is

$$\frac{1}{b}\sum_{d|b}\binom{\frac{n}{d}}{\frac{b}{d}}(-1)^{\frac{b}{d}(1+d)}\phi(d)\frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} \qquad (18)$$

which is precisely the number of strings of length n, weight b, and the first moment congruent to $k \mod b$.

Consider the set of strings described by $S_{w,k}^{m}$ for $m = n-1$ and $w = b-1$. Suppose we append '1' to each such string, and call the resulting set B. Let A denote the set of strings described by (18). By grouping the elements of A based on their periodicity one can show that a fraction b/n of strings in A describe the set B. Therefore, the cardinality of $S_{w,k}^{m}$ is

$$|S_{w,k}^{m}| = \frac{1}{m+1}\sum_{d|w+1}\binom{\frac{m+1}{d}}{\frac{w+1}{d}}(-1)^{\frac{w+1}{d}(1+d)}\phi(d)\frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} \; . \qquad (19)$$

Notice that the last expression is the same as the one proposed in Lemma 2 with $gcd(m+1, w+1) = w+1$.

Now suppose that b is not a factor of n. We work with $f_{g,n}(U,V)$ as in (8) where $g = gcd(n,b)$ and get

$$\begin{aligned} \sum_{s=0}^{n} g_{k,s}^{g}(n)U^s = \; & \frac{1}{g}\sum_{d|g}\sum_{l=0}^{\frac{n}{d}}\binom{\frac{n}{d}}{l}(-1)^{l(1+d)}U^{dl} \\ & \times \sum_{j:gcd(j,g)=g/d, 0\le j\le g-1} e^{-i\frac{2\pi jk}{g}} \; . \end{aligned}$$

Thus the coefficient of U^b here is

$$\frac{1}{g}\sum_{d|g}\binom{\frac{n}{\frac{g}{d}}}{\frac{b}{d}}(-1)^{\frac{b}{d}(1+d)}\phi(d)\frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} . \tag{20}$$

This is the number of strings of length n, weight b, and the first moment congruent to $k \mod g$. Let A denote the set of these strings. Consider the set of strings of length $n-1$, weight $b-1$, and the first moment congruent to $k \mod g$, and call this set B. If we append '1' to each element of B we will obtain a b/n fraction of the elements of A. Since B is comprised of strings of length $n-1$, weight $b-1$, and the first moment congruent to $k_u = k+ug \mod b$ for $0 \le u \le b/g-1$ and since the evaluation in (20) is the same for all such k_u, it follows by symmetry that a fraction $\frac{g}{b}$ of B represents the set $S^m_{w,k}$. Therefore $|S^m_{w,k}|$ is

$|S^m_{w,k}| =$

$$\frac{1}{m+1}\sum_{d|g}\binom{\frac{m+1}{d}}{\frac{w+1}{d}}(-1)^{(w+1+\frac{1}{d}(1+w))}\phi(d)\frac{\mu\left(\frac{d}{gcd(d,k)}\right)}{\phi\left(\frac{d}{gcd(d,k)}\right)} \tag{21}$$

which completes the proof of the lemma. ■

A. Connection with necklaces

It is interesting to briefly visit the relationship between optimal single insertion of a zero correcting codes and combinatorial objects known as necklaces [4].

A necklace consisting of n beads can be viewed as an equivalence class of strings of length n under cyclic shift (rotation).

Let us consider two-colored necklaces of length n with b black beads and $n-b$ white beads. It is known that the total number of distinct necklaces is

$$T(n) = \frac{1}{n}\sum_{d|gcd(n,b)}\binom{\frac{n}{d}}{\frac{b}{d}}\phi(d) . \tag{22}$$

In general necklaces may exhibit periodicity. However, consider, for example, the case $gcd(n,b)=1$. Then there are

$$\frac{1}{n}\binom{n}{b}$$

distinct necklaces, all of which are aperiodic. Now assume that $b+1|n$ and note that this implies $gcd(n+1,b+1)=1$. Suppose we label each necklace bead in the increasing order 1 through n and we rotate each necklace by one position at the time relative to this labelling. At each step we sum mod $(b+1)$ the positions of the b black beads. For each necklace, each residue k, $0 \le k \le b$ is encountered $n/(b+1)$ times. The total number of times each residue k is encountered is thus

$$\frac{1}{b+1}\binom{n}{b} = \frac{1}{n+1}\binom{n+1}{b+1},$$

which as expected equals the number of binary strings of weight b, length n, and the first moment congruent to k mod $(b+1)$ (same for all k).

IV. Multiple repetition case

Fix $r \ge 1$. Let $\mathbf{a} = (a_1, a_2, ..., a_r)$, and consider the set $\hat{S}(m,w,\mathbf{a},p)$ for $w \ge 1$:

$$\begin{aligned}\hat{S}(m,w,\mathbf{a},p) = \{ \quad & \mathbf{s} = (s_1, s_2, ...s_m) \in \{0,1\}^m : \\ & \textstyle\sum_{i=1}^m s_i = w, \\ & b_i = v_i - v_{i-1} - 1 \text{ for } v_i \\ & \text{ the position of the } i^{\text{th}} \text{ '1' in } \mathbf{s}, \\ & \textstyle\sum_{i=1}^{w+1} ib_i \equiv a_1 \bmod p, \\ & \textstyle\sum_{i=1}^{w+1} i^2 b_i \equiv a_2 \bmod p, \\ & \qquad\vdots \\ & \textstyle\sum_{i=1}^{w+1} i^r b_i \equiv a_r \bmod p \ \}.\end{aligned} \tag{23}$$

We say that $b_1, \ldots, b_{w+1}$ denote the sizes of the *bins* of 0's between successive 1's. Here $v_0 = 0$ and $v_{w+1} = m+1$. The set $\hat{S}(m,0,\mathbf{0},p)$ contains just the all-zeros string by convention. Let $\mathbf{a}_0 = \mathbf{0}$ and let $\hat{S}\left(m,(\mathbf{a}_1,p_1),(\mathbf{a}_2,p_2),...,(\mathbf{a}_\mathbf{m},p_m)\right)$ be defined as

$$\hat{S}\left(m,(\mathbf{a}_1,p_1),(\mathbf{a}_2,p_2),...,(\mathbf{a}_\mathbf{m},p_m)\right) = \bigcup_{l=0}^{m}\hat{S}(m,l,\mathbf{a}_\mathbf{l},p_l)., \tag{24}$$

Lemma 3: If each p_l is prime and $p_l > max(r,l)$, the set $\hat{S}\left(m,(\mathbf{a}_1,p_1),(\mathbf{a}_2,p_2),...,(\mathbf{a}_\mathbf{m},p_m)\right)$ is r-insertions of zeros correcting.

Proof: It suffices to show that each set $\hat{S}(m,l,\mathbf{a}_\mathbf{l},p_l)$ is r-insertions of zeros correcting. Suppose a string $\mathbf{x} \in \hat{S}(m,l,\mathbf{a}_\mathbf{l},p_l)$ is transmitted. After experiencing r insertions of zeros, it is received as a string $\mathbf{x}'$. We now show that $\mathbf{x}$ is always uniquely determined from $\mathbf{x}'$.

Let $i_1 \le i_2 \le ... \le i_r$ be the (unknown) indices of the bins of zeros that have experienced insertions. For each j, $1 \le j \le r$, compute $a'_j \equiv \sum_{i=1}^{w+1} i^j b'_i \bmod p_l$, where b'_i is the size of the ith bin of zeros of $\mathbf{x}'$,

$$\begin{aligned} a'_j \quad &\equiv \textstyle\sum_{i=1}^{w+1} i^j b'_i \bmod p_l \\ &\equiv a_j + (i_1^j + i_2^j + ... + i_r^j) \bmod p_l, \end{aligned} \tag{25}$$

where a_j is the jth entry in the residue vector $\mathbf{a}_\mathbf{l}$.

Using Newton's identities over $GF(p_l)$ which relate power sums to symmetric functions of the same variable set, the set $\{i_1, i_2, ..., i_r\}$ is uniquely determined from the set of equations

$$(i_1^j + i_2^j + ... + i_r^j) \equiv a'_j - a_j \bmod p_l$$

for $1 \le j \le r$. For the details of the proof please see [3]. ■

Let $\hat{S}^*\left(m,(\mathbf{a}_1,p_1),(\mathbf{a}_2,p_2),...,(\mathbf{a}_\mathbf{m},p_m)\right)$ be defined as

$$\hat{S}^*\left(m,(\mathbf{a}_1,p_1),(\mathbf{a}_2,p_2),...,(\mathbf{a}_\mathbf{m},p_m)\right) = \bigcup_{l=0}^{m}\hat{S}(m,l,\mathbf{a}_\mathbf{l}{}^*,p_l). \tag{26}$$

where $\hat{S}(m,l,\mathbf{a}_\mathbf{l}{}^*,p_l)$ has the largest cardinality among all sets $\hat{S}(m,l,\mathbf{a}_\mathbf{l},p_l)$ for $\mathbf{a}_\mathbf{l} \in \{0,1,\ldots,p_l-1\}^r$. The cardinality of $\hat{S}(m,l,\mathbf{a}_\mathbf{l}{}^*,p_l)$ is at least

$$\binom{m}{l}\frac{1}{p_l^r} .$$

Since for all n there exists a prime between n and $2n$ it follows that the cardinality of $\hat{S}(m, l, \mathbf{a_l}^*, p_l)$ for $l \geq r$ is at least

$$\binom{m}{l} \frac{1}{(2l)^r} .$$

Thus, the cardinality of $\hat{S}^* (m, (\mathbf{a_1}, p_1), (\mathbf{a_2}, p_2), ..., (\mathbf{a_m}, p_m))$ is at least

$$1 + \sum_{w=1}^{r-1} \binom{m}{w} \frac{1}{(2r)^r} + \sum_{w=r}^{m} \binom{m}{w} \frac{1}{(2w)^r} , \quad (27)$$

which is lower bounded by

$$1 + \frac{1}{(2r)^r} \sum_{w=1}^{r-1} \binom{m}{w} + \frac{1}{(2^r)(m+1)(m+2)\dots(m+r)} \left(2^{m+r} - \sum_{k=0}^{2r-1} \binom{m+r}{k} \right) . \quad (28)$$

The prime counting function $\pi(n)$ which counts the number of primes up to n, satisfies for $n \geq 67$ the inequalities [10]

$$\frac{n}{\ln(n) - 1/2} < \pi(n) < \frac{n}{\ln(n) - 3/2} .$$

With some algebra, it follows that for $n \geq 67$, there exists a prime between n and $(1+\epsilon)n$ for $\epsilon \geq \frac{3}{\ln(n)}$. Thus the lower bound on the asymptotic cardinality of $\hat{S}^* (m, (\mathbf{a_1}, p_1), (\mathbf{a_2}, p_2), ..., (\mathbf{a_m}, p_m))$ can be improved to

$$\frac{1}{(1+\epsilon)^r (m+1)(m+2)\dots(m+r)} \left(2^{m+r}\right) - P(m), \quad (29)$$

where ϵ is an arbitrarily small positive constant and $P(m)$ is a polynomial in m. In the limit $m \to \infty$, (29) is approximately

$$\frac{2^{m+r}}{(m+1)^r} . \quad (30)$$

A construction proposed by Levenshtein [7] has the lower asymptotic bound on the cardinality given by

$$\frac{1}{(\log_2 2r)^r} \frac{2^m}{m^r} . \quad (31)$$

Note that both (27) and the improved bound (29) improve on (31) by at least a constant factor.

The upper bound $U_r(m)$ on any set of strings each of length m capable of overcoming r insertions of zero is

$$U_r(m) = c(r) \frac{2^m}{m^r},$$

as obtained in [7], where

$$c(r) = \begin{cases} 2^r r! & \text{odd } r \\ 8^{r/2}((r/2)!)^2 & \text{even } r \end{cases}$$

which makes the proposed construction be within a factor of this bound. By applying the inverse T_n transformation for $n = m+1$ to $\hat{S}^* (m, (\mathbf{a_1}, p_1), (\mathbf{a_2}, p_2), ..., (\mathbf{a_m}, p_m))$ and noting that both strings under the inverse T_n transformation can simultaneously belong to the repetition error correcting set, we obtain a code of length n capable of overcoming r repetitions, with an asymptotic lower bound on its size being

$$\frac{2^{n+r}}{n^r} . \quad (32)$$

An interesting related problem is that of interactive communication when user 1 owns an uncorrupted copy of a data stream and user 2 owns a corrupted copy of the same data. A method to communicate the minimal number of bits from user 1 to user 2 when the data stream is corrupted by modifying the sizes of the runs of equal symbols is proposed in [9]. In contrast to our model, the model considered in [9] assumes that these communicated bits are transmitted without themselves being subjected to repetition errors.

V. Conclusion

In this paper we discussed the problem of constructing repetition error correcting codes (subsets of binary strings). We presented some explicit number-theoretic constructions and provided some results on the cardinalities of these constructions. Specific contributions included a generalization of a generating function calculation of Sloane [11] and a construction of multiple repetition error correcting codes that is asymptotically a constant factor better than the previously best known construction due to Levenshtein [7].

Acknowledgment

This research was supported in part by NSF award CCF-0635372, Marvell Semiconductor Inc., and the University of California MICRO program.

References

[1] T. M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, NY, 1976.

[2] P. A. H. Bours, "Construction of fixed-length insertion/deletion correcting runlength-limited code", *IEEE Trans. Inform. Th.*, vol. 40(6), pp. 1841–1856, Nov. 1994.

[3] L. Dolecek and V. Anantharam, "A synchonization technique for array-based LDPC codes", *Int. Symp. Info. Th.*, Seattle, WA, July 9-13, 2006.

[4] E. N. Gilbert and J. Riordan, "Symmetry types of periodic sequences", *Ill. Jour. Math.*, Vol. 5, pp. 657–665, 1961.

[5] H. C. Ferreira, W. A. Clarke, A. S. J. Helberg, K. A. S. Abdel-Gaffar and A. J. Han Vinck, "Insertion/deletion correction with spectral nulls", *IEEE Trans. Inform. Th.*, vol. 43(2), pp. 722–732, March 1997.

[6] A. S. J. Helberg and H. C. Ferreira, "On multiple insertion/deletion correcting codes", *IEEE Trans. Inform. Th.* vol. 48(1), pp. 305–308, Jan. 2002.

[7] V. I. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones", *Probl. Inform. Trans.*, vol. 1(1),pp. 8–17, Jan. 1965.

[8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", *Sov. Phys.-Dokl.*, vol. 10(8), pp. 707–710, Feb. 1966.

[9] A. Orlitsky, "Interactive communication of balanced distributions and of correlated files", *SIAM Jour. Disc. Math.*, vol. 6(4), pp. 548–564, Nov. 1993.

[10] J. B. Rosser and L. Schoenfeld, "Approximate formulas for some functions of prime numbers", *Illinois Jour. Math.*, vol. 6(1), pp. 64–94, March 1962.

[11] N. J. A. Sloane, "On single deletion correcting codes", 2000. Available at http://www.research.att.com/~njas/doc/dijen.pdf

[12] T. G. Swart and H.C. Ferreira, "A note on double insertion/deletion correcting codes", *IEEE Trans. Inform. Th.* vol. 49(1), pp. 269–273, Jan. 2003.

[13] R. R. Varshamov and G.M. Tenengolts, "Codes which correct single asymmetric errors", *Avtom. i Telemehk.*, vol. 26(2), pp. 288–292, 1965.