REPLICATION ERROR CORRECTING SETS: EXPLICIT CONSTRUCTIONS AND PREFIXING METHODS

LARA DOLECEK† AND VENKAT ANANTHARAM‡

Abstract. In this paper we study the problem of finding maximally sized subsets of binary strings (codes) of equal length 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. We propose explicit number theoretic constructions of such subsets. In the case of $r = 1$ repetition, the proposed construction is asymptotically optimal. For $r \geq 1$, the proposed construction is within a constant factor of the best known upper bound on the cardinality of a set of strings immune to $r$ repetitions. Inspired by these constructions, we then develop a prefixing method for correcting any prescribed number $r$ of repetition errors in an arbitrary binary linear block code. The proposed method constructs for each string in the given code a carefully chosen prefix such that the resulting strings are all of the same length and such that despite up to any $r$ repetitions in the concatenation of the prefix and the codeword, the original codeword can be recovered. In this construction, the prefix length is made to scale logarithmically with the length of strings in the original code. As a result, the guaranteed immunity to repetition errors is achieved while the added redundancy is asymptotically negligible.

Key words. synchronization error correcting codes, enumeration problems, generating functions, congruencies, residue systems

AMS subject classifications. 94B50, 05A15, 11A07

DOI. 10.1137/080730093

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

$$s(t) = \sum_i c_i h(t - iT) + n(t), \quad (1.1)$$

where $c_i$ is the $i$th bit of $c$, $h(t)$ 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 $c$ (or $x$). 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, an alternative solution is to operate under a poorer timing recovery while oversampling the received waveform in order 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 signal-to-noise ratio limit of a pulse amplitude modulation (PAM) system, it 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. In particular, we develop explicit number-theoretic constructions of sets of binary strings immune to multiple repetitions and provide results on their cardinalities. We then use these constructions to develop a prefixing method which transforms a given set of binary strings into another set that itself satisfies number-theoretic constraints of the proposed constructions. The redundancy introduced by this carefully chosen prefix is shown to be logarithmic in the length of the strings in the given set.

The remainder of the paper is organized as follows. In section 2 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 3 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 4 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 [4]. Inspired by these number-theoretic constructions, in section 5 we develop a general prefixing-based method which injectively converts a given set of binary strings of the same length into another set such that the resulting set is immune to a prescribed number of repetition errors. The method produces for each string in the original set a carefully chosen prefix such that the result of the concatenation of the prefix and this string satisfies number-theoretic congruential constraints previously developed in section 4 (where these constraints were shown to be sufficient to provide immunity to repetition errors). The prefix length in the proposed method is shown to scale logarithmically with the length of the strings in the original given set. Thus, the proposed construction guarantees immunity to a prescribed number of repetition errors, while the incurred redundancy becomes asymptotically negligible.

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

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

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


Now, the repetition in \( c \) in position \( p \) corresponds to the insertion of “0” in position \( p-1 \) in \( \tilde{c} \), and \( \text{weight}(\tilde{c}) \) is the number of runs in \( c - 1 \). We let \( \tilde{C} \) be the collection of strings of length \( n - 1 \) obtained by applying \( T_n \) to all strings \( C \). Note that \( c \) 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 \( n \).

Since the strings starting with runs of different types cannot be confused under repetition errors, both preimages under \( T_n \) may be included in such a code immune to repetition errors.

3. Single repetition error correcting set. Following the analysis of Sloane [9] and Levenshtein [5] of the related so-called Varshamov–Tenengolts codes [11] 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 with \( w \) ones (1's) for \( 0 \leq w \leq m \). Partition \( A_w^m \) based on the value of the first moment of each string. More specifically, let \( S_w,k \) be the subset of \( A_w^m \) such that

\[
S_{w,k}^{m,t} = \left\{ (s_1, s_2, \ldots, s_m) \left| \sum_{i=1}^{m} i \times s_i \equiv k \mod t \right. \right\}.
\]

In the subsequent analysis we say that an element of \( S_{w,k}^{m,t} \) has the first moment congruent to \( k \) mod \( t \).

**Lemma 3.1.** Each subset \( S_{w,k}^{m,w+1} \) is a single “0”-insertion correcting code.

**Proof.** Suppose the string \( s' \) is received. We want to uniquely determine the codeword \( s = (s_1, s_2, \ldots, s_m) \in S_{w,k}^{m,w+1} \) such that \( s' \) is the result of inserting at most one zero in \( s \).

If the length of \( s' \) is \( m \), conclude that no insertion occurred and that \( s = s' \).

If the length of \( s' \) is \( m+1 \), a zero has been inserted. For \( s' = (s'_1, s'_2, \ldots, s'_m, s'_{m+1}) \), compute \( i \times s'_i \mod (w+1) \). 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 \( \sum_{i=1}^{m+1} i \times s'_i \mod (w+1) \). If \( k' = k \), the insertion occurred after the rightmost one, so we declare \( s \) to be the \( m \) leftmost bits in \( s' \). If \( k' > k \), \( R_1 \) is equal to \( k' - k \), and we declare \( s \) to be the string obtained by deleting the zero immediately preceding the rightmost \( k' - k \) ones. Finally, if \( k' < k \), \( R_1 \) is \( w+1-k+k' \), and we declare \( s \) to be the string obtained by deleting the zero immediately preceding the rightmost \( w+1-k+k' \) ones.

A related, number- and group-theoretic–based construction of a code for correcting a single error and for detecting all unidirectional (asymmetric) errors is presented in [6].

3.1. Cardinality results. Since \( |A_w^m| = \binom{m}{w} \), there exists \( k \) such that

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

Since two codewords of different weights cannot result in the same string when at most one zero is inserted, we may let \( \tilde{C} \) be the union of the largest sets \( S_{w,k}^{m,w+1} \) over
different weights \( w \), i.e.,

\[
\tilde{C} = \bigcup_{w=1}^{m} S_{w,k_{w}^*}^{m,w+1},
\]

where \( S_{w,k_{w}^*}^{m,w+1} \) is the set of the largest cardinality among all sets \( S_{w,k}^{m,w+1} \) for \( 0 \leq k \leq w \). 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 [4] to be

\[
U_1(m) = \frac{2^{m+1}}{m}.
\]

Hence the proposed construction is asymptotically optimal in the sense that the ratio of its cardinality to the largest possible cardinality approaches 1 as \( m \to \infty \).

By applying inverse \( T_n \) transformation for \( n = m + 1 \) to \( \tilde{C} \) and noting that both preimages under \( T_n \) can simultaneously belong to a repetition correcting set, we obtain a code of length \( n \) and of size at least

\[
\frac{1}{n} \left( 2^n - 2 \right)
\]

capable of correcting one repetition.

The cardinalities of the sets \( S_{w,k}^{m,w+1} \) may be computed explicitly as we now show.

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

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

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 3.2.** Let \( g = \gcd(m + 1, w + 1) \). The cardinality of \( S_{w,k}^{m,w+1} \) is

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

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 [9] of the Varshamov–Tenengolts codes [11], let us introduce the function \( f_{b,n}(U,V) \) in which the coefficient of \( U^sV^k \) (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 \( b \) (i.e., \( g_{k,s}^{b}(n) = |S_{n,k}^{s,b}| \)),

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

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

\[
f_{b,n}(U,V) = \prod_{t=1}^{n} (1 + U^tV^t) \mod (V^b - 1).
\]
Let $a = e^{i\frac{2\pi}{b}}$ so that for $V = a^j$

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

By inverting this expression we can write

\begin{equation}
(3.11) \quad \sum_{s=0}^{n} g_{k,s}^b(n) U^s = \frac{1}{b} \sum_{j=1}^{b-1} f_{b,n}(U, e^{i\frac{2\pi j}{b}}) e^{-i\frac{2\pi s j}{b}} = \frac{1}{b} \sum_{j=1}^{b-1} \prod_{t=1}^{n} (1 + U e^{i\frac{2\pi j}{b}}) e^{-i\frac{2\pi s j}{b}}.
\end{equation}

Our next goal is to evaluate the coefficient $U^b$ on the right-hand side in (3.11). To do so we first evaluate the following expression:

\begin{equation}
(3.12) \quad \prod_{t=1}^{b} (1 + U e^{i\frac{2\pi j}{b}}).
\end{equation}

Let $d_j = b / \text{gcd}(b, j)$ and $s_j = j / \text{gcd}(b, j)$, and write

\begin{equation}
(3.13) \quad \prod_{t=1}^{b} (1 + U e^{i\frac{2\pi j}{b}}) = \left( \prod_{t=1}^{d_j} (1 + U e^{i\frac{2\pi j}{s_j}}) \right)^{\text{gcd}(b, j)} = 1 + U \sum_{t_1=1}^{d_j} e^{i\frac{2\pi s_j}{s_j}} + U^2 \sum_{t_2=t_1+1}^{d_j} \sum_{t_3=t_2}^{d_j} e^{i\frac{2\pi s_j(t_3-t_2)}{s_j}} + \cdots + U^{d_j} e^{i\frac{2\pi s_j}{s_j} (1+2+\cdots+d_j)}.
\end{equation}

Since $\text{gcd}(d_j, s_j) = 1$, the set

\begin{equation}
(3.14) \quad V = \{ e^{i\frac{2\pi s_j}{s_j}}, e^{i\frac{2\pi s_j}{s_j}} e^{i\frac{2\pi s_j}{s_j}}, \ldots, e^{i\frac{2\pi s_j}{s_j} d_j} \}
\end{equation}

represents all distinct solutions of the equation

\begin{equation}
(3.15) \quad x^{d_j} - 1 = 0.
\end{equation}

For a polynomial equation $P(x)$ of degree $d$, the coefficient multiplying $x^k$ is a scaled symmetric function of $d - k$ roots. Hence, by (3.15), 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,

\begin{equation}
(3.16) \quad \prod_{t=1}^{b} (1 + U e^{i\frac{2\pi j}{b}}) = (1 + (-1)^{1+d_j} U^{d_j})^{\text{gcd}(b, j)}.
\end{equation}

Returning to the inner product in (3.11), let us first suppose that $b | n$. Then

\begin{equation}
(3.17) \quad \prod_{t=1}^{n} (1 + U e^{i\frac{2\pi j}{b}}) = \left( \prod_{t=1}^{b} (1 + U e^{i\frac{2\pi j}{b}}) \right)^{n/b} = (1 + (-1)^{1+d_j} U^{d_j})^{\text{gcd}(b, n)/b} = \sum_{d_j=0}^{n} \left( \frac{n}{d_j} \right) (-1)^{l(1+d_j)} U^{ld_j}.
\end{equation}
Thus (3.11) becomes
\begin{equation}
\sum_{s=0}^{n} g_{b,s}^b(n) U^s = \frac{1}{b}\sum_{d|b} \sum_{l=0}^{\frac{n}{d}} \left( \frac{n}{d} \right) (-1)^l (1+d) U^d l e^{-i \frac{2\pi j k}{d}}.
\end{equation}

We now regroup the terms whose \(j\)'s yield the same \(d_j\)'s:
\begin{equation}
\sum_{s=0}^{n} g_{b,s}^b(n) U^s = \frac{1}{b}\sum_{d|b} \sum_{l=0}^{\frac{n}{d}} \left( \frac{n}{d} \right) (-1)^l (1+d) U^d l e^{-i \frac{2\pi j k}{d}}.
\end{equation}

The rightmost sum can also be written as
\begin{equation}
\sum_{j: \gcd(j,b)=b/d, 0 \leq j \leq b-1} e^{-i \frac{2\pi j k}{d}} = \sum_{s:0 \leq s \leq d-1, \gcd(s,d)=1} e^{-i \frac{2\pi s k}{d}}.
\end{equation}

This last expression is known as the Ramanujan sum \([1]\) and simplifies to
\begin{equation}
\mu\left( \frac{d}{\gcd(d,k)} \right)
\end{equation}

Now the coefficient of \(U^b\) in (3.11) is
\begin{equation}
\frac{1}{b}\sum_{d|b} \left( \frac{n}{d} \right) (-1)^l (1+d) \phi(d) \mu\left( \frac{d}{\gcd(d,k)} \right),
\end{equation}

which is precisely the number of strings of length \(n\), weight \(b\), and the first moment congruent to \(k\) mod \(b\), i.e., \(|S_{b,k}^{n,b}|\).

Consider the set of strings described by \(S_{w,k}^{m,w+1}\) for \(m = n - 1\) and \(w = b - 1\), i.e., \(S_{w,k}^{m,w+1} = S_{b-1,k}^{n-1,b}\). If we append “1” to each such string, we would obtain a fraction of \(b/n\) of all strings that belong to the set \(S_{b,k}^{n,b}\). To see why this is true, first note that the cardinality of the set \(S_{b-1,k}^{n-1,b}\) and of the subset \(T_{b,k}^n\) of \(S_{b,k}^{n,b}\) which contains all strings ending in “1” is the same (since, when a “1” is appended to each element of the set \(S_{b-1,k}^{n-1,b}\), the resulting set contains strings of length \(n\), weight \(b\), and first moment congruent to \((k + n)\) mod \(b\), which is also congruent to \(k\) mod \(b\) since by assumption \(b|n\)). It is thus sufficient to show that \(|T_{b,k}^n| = \frac{1}{b}|S_{b,k}^{n,b}|\). Let \(A_k = |S_{b,k}^{n,b}|\). Write \(A_k = \sum_{u|b} A_k(n, b, \frac{u}{n})\), where \(A_k(n, b, v)\) denotes the number of strings of length \(n\), weight \(b\), first moment congruent to \(k\) mod \(b\), and with period \(v\). Consider a string accounted for in \(A_k(n, b, \frac{u}{n})\). Its single cyclic shift has the first moment congruent to \((k + b)\) mod \(b\) and is thus also accounted for in \(A_k(n, b, \frac{u}{n})\). Since \(\frac{u}{n}\) is the period, and since \(\frac{b}{u}\) is the weight per period, fraction \(\frac{b}{n/u}\) of \(A_k(n, b, \frac{u}{n})\) represents distinct strings that end in “1,” have length \(n\), weight \(b\), first moment congruent to \(k\) mod \(b\), and period \(\frac{u}{n}\). Thus, \(|T_{b,k}^n| = \sum_{u|b} \frac{b}{n/u} A_k(n, b, \frac{u}{n}) = \frac{1}{b} A_k\), as required.

Therefore, the cardinality of \(S_{w,k}^{m,w+1}\) is \(b/n\) times the expression in (3.22),
\begin{equation}
|S_{w,k}^{m,w+1}| = \frac{1}{m+1} \sum_{d|w+1} \left( \frac{m+1}{w+1} \right) (-1)^{\frac{w+1}{d}(1+d)} \phi(d) \mu\left( \frac{d}{\gcd(d,k)} \right).
\end{equation}
Notice that the last expression is the same as the one proposed in Lemma 3.2 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 (3.9), where \( g = \gcd(n,b) \), and get
\[
(3.24) \quad \sum_{s=0}^{n} g_{n,s}^g(n)U^s = \frac{1}{g} \sum_{d|g} \sum_{l=0}^{\frac{n}{d}} \left( \frac{n}{d} \right) \frac{\binom{d}{l}}{\phi(g)} \left( -1 \right)^{l(1+d)} U^{dl} \times \sum_{j: \gcd(j,g)=g/\delta,0 \leq j \leq g-1} e^{-\frac{2\pi i jk}{g}}.
\]

Thus the coefficient of \( U^b \) here is
\[
(3.25) \quad \frac{1}{g} \sum_{d|g} \left( \frac{n}{d} \right) \frac{\binom{d}{1}}{\phi(g)} \left( -1 \right)^{\frac{b}{d}(1+d)} \phi(d) \frac{\mu(d)}{\gcd(d,k)} \frac{d}{\phi(d,k)}.
\]

This is the number of strings of length \( n \), weight \( b \), and the first moment congruent to \( k \) mod \( g \); namely, it is the cardinality of the set \( S_{b,k}^g \). Let \( B_k = |S_{b,k}^g| \). Write \( B_k = \sum_{u,v} B_k(n,b,\frac{u}{v}) \), where \( B_k(n,b,\frac{u}{v}) \) denotes the number of strings of length \( n \), weight \( b \), first moment congruent to \( k \) mod \( g \), and with period \( v \). By cyclically shifting a string of length \( n \), weight \( b \), first moment congruent to \( k \) mod \( g \), and with period \( u \) and \( v \), we have the first moment congruent to \( k \) mod \( g \), it follows that a fraction \( \frac{b}{b/u} \) of \( B_k(n,b,\frac{u}{v}) \) represents the number of strings that end in \( "1" \) have length \( n \), weight \( b \), first moment congruent to \( k \) mod \( g \), and period \( \frac{u}{v} \). Thus a fraction \( b/v \) of \( B_k \) denotes the number of strings that end in \( "1" \) are of length \( n \), weight \( b \), and have the first moment congruent to \( k \) mod \( g \). Since each string of length \( n-1 \), weight \( b-1 \), and the first moment congruent to \( k \) mod \( g \) produces a unique string that ends in \( "1" \), is of length \( n \), weight \( b \), and has the first moment congruent to \( k \) mod \( g \) by appending \( "1" \), it follows that \( \frac{b}{n} B_k \) is also the number of strings of length \( n-1 \), weight \( b-1 \), and the first moment congruent to \( k \) mod \( g \). Thus the number of strings given by \( S_{b-1,k}^g \) is also \( \frac{b}{n} B_k \).

Consider again cyclic shifts of a string of length \( n \), weight \( b \), the first moment congruent to \( k \) mod \( g \) and with period \( n/u \). A fraction \( b/u \) of these shifts produce strings with a \( "1" \) in the last position. Let us consider one such string \( s_0 \). Its first \( n-1 \) bits correspond to a string of length \( n-1 \), weight \( b-1 \), and the first moment congruent to \( k \) mod \( g \). This \( n-1 \)-bit string has the first moment congruent to \( k_0 \) mod \( b \) for some \( k_0 \). Cyclically shift \( s_0 \) for \( t_1 \) places until the first time \( "1" \) again appears in the \( n \)th position, and call the resulting string \( s_1 \). (Since \( b > g \) and \( u/g, b/u > 1 \), and thus \( s_1 \neq s_0 \).) The first \( n-1 \) bits of \( s_1 \) correspond to a string of length \( n-1 \), weight \( b-1 \), and the first moment congruent to \( k_1 \equiv k_0 + t_1 (b-1) + t_1 - \mod g \equiv k_0 + t_1 b - n \mod b \equiv k_0 - gy \mod b \), where \( y = \frac{n}{\phi(g)} \). Cyclically shift \( s_1 \) for \( t_2 \) places until the first time \( "1" \) again appears in the \( n \)th position, and call the resulting string \( s_2 \). The first \( n-1 \) bits of \( s_2 \) correspond to a string of length \( n-1 \), weight \( b-1 \), and the first moment congruent to \( k_2 \equiv k_0 - gy + t_2 (b-1) + t_2 - n \mod g \equiv k_0 - gy + t_2 b - n \mod b \equiv k_0 - 2gy \mod b \). Each successive cyclic shift with \( "1" \) in the last place gives a string \( s_i \) whose first \( n-1 \) bits have the first moment congruent to \( k_i \equiv k_0 - igy \mod b \). The last such string, \( s_{b/u-1} \), before the string \( s_0 \) is encountered again, has the left \( n-1 \) bit substring whose first moment is congruent to \( k_{b/u-1} \equiv k_0 - (\frac{b}{u} - 1)gy \mod b \). Note that the sequence \( \{k_0, k_1, k_2, \ldots, k_{b/u-1}\} \) is periodic with period \( z \) (here \( \gcd(y,g) = 1 \) by construction), where \( z = \frac{b}{g} \). Since \( z|u \), each of \( k_0, k_1 \) through \( k_{\frac{b}{g}-1} \) appears an equal number of times in this sequence.
Consequently, the number of strings in the set \( S_{b-1,k}^{m-1,b} \) is \( \frac{n}{k} \) of the size of the set \( S_{b-1,k}^{n-1,g} \) for every \( k_i \equiv ig + k \mod b \).

Therefore, \( |S_{w,k}^{m,w+1}| \) is

\[
|S_{w,k}^{m,w+1}| = \frac{k}{n} \frac{g}{b} |S_{b,k}^{n,g}|
\]

which completes the proof of the lemma. \( \square \)

### 3.2. Connection with necklaces

It is interesting to briefly visit the relationship between optimal single insertion of a zero correcting code and combinatorial objects known as necklaces [2].

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 [2]

\[
T(n) = \frac{1}{n} \sum_{d|gcd(n,b)} \left( \frac{\varphi(d)}{d} \right) \phi(d).
\]

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

\[
\frac{1}{b+1} \binom{n}{b} = \frac{1}{n+1} \binom{n+1}{b+1},
\]

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 \( k \)).

### 4. Multiple repetition error correcting set

We now present an explicit construction of a multiple repetition error correcting set and discuss its cardinality.

Let \( a = (a_1, a_2, \ldots, a_r) \) for \( r \geq 1 \), and consider the set \( \hat{S}(m,w,a,p) \) for \( w \geq 1 \) defined as

\[
\hat{S}(m,w,a,p) = \{s = (s_1, s_2, \ldots, s_m) \in \{0,1\}^m: v_0 = 0, v_{w+1} = m + 1, \text{ and } v_i \text{ is the position of the } i\text{th } 1 \text{ in } s \text{ for } 1 \leq i \leq w, \]

\[
b_i = v_i - v_{i-1} - 1 \text{ for } 1 \leq i \leq w + 1, \]

\[
\sum_{i=1}^{w+1} s_i = w, \]

\[
\sum_{i=1}^{w+1} ib_i \equiv a_1 \mod p, \]

\[
\sum_{i=1}^{w+1} i^2b_i \equiv a_2 \mod p, \]

\[
\vdots \]

\[
\sum_{i=1}^{w+1} i^rb_i \equiv a_r \mod p \}. \]
The set $\hat{S}(m, 0, 0, p)$ contains just the all-zeros string. Let $a_0 = 0$, and let $\hat{S}(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m))$ be defined as

$$
\hat{S}(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) = \bigcup_{l=0}^{m} \hat{S}(m, l, a_1, p_1),
$$

where $b_1, \ldots, b_{w+1}$ denote the sizes of the bins of 0’s between successive 1’s.

**Lemma 4.1.** If each $p_i$ is prime and $p_i > \max(r, l)$, the set $\hat{S}(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m))$, provided it is nonempty, is $r$-insertions of zeros correcting.

**Proof.** It suffices to show that each nonempty set $\hat{S}(m, l, a_1, p_1)$ is $r$-insertions of zeros correcting. This is obvious for $l = 0$. For $l > 0$ suppose a string $x \in \hat{S}(m, l, a_1, p_1)$ is transmitted. After experiencing $r$ insertions of zeros, it is received as a string $x'$. We now show that $x$ is always uniquely determined from $x'$.

Let $i_1 \leq i_2 \leq \cdots \leq i_r$ be the (unknown) indices of the bins of zeros that have experienced insertions. For each $j$, $1 \leq j \leq r$, compute $a_j' \equiv \sum_{i=1}^{w+1} i' b_i' \mod p_i$, where $b_i'$ is the size of the $i$th bin of zeros of $x'$,

$$
a_j' = \sum_{i=1}^{w+1} i' b_i' \mod p_i = a_j + (i_1' + i_2' + \cdots + i_r') \mod p_i,
$$

where $a_j$ is the $j$th entry in the residue vector $a_1$ (to lighten the notation the subscript $l$ in $a_i$ is omitted).

By collecting the resulting expressions over all $j$ and setting $a_j'' \equiv a_j' - a_j \mod p_i$, we arrive at

$$
E_r = \left\{ \begin{array}{l}
  a_1'' \equiv i_1 + i_2 + \cdots + i_r \mod p_1, \\
  a_2'' \equiv i_1^2 + i_2^2 + \cdots + i_r^2 \mod p_1, \\
  \vdots \\
  a_r'' \equiv i_1^r + i_2^r + \cdots + i_r^r \mod p_1.
\end{array} \right.
$$

The terms on the right-hand side of the congruency constraints are known as power sums in $r$ variables. Let $S_k$ denote the $k$th power sum mod $p_i$ of $\{i_1, i_2, \ldots, i_r\}$, and let $\Lambda_k$ denote the $k$th elementary symmetric function of $\{i_1, i_2, \ldots, i_r\}$ mod $p_i$,

$$
S_k \equiv i_1^k + i_2^k + \cdots + i_r^k \mod p_i,
$$

and

$$
\Lambda_k \equiv \sum_{v_1 < v_2 < \cdots < v_k} i_{v_1} i_{v_2} \cdots i_{v_k} \mod p_i.
$$

Using Newton’s identities over $GF(p_i)$, which relate power sums to symmetric functions of the same variable set and are of the type

$$
S_k - \Lambda_1 S_{k-1} + \Lambda_2 S_{k-2} - \cdots + (-1)^{k-1} \Lambda_{k-1} S_1 + (-1)^k \Lambda_k = 0
$$

for $k \leq r$, we can obtain an equivalent system of $r$ equations:

$$
\bar{E}_r = \left\{ \begin{array}{l}
  d_1 \equiv \sum_{j=1}^{r} i_j \mod p_1, \\
  d_2 \equiv \sum_{j<k} i_j i_k \mod p_1, \\
  \vdots \\
  d_r \equiv \prod_{j=1}^{r} i_j \mod p_1.
\end{array} \right.
$$
where each residue \( d_k \) is computed recursively from \( \{d_1, \ldots, d_{k-1}\} \) and \( \{a''_1, a''_2, \ldots, a''_r\} \). Specifically, since the largest coefficient in (4.7) is \( r \), and \( r < p_t \) by construction, the last term in (4.7) never vanishes due to the multiplication by the coefficient \( k \).

Consider now the equation

\[
(4.9) \quad \prod_{j=1}^{r} (x - i_j) \equiv 0 \mod p_t,
\]

and expand it into the standard form

\[
(4.10) \quad x^r + c_{r-1}x^{r-1} + \cdots + c_1x + c_0 \equiv 0 \mod p_t.
\]

By collecting the same terms in (4.9) and (4.10), it follows that \( d_k \equiv (-1)^k c_{r-k} \) mod \( p_t \). Furthermore, by the Lagrange theorem, (4.10) has at most \( r \) solutions. Since \( i_r \leq p_t \), all incongruent solutions are distinguishable, and thus the solution set of (4.10) is precisely the set \( \{i_1, i_2, \ldots, i_r\} \).

Therefore, since the system \( E_r \) of \( r \) equations uniquely determines the set \( \{i_1, i_2, \ldots, i_r\} \), the locations of the inserted zeros (up to the position within the bin they were inserted in) are uniquely determined, and thus \( x' \) is always uniquely recovered from \( x' \).

Hence, \( \hat{S}(m,(a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) is \( r \)-insertions of zeros correcting for \( p_t \) prime and \( p_t > \max(r, l) \).

**Remark.** In fact, when nonempty, the set \( \hat{S}(m,(a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) in (4.2) is \( u \)-insertions of zeros correcting for any \( u \) up to \( r \), since each nonempty constituent set \( \hat{S}(m,(a_1, p_1)) \) is itself \( u \)-insertions of zeros correcting. Note that under the current set-up the actual number of insertions, \( u \), can be inferred from the length of the received string, which is \( m + u \). It then suffices to set up a system of equations \( E_u \) consisting of the first \( u \) power sums of the unknown indices, as in (4.3) and (4.4). While the collection in (4.1) is defined in terms of \( r \) congruential constraints, the remaining \( r - u \) power sums are redundant for determining these \( u \) unknown indices.

With \( r \) replaced by \( u \), the rest of the proof is as in Lemma 4.1.

In particular, for \( r = 1 \), the constructions in (3.1) and (4.1) are related as follows.

**Lemma 4.2.** For prime and \( p > w \), the set \( S_{w,p}^m \) defined in (3.1) equals the set \( \hat{S}(m,(a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) in (4.2) is \( u \)-insertions of zeros correcting for any \( u \) up to \( r \), since each nonempty constituent set \( \hat{S}(m,(a_1, p_1)) \) is itself \( u \)-insertions of zeros correcting. Note that under the current set-up the actual number of insertions, \( u \), can be inferred from the length of the received string, which is \( m + u \). It then suffices to set up a system of equations \( E_u \) consisting of the first \( u \) power sums of the unknown indices, as in (4.3) and (4.4). While the collection in (4.1) is defined in terms of \( r \) congruential constraints, the remaining \( r - u \) power sums are redundant for determining these \( u \) unknown indices.

With \( r \) replaced by \( u \), the rest of the proof is as in Lemma 4.1.

In particular, for \( r = 1 \), the constructions in (3.1) and (4.1) are related as follows.

For prime and \( p > w \), the set \( S_{w,p}^m \) defined in (3.1) equals the set \( \hat{S}(m,(a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) in (4.2) is \( u \)-insertions of zeros correcting for any \( u \) up to \( r \), since each nonempty constituent set \( \hat{S}(m,(a_1, p_1)) \) is itself \( u \)-insertions of zeros correcting. Note that under the current set-up the actual number of insertions, \( u \), can be inferred from the length of the received string, which is \( m + u \). It then suffices to set up a system of equations \( E_u \) consisting of the first \( u \) power sums of the unknown indices, as in (4.3) and (4.4). While the collection in (4.1) is defined in terms of \( r \) congruential constraints, the remaining \( r - u \) power sums are redundant for determining these \( u \) unknown indices.

With \( r \) replaced by \( u \), the rest of the proof is as in Lemma 4.1.

In particular, for \( r = 1 \), the constructions in (3.1) and (4.1) are related as follows.

**Lemma 4.2.** For prime and \( p > w \), the set \( S_{w,p}^m \) defined in (3.1) equals the set \( \hat{S}(m,(a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) in (4.2) is \( u \)-insertions of zeros correcting for any \( u \) up to \( r \), since each nonempty constituent set \( \hat{S}(m,(a_1, p_1)) \) is itself \( u \)-insertions of zeros correcting. Note that under the current set-up the actual number of insertions, \( u \), can be inferred from the length of the received string, which is \( m + u \). It then suffices to set up a system of equations \( E_u \) consisting of the first \( u \) power sums of the unknown indices, as in (4.3) and (4.4). While the collection in (4.1) is defined in terms of \( r \) congruential constraints, the remaining \( r - u \) power sums are redundant for determining these \( u \) unknown indices.

With \( r \) replaced by \( u \), the rest of the proof is as in Lemma 4.1.

In particular, for \( r = 1 \), the constructions in (3.1) and (4.1) are related as follows.
proof to hold. We can therefore replace each $i$ in (4.1) with the weighting $f_i$ with the property that each $f_i$ is a residue mod $p$ and that $f_i \neq f_j$ for $i \neq j$. Let $\hat{S}(m, w, a, f, p)$ for $w \geq 1$ be defined as

$$\hat{S}(m, w, a, f, p) = \{s = (s_1, s_2, \ldots, s_m) \in \{0, 1\}^m : v_0 = 0, v_{w+1} = m + 1, \text{and}$$
$$v_i \text{ is the position of the } i\text{th } 1 \text{ in } s \text{ for } 1 \leq i \leq w, \quad b_i = v_i - v_{i-1} - 1 \text{ for } 1 \leq i \leq w + 1, \quad \sum_{i=1}^{w+1} s_i = w,$$
$$f_i \mod p \neq f_j \mod p \text{ for } i \neq j,$$
$$\sum_{i=1}^{w+1} f_i b_i \equiv a_1 \mod p,$$
$$\sum_{i=1}^{w+1} (f_i)^2 b_i \equiv a_2 \mod p,$$
$$\vdots$$
$$\sum_{i=1}^{w+1} (f_i)^r b_i \equiv a_r \mod p\}.$$ (4.12)

The set $\hat{S}(m, 0, 0, 0, p)$ contains just the all-zeros string. Let $a_0 = 0$, and let $\hat{S}(m, (a_1, f_1, p_1), (a_2, f_2, p_2), \ldots, (a_m, f_m, p_m))$ be defined as

$$\hat{S}(m, (a_1, f_1, p_1), (a_2, f_2, p_2), \ldots, (a_m, f_m, p_m)) = \bigcup_{l=0}^{w} \hat{S}(m, l, a_1, f_1, p_1).$$ (4.13)

We note that $\hat{S}(m, w, a, f, p) = \hat{S}(m, w, a, p)$ when $f = (1, 2, \ldots, (w + 1))$.

**Lemma 4.3.** If each $p_i$ is prime and $p_i > \max(r, l)$, the set $\hat{S}(m, (a_1, f_1, p_1), (a_2, f_2, p_2), \ldots, (a_m, f_m, p_m))$ is $r$-insertions of zeros correcting.

**Proof.** The proof follows that of Lemma 4.1 with appropriate substitutions of $f_i$ for $i$.

The object $\hat{S}(m, w, a, f, p)$ will be of further interest to us in section 5.2 when we discuss a prefixing method for improved immunity to repetition errors.

We now present some cardinality results for the construction of present interest. For simplicity we focus on the set $\hat{S}(m, w, a, p)$ as the results hold verbatim for $\hat{S}(m, w, a, f, p)$ with appropriate weighting assignments.

**4.1. Cardinality results.** Let $\hat{S}^*(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m))$ be defined as

$$\hat{S}^*(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) = \bigcup_{l=0}^{w} \hat{S}(m, l, a_1^*, p_1),$$ (4.14)

where $\hat{S}(m, l, a_1^*, p_1)$ is the largest among all sets $\hat{S}(m, l, a_1, p_1)$ for $a_1 \in \{0, 1, \ldots, p_1\}^r$. The cardinality of $\hat{S}(m, l, a_1^*, p_1)$ is at least

$$\binom{m}{l} \frac{1}{p_1^r}.$$ (4.15)

Since for all $n$ there exists a prime between $n$ and $2n$, it follows that one can choose the $p_i$, $1 \leq l \leq m$, so that cardinality of $\hat{S}(m, l, a_1^*, p_1)$ for $l \geq r$ is at least

$$\binom{m}{l} \frac{1}{(2l)^r},$$ (4.16)
Thus $p_1, \ldots, p_m$ can be chosen so that the cardinality of $\hat{S}^*(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m))$ is at least

\begin{equation}
1 + \sum_{w=1}^{r-1} \binom{m}{w} \frac{1}{(2^r)^r} + \sum_{w=r}^{m} \frac{1}{(2^r)^r},
\end{equation}

which is lower bounded by

\begin{equation}
1 + \frac{1}{(2^r)^r} \sum_{w=1}^{r-1} \binom{m}{w} + \frac{1}{(2^r)^r} \left( 2^{m+r} - \sum_{k=0}^{2r-1} \binom{m+r}{k} \right).
\end{equation}

The prime counting function $\pi(n)$, which counts the number of primes up to $n$, satisfies for $n \geq 67$ the inequalities [8]

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

From (4.19) it follows that

\begin{equation}
\frac{(1+\epsilon)n}{\ln((1+\epsilon)n) - 1/2} < \pi((1+\epsilon)n) < \frac{(1+\epsilon)n}{\ln((1+\epsilon)n) - 3/2}.
\end{equation}

For a prime number to exist between $n$ and $(1+\epsilon)n$, it is sufficient to have

\begin{equation}
\pi((1+\epsilon)n) > \pi(n).
\end{equation}

Using (4.19) and (4.20), it is sufficient to have

\begin{equation}
\pi((1+\epsilon)n) > \frac{(1+\epsilon)n}{\ln((1+\epsilon)n) - 1/2} \geq \frac{n}{\ln(n) - 3/2} > \pi(n).
\end{equation}

Comparing the innermost terms in (4.22), it follows that it is sufficient for $\epsilon$ to satisfy

\begin{equation}
\epsilon \ln(n) \geq \ln(1+\epsilon) + \frac{3\epsilon}{2} + 1
\end{equation}

for (4.21) to hold.

For $n \geq 67$ and $\epsilon = \frac{3}{\ln(n)}$, the left-hand side of (4.1) evaluates to 3 while the right-hand side of (4.1) is upper bounded by $(0.539 + 1.071 + 1) < 3$. Since $\pi(n)$ is a nondecreasing function of $n$, 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 the best choice over $p_1, \ldots, p_m$ of $\hat{S}^*(m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m))$ can be improved to

\begin{equation}
\frac{1}{(1+\epsilon)^r(m+1)(m+2)\ldots(m+r)} (2^{m+r}) - P(m),
\end{equation}

where $\epsilon = \frac{3}{\ln m}$ and $P(m)$ is a polynomial in $m$. In the limit $m \to \infty$, (4.24) is approximately

\begin{equation}
\frac{2^{m+r}}{(m+1)^r}.
\end{equation}
A construction proposed by Levenshtein [4] has the lower asymptotic bound on the cardinality given by

\[ \frac{1}{(\log_2 2^r)^r} m^r. \]  

(4.26)

Note that both (4.17) and the improved bound (4.24) improve on (4.26) 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 [4], where

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

(4.28)

which makes the proposed construction be within a factor of this bound. By applying the inverse \( T_n \) transformation for \( n = m + 1 \) to \( \hat{S}^* \) \((m, (a_1, p_1), (a_2, p_2), \ldots, (a_m, p_m)) \) 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 and of asymptotic size at least

\[ \frac{2^n + r}{n^r}. \]  

(4.29)

5. Prefixing-based method for multiple repetition error correction. In this section we develop a general prefixing method which injectively transforms a given collection \( S \) of binary strings of length \( n \) into another collection \( T_S \) of binary strings of equal length, such that the collection \( T_S \) is guaranteed to be immune to the prescribed number of repetition errors. The proposed method is inspired by the number-theoretic construction developed in the previous section. Given an element \( s \) of \( S \), a string \( t_s = [p_s s] \), \( t_s \in T_S \), is created; that is, the prefix \( p_s \) is prepended to \( s \) to produce \( t_s \), such that the string \( t_s \) under transformation (2.1) satisfies the set of conditions given by (4.12). In the proposed method, the set \( T_S \) has the property that the length of the prefix \( p_s \) is \( \Theta(\log(n)) \). Thus, if the set \( S \) is used for transmission, the proposed method provides increased immunity to repetition errors with asymptotically vanishing loss in the rate.

We start with some auxiliary results.

5.1. Auxiliary results. Consider a prime number \( P \) with the property that \( \text{lcm}(2, 3, \ldots, r) \mid (P - 1) \) for a given positive integer \( r \). Since each \( i, 1 \leq i \leq r \), satisfies \( i \mid (P - 1) \), it follows that in the residue set mod \( P \), there are \( \frac{P - 1}{i} \) elements that are \( i \)th power residues, each having \( i \) distinct roots (an \( i \)th power residue \( x \) satisfies \( y^i \equiv x \mod P \) for some \( y \)) [1]. For convenience, let \( G = \lfloor \log_2(P) \rfloor \).

For each \( i, 1 \leq i \leq r \), we will construct a specific subset \( V_i \) of the \( i \)th power residues mod \( P \) such that all other residues can be expressed as a sum of a subset of elements of \( V_i \), and such that each \( V_i \) has size that is logarithmic in \( P \). The set of the \( i \)th roots of the elements of the set \( V_i \) will be denoted \( F_i \). Thus, \( F_i \) will also have size logarithmic in \( P \). The elements of \( M = \bigcup_{i=1}^r F_i \cup \{0\} \) (the sets \( F_i \) will be made disjoint) will be reserved for the weightings \( f_i \) of the bins of zeros of the prefix string.
in the transformed domain (see the construction (4.12)). Note that \( M \) also has size that is logarithmic in \( P \). Since each bin in the prefix in the transformed domain will have at most one zero apart possibly from the zero-weighting bin, the length of the prefix in the transformed domain is \( 2M \), and the length of the prefix in the original domain is \( l_n = 2|M| + 1 \) and thus also logarithmic in \( P \).1 The sets \( V_i \) will serve to satisfy the \( i \)th congruency constraint of the type given in (4.12) for the string \( t_n \) in the transformed domain, as further explained below.

In the remainder of this section we will first show how to construct sets \( V_i \), and then we will provide the proof that it is possible to construct sets \( V_i \) with all distinct elements as well as sets \( F_i \) (from sets \( V_i \)) that have distinct elements and are nonintersecting, for the prime \( P \) large enough. We will also provide a proof that for a given integer \( n \), for \( n \) large enough, there exists a prime \( P \) for which we can construct nonintersecting sets \( F_i \) containing distinct elements, where the prime \( P \) lies in an interval that linearly depends on \( n \).

Combined with the encoding method described in the next section we will therefore have constructed a prefix whose length is logarithmic in \( n \) such that the overall string (which is a concatenation of the prefix and original string) in the transformed domain satisfies equations of congruential type given in (4.12), which we have already proved in section 4 are sufficient for the immunity to \( r \) repetition errors.

We now provide some auxiliary results. Let \( [x]_P \) indicate the residue mod \( P \) congruent to \( x \).

**Lemma 5.1.** For an integer \( P \), each residue \( v \) mod \( P \) can be expressed as a sum of a subset of elements of the set \( T_{z,P} = \{ [z]_P, [2z]_P, [2^2z]_P, \ldots, [2^Gz]_P \} \), where \( G = \lfloor \log_2 P \rfloor \) and \( z \) is an arbitrary nonzero residue mod \( P \).

**Proof.** Observe that \( T_{1,P} = \{ 1, 2, 2^2, \ldots, 2^G \} \). We first show that each residue \( v \) mod \( P \) can be expressed as a sum of a subset of elements of the set \( T_{1,P} \). Note that each residue \( i, 0 \leq i \leq 2^G - 1 \) (mod \( P \)), can be expressed as a sum of a subset (call this subset \( K_i \)) of the set \( \{ 1, 2, 2^2, \ldots, 2^{G-1} \} \). Here \( K_0 \) is the empty set. Adding \( 2^G \) to the sum of each \( K_i \) for \( 0 \leq i \leq 2^G - 1 \) mod \( P \) generates the remaining residues \( \{ 2^G, 2^G + 1, \ldots, P - 1 \} \). As a result every residue mod \( P \) can be expressed as a sum of a subset of \( T_{1,P} = \{ 1, 2, 2^2, \ldots, 2^G \} \).

Suppose there exists an element \( v \) which cannot be expressed as a sum of a subset of elements of \( T_{z,P} \) for \( z > 1 \) that is, \( v \neq \sum_{i=0}^{G} \epsilon_i z^{2^i} \) mod \( P \), for all choices of \( \{ \epsilon_0, \ldots, \epsilon_G \} \), \( \epsilon_i \in \{ 0, 1 \} \). Let \( z^{-1} \) be the inverse element of \( z \) under multiplication mod \( P \). Then the residue \( v' = vz^{-1} \neq \sum_{i=0}^{G} \epsilon_i z^{2^i} \) mod \( P \) for all choices of \( \{ \epsilon_0, \ldots, \epsilon_G \} \), \( \epsilon_i \in \{ 0, 1 \} \), which contradicts the result from the previous paragraph. \( \square \)

For a prime number \( P \) for which \( i|P - 1 \) and \( i < P - 1 \), let \( Q_i(P) \) be the set of distinct \( i \)-th power residues mod \( P \). We also state the following convenient result.

**Lemma 5.2.** For a prime \( P \) such that \( i|(P-1) \), each residue \( u \) mod \( P \) can be expressed as a sum of two distinct elements of \( Q_i(P) \) in at least \( P/(2i^2) - \sqrt{P}/2 - 3 \) ways.

**Proof.** The result follows from Theorem II in [3], which states that over \( GF(P) \) the equation

\[
(5.1) \quad x^i + y^i = a,
\]

---

1Since each bin stemming from \( F_i \)'s in the transformed domain has size at most 1, the special zero-weighting bin ensures that the total of \( \left| \bigcup_{i=1}^n F_i \right| = |M| - 1 \) zeros is allocated to bins of zeros in the prefix in the transformed domain (in particular, it is empty if all \( F_i \) bins have size 1). Accounting for the 1's separating adjacent bins of zeros, the total length of the prefix in the transformed domain is then \( 2(|M| - 1) + 2 \), and so the prefix in the original domain has size \( 2|M| + 1 \).
where \( x, y, a \in GF(P) \) and nonzero and \( 0 < i < P - 1 \), has at least
\[
\frac{(P - 1)^2}{P} - P^{1/2} \left( 1 + (i - 1)P^{1/2} \right)^2
\]
solutions. Rearrange the terms in (5.2) to conclude that (5.1) has at least
\[
P - (i - 1)^2 \sqrt{P} - 2(i - 1) - 2 + \frac{1}{P} - \frac{1}{\sqrt{P}}
\]
solutions. Noting that \( i \) distinct values of \( x \) result in the same \( x^i \), accounting for the
symmetry of \( x \) and \( y \) and omitting the case \( x^i = y^i \), we obtain a lower bound on
the number of ways a residue \( a \) can be expressed as a sum of two distinct \( i \)th power
residues to be \( P/(2^i) - \sqrt{P}/2 - 3 \). \( \square \)

Equations of the type in (5.1) were also studied by Weil [12].

We now continue with the introduction of some convenient notation. For \( x_{i,1} \) an
\( i \)th power residue define the set \( A_{i,1}(x_{i,1}) \) to be
\[
A_{i,1}(x_{i,1}) = \left\{ |2^{ik}x_{i,1}| \cdot 0 \leq k \leq \left\lfloor \frac{G}{i} \right\rfloor \right\}.
\]
Let \( x_{i,2} \) and \( x_{i,3} \) be distinct \( i \)th power residues such that \( x_{i,2} + x_{i,3} \equiv 2x_{i,1} \mod P \).
These two power residues generate sets \( A_{i,2}(x_{i,2}) \) and \( A_{i,3}(x_{i,3}) \), where
\[
A_{i,2}(x_{i,2}) = \left\{ |2^{ik}x_{i,2}| \cdot 0 \leq k \leq \left\lfloor \frac{G - 1}{i} \right\rfloor \right\} \text{ and}
\]
\[
A_{i,3}(x_{i,3}) = \left\{ |2^{ik}x_{i,3}| \cdot 0 \leq k \leq \left\lfloor \frac{G - 1}{i} \right\rfloor \right\}.
\]

Likewise, for each \( 2^l x_{i,1} \), for \( 1 \leq l \leq i - 1 \), let \( x_{i,2l} \) and \( x_{i,2l+1} \) be distinct \( i \)th
power residues such that \( x_{i,2l} + x_{i,2l+1} \equiv 2^l x_{i,1} \mod P \). These residues generate sets
\( A_{i,2l}(x_{i,2l}) \) and \( A_{i,2l+1}(x_{i,2l+1}) \), where
\[
A_{i,2l}(x_{i,2l}) = \left\{ |2^{ik}x_{i,2l}| \cdot 0 \leq k \leq \left\lfloor \frac{G - l}{i} \right\rfloor \right\} \text{ and}
\]
\[
A_{i,2l+1}(x_{i,2l+1}) = \left\{ |2^{ik}x_{i,2l+1}| \cdot 0 \leq k \leq \left\lfloor \frac{G - l}{i} \right\rfloor \right\}.
\]

By introducing sets \( A_{i,j}(x_{i,j}) \) we have effectively decomposed all residues of the
type \( 2^{ik+l}x_{i,1} \), \( 0 \leq ik + l \leq G \), \( 1 \leq l \leq i - 1 \), into a sum of two \( i \)th power residues,
namely, \( |2^{ik}x_{i,2l}| \) and \( |2^{ik}x_{i,2l+1}| \). For each set \( A_{i,j}(x_{i,j}) \), \( 1 \leq j \leq 2i - 1 \), we let
\( B_{i,j}(x_{i,j}) \) be the set of all \( i \)th power roots of elements of \( A_{i,j}(x_{i,j}) \),
\[
B_{i,j}(x_{i,j}) = \left\{ |2^{ik}y_{i,j}^{(t)}| \cdot (y_{i,j}^{(t)})^i \equiv x_{i,j} \mod P, 1 \leq t \leq i, 0 \leq k \leq \left\lfloor \frac{G - \left\lfloor \frac{j}{i} \right\rfloor}{i} \right\rfloor \right\}.
\]
First note that all elements in \( A_{i,j}(x_{i,j}) \) are \( i \)th power residues by construction.
Moreover, they are all distinct since \( 2^{ij_1} \neq 2^{ij_2} \mod P \) for \( 1 \leq j_1, j_2 \leq \left\lfloor \frac{G - \left\lfloor \frac{j}{i} \right\rfloor}{i} \right\rfloor \),
for \( j_1 \neq j_2 \) implies \( x_{i,j_1}2^{ij_1} \neq x_{i,j_2}2^{ij_2} \mod P \). Thus, \( |A_{i,j}(x_{i,j})| = \left\lfloor \frac{G - \left\lfloor \frac{j}{i} \right\rfloor}{i} \right\rfloor + 1 \),
and since the \( i \)th power roots of distinct \( i \)th power residues are themselves distinct,
\( |B_{i,j}(x_{i,j})| = i \left( \left\lfloor \frac{G - \left\lfloor \frac{j}{i} \right\rfloor}{i} \right\rfloor + 1 \right) \).
**Lemma 5.3.** Suppose \( P \) is a prime number such that \( i| (P - 1) \). Let \( x_{i,1} \) be an \( i \)th power residue. Suppose \( x_{i,j} \) for \( 2 \leq j \leq 2i - 1 \) are \( i \)th power residues such that \( 2^k x_{i,1} \equiv x_{i,2k} + x_{i,2k+1} \mod P \) for \( 1 \leq k \leq (i - 1) \). Let \( A_{i,j}(x_{i,j}) = \{ [2^d x_{i,j}]_P \mid 0 \leq d \leq \left\lfloor \frac{P-1}{\log_2 P} \right\rfloor \} \) for \( 1 \leq j \leq 2i - 1 \) and \( G = \left\lfloor \log_2 P \right\rfloor \). If the sets \( A_{i,j}(x_{i,j}) \) are disjoint for \( 1 \leq j \leq 2i - 1 \), each residue \( u \mod P \) can be expressed as a sum of a subset of elements of the set \( L_z, P = \bigcup_{j=1}^{2i-1} A_{i,j}(x_{i,j}) \), where \( z \) denotes \( x_{i,1} \).

**Proof.** The proof follows immediately from Lemma 5.1 by observing that, with \( z \) denoting \( x_{i,1} \), we have in fact decomposed elements \( [2^k z]_P \) in the set \( T_z, P \) for \( k \) not a multiple of \( i \) into a sum of two component elements such that all component elements are distinct from one another and distinct from \( [2^k z]_P \) for \( i|k \). \( \square \)

The following lemma proves that it is possible to construct subsets \( A_{i,j}(x_{i,j}) \), and subsets \( \beta_{ij}(x_{i,j}) \) from them, of the set of residues mod \( P \) for \( P \) prime that satisfies \( lcm(2, 3, \ldots, r)|(P - 1) \) for a given positive integer \( r \), provided that \( P \) is large enough, such that for fixed \( i \) the subsets \( A_{i,j}(x_{i,j}) \) are disjoint, and such that all subsets \( A_{i,j}(x_{i,j}) \) for \( 1 \leq i \leq r, 1 \leq j \leq 2i - 1 \) are also disjoint. Let \( W_i \) denote the number of ways any residue mod \( P \) can be expressed as a sum of two distinct nonzero \( i \)th power residues mod \( P \). A universal lower bound on \( W_i \) that holds for all residues was given in Lemma 5.2.

**Lemma 5.4.** For a given integer \( r \), suppose a prime number \( P \) satisfies \( lcm(2, 3, \ldots, r)|(P - 1) \). Let \( G = \left\lfloor \log_2 P \right\rfloor \). If \( P - 1 > (G + r)(G + r - 1)(r - 1)^2 \) and \( W_i > 2i(G + i)(G + i - 1) \) for each \( i \) in the range \( 2 \leq i \leq r \), there exist subsets \( A_{i,j}(x_{i,j}) \) of the type given in (5.7) and (5.8) and \( \beta_{ij}(x_{i,j}) \) of the type given in (5.9) such that for fixed \( i \) subsets \( A_{i,j}(x_{i,j}) \) for \( 1 \leq j \leq 2i - 1 \) are disjoint, and for \( 1 \leq i \leq r, 1 \leq j \leq 2i - 1 \) all subsets \( A_{i,j}(x_{i,j}) \) are disjoint.

**Proof.** We inductively build the sets \( A_{i,j}(x_{i,j}) \) and \( \beta_{ij}(x_{i,j}) \) for \( 1 \leq i \leq r \) and \( 1 \leq j \leq 2i - 1 \), starting with the level \( i = 1 \). We then increment \( i \) by one to reach the next collection of sets \( A_{i,j}(x_{i,j}) \) and \( \beta_{ij}(x_{i,j}) \), while making sure the sets \( A_{i,j}(x_{i,j}) \) at the current level are disjoint from one another and with all previously constructed sets at lower levels.

Consider \( i = 1 \). Let \( x_{1,1} \) be an arbitrary residue mod \( P \), and let

\[
A_{1,1}(x_{1,1}) = \{ [2^k x_{1,1}]_P \mid 0 \leq k \leq G \}.
\]

Let \( z_1 = x_{1,1} \) and \( y_{1,1}^{(1)} = x_{1,1} \). Here \( B_{1,1}(z_1) \) is simply \( A_{1,1}(x_{1,1}) \) for \( i = 1 \). All elements in \( B_{1,1}(z_1) \) are distinct and \( |B_{1,1}(z_1)| = (G + 1) \). If \( r = 1 \), we are done, as we did not even appeal to the condition on the lower bound on \( P - 1 \) (it is simply \( P - 1 > 0 \)).

If \( r \geq 2 \), let us consider \( i = 2 \). Consider quadratic residues \( x_{2,1}, x_{2,2}, \) and \( x_{2,3} \). Let their respective distinct quadratic roots be \( y_{2,1}^{(1)}, y_{2,1}^{(2)} \) (so that \( (y_{2,1}^{(1)})^2 \equiv (y_{2,1}^{(2)})^2 \equiv x_{2,1} \mod P \)), \( y_{2,2}^{(1)}, y_{2,2}^{(2)} \) (so that \( (y_{2,2}^{(1)})^2 \equiv (y_{2,2}^{(2)})^2 \equiv x_{2,2} \mod P \)), and \( y_{2,3}^{(1)}, y_{2,3}^{(2)} \) (so that \( (y_{2,3}^{(1)})^2 \equiv (y_{2,3}^{(2)})^2 \equiv x_{2,3} \mod P \)). These quadratic residues give rise to sets

\[
A_{2,1}(x_{2,1}) = \left\{ [2^{2k} x_{2,1}]_P \mid 0 \leq k \leq \left\lfloor \frac{G}{2} \right\rfloor \right\},
\]

\[
A_{2,2}(x_{2,2}) = \left\{ [2^{2k} x_{2,2}]_P \mid 0 \leq k \leq \left\lfloor \frac{G - 1}{2} \right\rfloor \right\}, \text{ and}
\]

\[
A_{2,3}(x_{2,3}) = \left\{ [2^{2k} x_{2,3}]_P \mid 0 \leq k \leq \left\lfloor \frac{G - 1}{2} \right\rfloor \right\}.
\]
Quadratic roots of elements of sets $A_{2,1}(x_{1,1})$, $A_{2,2}(x_{2,2})$, and $A_{2,3}(x_{2,3})$ give rise to sets $B_{2,1}(x_{2,1})$, $B_{2,2}(x_{2,2})$, and $B_{2,3}(x_{2,3})$:

\begin{align}
B_{2,1}(x_{2,1}) &= \left\{ 2^k y_{2,1}^{(t)} | 1 \leq t \leq 2, 0 \leq k \leq \left\lfloor \frac{G}{2} \right\rfloor \right\}, \\
B_{2,2}(x_{2,2}) &= \left\{ 2^k y_{2,2}^{(t)} | 1 \leq t \leq 2, 0 \leq k \leq \left\lfloor \frac{G-1}{2} \right\rfloor \right\}, \text{ and} \\
B_{2,3}(x_{2,3}) &= \left\{ 2^k y_{2,3}^{(t)} | 1 \leq t \leq 2, 0 \leq k \leq \left\lfloor \frac{G-1}{2} \right\rfloor \right\}.
\end{align}

Having fixed the set $B_{1,1}(x_{1,1})$ based on the earlier selection of the residue $x_{1,1}$, we want to show that it is possible to find quadratic residues $x_{2,1}$, $x_{2,2}$, and $x_{2,3}$ such that $x_{2,2} + x_{2,3} \equiv 2x_{2,1} \mod P$ and such that the resulting sets $B_{1,1}(x_{1,1})$, $B_{2,1}(x_{2,1})$, $B_{2,2}(x_{2,2})$, and $B_{2,3}(x_{2,3})$ are all disjoint.

In particular, we require that $x_{2,1}$ is a quadratic residue mod $P$ (there are $(P-1)/2$ quadratic residues) with the property that the set $B_{2,1}(x_{2,1})$ is disjoint from $B_{1,1}(x_{1,1})$. That is, we require

\begin{equation}
\boxed{y_{2,1}^{(1)} 2^k \not\equiv y_{1,1}^{(1)} 2^l \mod P}
\end{equation}

and

\begin{equation}
\boxed{y_{2,1}^{(2)} 2^k \not\equiv y_{1,1}^{(1)} 2^l \mod P}
\end{equation}

for $0 \leq k \leq \left\lfloor \frac{G}{2} \right\rfloor$ and $0 \leq l \leq G$. By squaring the expressions, these two conditions can be combined into

\begin{equation}
\boxed{x_{2,1}^{2} 2^{2k} \not\equiv (x_{1,1})^{2} 2^{2l} \mod P}
\end{equation}

for $0 \leq k \leq \left\lfloor \frac{G}{2} \right\rfloor$ and $0 \leq l \leq G$. For the already chosen $y_{1,1}^{(1)} (= x_{1,1})$ at most $(G+1)(\left\lfloor \frac{G}{2} \right\rfloor + 1)$ candidate quadratic residues out of the total $(P-1)/2$ quadratic residues violate (5.19). Observe that the condition $(G+i)(G+i-1)(i-1)^2$ is strictly increasing for positive $i$, $2 \leq i \leq r$, and thus the condition $P-1 > (G+r)(G+r-1)(r-1)^2$ in the statement of the lemma implies $P-1 > (G+2)(G+1)$. Since $\frac{G+1}{2} > \frac{G+1}{2} \frac{G+2}{2} = (G+1)(\left\lfloor \frac{G}{2} \right\rfloor + 1)$, such $x_{2,1}$ exists.

Fix $x_{2,1}$ such that (5.19) holds. Having chosen such $x_{2,1}$, we now look for $x_{2,2}$ and $x_{2,3}$ as distinct quadratic residues that satisfy $x_{2,2} + x_{2,3} \equiv 2x_{2,1} \mod P$. We require that $B_{2,2}(x_{2,2})$ be disjoint from both $B_{1,1}(x_{1,1})$ and $B_{2,1}(x_{2,1})$ (by construction, if $B_{2,2}(x_{2,2})$ and $B_{2,1}(x_{2,1})$ are disjoint, so are $A_{2,2}(x_{2,2})$ and $A_{2,1}(x_{2,1})$) so that

\begin{align}
y_{2,2}^{(1)} 2^{k_3} &\not\equiv y_{1,1}^{(1)} 2^{k_1} \mod P, \\
y_{2,2}^{(2)} 2^{k_3} &\not\equiv y_{1,1}^{(1)} 2^{k_1} \mod P, \\
y_{2,2}^{(1)} 2^{k_3} &\not\equiv y_{2,1}^{(1)} 2^{k_2} \mod P, \\
y_{2,2}^{(2)} 2^{k_3} &\not\equiv y_{2,1}^{(1)} 2^{k_2} \mod P, \\
y_{2,2}^{(1)} 2^{k_3} &\not\equiv y_{2,2}^{(1)} 2^{k_2} \mod P, \\
y_{2,2}^{(2)} 2^{k_3} &\not\equiv y_{2,2}^{(2)} 2^{k_2} \mod P,
\end{align}

where $0 \leq k_1 \leq G$, $0 \leq k_2 \leq \left\lfloor \frac{G}{2} \right\rfloor$, and $0 \leq k_3 \leq \left\lfloor \frac{G-1}{2} \right\rfloor$. 

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
Alternatively, by squaring both sides in each expression in (5.20),

\[
\begin{align*}
    x_{2,2}2^{2k_3} & \neq (x_{1,1})2^{2k_1} \mod P, \\
    x_{2,2}2^{2k_3} & \neq x_{2,1}2^{2k_2} \mod P,
\end{align*}
\]

where \(0 \leq k_1 \leq G, 0 \leq k_2 \leq \left\lfloor \frac{G}{2} \right\rfloor, \) and \(0 \leq k_3 \leq \left\lfloor \frac{G-1}{2} \right\rfloor.\)

Likewise, we require that \(B_{2,3}(x_{2,3})\) be disjoint from \(B_{1,1}(x_{1,1}), B_{2,1}(x_{2,1}),\) and \(B_{2,2}(x_{2,2})\) (again, if \(B_{2,3}(x_{2,3})\) is disjoint from \(B_{2,2}(x_{2,2})\) and \(B_{2,1}(x_{2,1}),\) then \(A_{2,3}(x_{2,3})\) is disjoint from \(A_{2,2}(x_{2,2})\) and \(A_{2,1}(x_{2,1})\)) so that

\[
\begin{align*}
    x_{2,3}2^{2k_4} & \neq (y_{1,1})2^{2k_1} \mod P, \\
    x_{2,3}2^{2k_4} & \neq x_{2,1}2^{2k_2} \mod P, \\
    x_{2,3}2^{2k_4} & \neq x_{2,2}2^{2k_3} \mod P,
\end{align*}
\]

where \(0 \leq k_1 \leq G, 0 \leq k_2 \leq \left\lfloor \frac{G}{2} \right\rfloor, \) and \(0 \leq k_3 \leq \left\lfloor \frac{G-1}{2} \right\rfloor.\)

For the already chosen values of \(x_{2,1}\) and \(y_1,1\) at most \(N_2 = 2 \left( \left\lfloor \frac{G}{2} \right\rfloor + 1 \right) \left( \left\lfloor \frac{G-1}{2} \right\rfloor + 1 \right) + (G+1) \left( \left\lfloor \frac{G-1}{2} \right\rfloor + 1 \right)^2 \) choices for \(x_{2,2}\) and \(x_{2,3}\) violate (5.21) and (5.22).

We thus require that \(W_2\) be strictly larger than \(N_2.\) Dropping floor operations it is sufficient that \(W_2 > \frac{7(G+1)(G+2)}{4}\). Further simplification yields that

\[
W_2 > \frac{7(G+1)(G+2)}{4}
\]

is sufficient to ensure that there exist \(x_{2,2}, x_{2,3}\) that make the respective sets disjoint.

Note that this last condition follows from the requirement in the statement of the lemma for \(i = 2,\) namely, that \(W_2 > 4(G+1)(G+2).\) If \(r = 2,\) we are done; else we consider \(i = 3.\) Before considering general level \(i,\) let us present the \(i = 3\) case.

For \(i = 3\) we seek distinct cubic residues \(x_{3,1}, x_{3,2}, x_{3,3}, x_{3,4},\) and \(x_{3,5}\) with the property that \(x_{3,2} + x_{3,3} \equiv 2x_{3,1} \mod P\) and \(x_{3,4} + x_{3,5} \equiv 2^2x_{3,1} \mod P,\) and such that the respective sets \(B_{3,j}(x_{3,j})\) for \(1 \leq j \leq 5\) generated from the cubic roots of these residues are disjoint and are disjoint from previously constructed sets \(B_{1,1}(x_{1,1}), B_{2,1}(x_{2,1}), B_{2,2}(x_{2,2}),\) and \(B_{3,1}(x_{3,1}).\)

We start with \(x_{3,1},\) a cubic residue \(\mod P\) (there are \((P-1)/3\) cubic residues) with the property that the set \(B_{3,1}(x_{3,1})\) is disjoint from each of \(B_{1,1}(x_{1,1}), B_{2,1}(x_{2,1}), B_{2,2}(x_{2,2}),\) and \(B_{3,1}(x_{3,1}).\) That is, after raising the elements of these sets to the third power, we require

\[
\begin{align*}
    x_{3,1}2^{3k_5} & \neq (y_{1,1})3^{3k_1} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,1})3^{2k_2} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,1})3^{2k_2} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,2})3^{3k_3} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,2})3^{3k_3} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,3})3^{3k_4} \mod P, \\
    x_{3,1}2^{3k_5} & \neq (y_{2,3})3^{3k_4} \mod P,
\end{align*}
\]

where \(0 \leq k_1 \leq G, 0 \leq k_2 \leq \left\lfloor \frac{G}{2} \right\rfloor, \) and \(0 \leq k_3 \leq \left\lfloor \frac{G-1}{2} \right\rfloor.\)

For the already chosen values of \(x_{1,1}\) through \(x_{2,3},\) which in turn determine \(y_{1,1}\) through \(y_{2,3},\) the condition in (5.24) prevents \(N_3 = \left( \left\lfloor \frac{G}{2} \right\rfloor + 1 \right) \left( G+1 \right) + 2 \left( \left\lfloor \frac{G}{2} \right\rfloor + 1 \right) + \)
4 \left( \left\lceil \frac{G}{3} \right\rceil + 1 \right) \) choices for \( x_{3,1} \). Since there are \( \frac{P-1}{3} \) cubic residues, after simplifying and upper bounding the expression for \( N_3 \), it follows that it is sufficient that \( \frac{P-1}{3} \) be strictly larger than \( \frac{4(G+2)(G+3)}{3} \). Note that this condition is implied by the requirement that \( P - 1 > (r - 1)^2(G + r)(G + r - 1) \) (again, since the function \((i - 1)^2(G + i)(G + i - 1)\) is strictly increasing for positive \( i \)).

Fix \( x_{3,1} \) such that (5.24) holds. Having chosen such \( x_{3,1} \), we now look for distinct \( x_{3,2}, x_{3,3}, x_{3,4}, x_{3,5} \) cubic residues that satisfy \( x_{3,2} + x_{3,3} \equiv 2x_{3,1} \mod P \) and \( x_{3,4} + x_{3,5} \equiv 2^2x_{3,1} \mod P \) that make all sets \( B_{i,j}(x_{i,j}) \), \( 1 \leq i \leq 3, 1 \leq j \leq 2i - 1 \) disjoint.

In order that residue \( x_{3,2} \) generates set \( B_{3,2}(x_{3,2}) \) with the property that \( B_{3,2}(x_{3,2}) \) is disjoint from each of \( B_{1,1}(x_{1,1}), B_{2,1}(x_{2,1}), B_{2,2}(x_{2,2}), B_{2,3}(x_{2,3}) \), and \( B_{3,1}(x_{3,1}) \), we require that their respective elements raised to the third power be distinct,

\[
\begin{align*}
x_{3,2}^{3k_6} & \neq (y_{1,1})^{3k_1} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,1})^{3k_2} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,1})^{3k_2} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,2})^{3k_3} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,2})^{3k_3} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,3})^{3k_4} \mod P, \\
x_{3,2}^{3k_6} & \neq (y_{2,3})^{3k_4} \mod P, \\
x_{3,2}^{3k_6} & \neq x_{3,1}^{2k_5} \mod P,
\end{align*}
\]

where \( 0 \leq k_1 \leq G, 0 \leq k_2 \leq \left\lceil \frac{G}{2} \right\rceil, 0 \leq k_3 \leq \left\lceil \frac{G-1}{2} \right\rceil, 0 \leq k_4 \leq \left\lceil \frac{G-1}{2} \right\rceil, 0 \leq k_5 \leq \left\lceil \frac{G}{4} \right\rceil, \) and \( 0 \leq k_6 \leq \left\lceil \frac{G}{3} \right\rceil \).

Likewise, we require that \( B_{3,3}(x_{3,3}) \) be disjoint from all of \( B_{1,1}(x_{1,1}), B_{2,1}(x_{2,1}), B_{2,2}(x_{2,2}), B_{2,3}(x_{2,3}), B_{3,1}(x_{3,1}), \) and \( B_{3,2}(x_{3,2}) \), so that

\[
\begin{align*}
x_{3,3}^{3k_7} & \neq (y_{1,1})^{3k_1} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,1})^{3k_2} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,1})^{3k_2} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,2})^{3k_3} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,2})^{3k_3} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,3})^{3k_4} \mod P, \\
x_{3,3}^{3k_7} & \neq (y_{2,3})^{3k_4} \mod P, \\
x_{3,3}^{3k_7} & \neq x_{3,2}^{2k_6} \mod P,
\end{align*}
\]

where \( 0 \leq k_1 \leq G, 0 \leq k_2 \leq \left\lceil \frac{G}{2} \right\rceil, 0 \leq k_3 \leq \left\lceil \frac{G-1}{2} \right\rceil, 0 \leq k_4 \leq \left\lceil \frac{G-1}{2} \right\rceil, 0 \leq k_5 \leq \left\lceil \frac{G}{4} \right\rceil, \) and \( 0 \leq k_6 \leq \left\lceil \frac{G}{3} \right\rceil \), and \( 0 \leq k_7 \leq \left\lceil \frac{G}{4} \right\rceil \).

From (5.25) and (5.26) it follows that at most

\[
N_3 = \frac{2 \left( \left\lceil \frac{G}{3} \right\rceil + 1 \right) \left( \left\lceil \frac{G}{2} \right\rceil + 2 \left( \left\lceil \frac{G}{2} \right\rceil + 1 \right) + 4 \left( \left\lceil \frac{G-1}{2} \right\rceil + 1 \right) + \left( \left\lceil \frac{G}{4} \right\rceil + 1 \right) \right)^2}{\left( \left\lceil \frac{G}{4} \right\rceil + 1 \right)^2}
\]

candidate pairs \( (x_{3,2}, x_{3,3}) \) do not make the respective \( B_{i,j}(x_{i,j}) \) sets disjoint. Since

\[
\begin{align*}
N_3 & \leq 2 \left( \left\lceil \frac{G+2}{3} \right\rceil \left( \left\lceil \frac{G}{3} \right\rceil + 2 \left( \left\lceil \frac{G}{2} \right\rceil + 4 \left( \left\lceil \frac{G}{4} \right\rceil + 1 \right) \right) + \left( \left\lceil \frac{G}{4} \right\rceil + 2 \right)^2 \right) \\
& < 2 \left( \frac{G+2}{3} \right) \cdot 13 \left( \frac{G+3}{3} \right) + \left( \frac{G+4}{2} \right)^2 \\
& < 3(G+2)(G+3),
\end{align*}
\]
it follows that it is sufficient that

(5.29) \[ W_3 > 3(G + 2)(G + 3), \]

where \( W_3 \) is the number of ways a residue \( \mod P \) can be expressed as a sum of two different cubic residues. The same is true for the cubic residues \( x_{3,4} \) and \( x_{3,5} \) for which the respective disjoint \( B_{i,j}(x_{i,j}) \) sets exist, provided that

(5.30) \[ W_3 > 2 \left( \left\lceil \frac{G-2}{3} \right\rceil + 1 \right) \left( G + 1 \right) + 2 \left( \left\lceil \frac{G-1}{3} \right\rceil + 1 \right) + 4 \left( \left\lceil \frac{G}{3} \right\rceil + 1 \right) + \left( \left\lceil \frac{G}{2} \right\rceil + 1 \right)^2. \]

Some simplification of (5.30) yields

(5.31) \[ W_3 > \frac{31}{9}(G + 2)(G + 3), \]

which subsumes the lower bound on \( W_3 \) given in (5.29). Note that (5.31) is implied by the condition in the statement of the lemma, namely, \( W_3 > 6(G + 2)(G + 3) \).

We now inductively show the existence of the appropriate \( r \)-th power residues and their sets, assuming that we have successfully identified power residues at lower levels for which all the sets \( B_{k,j}(x_{k,j}) \) for \( 1 \le k < i, 1 \le j \le 2k-1 \) are disjoint.

Consider \( x_{i,1} \) an \( i \)-th power residue \( \mod P \) (there are \( P - 1 \) such residues) with the property that the set \( B_{i,1}(x_{i,1}) \) is disjoint from all of \( B_{k,j}(x_{k,j}) \) for \( 1 \le k < i, 1 \le j \le 2k-1 \). These constraints on disjointness (an example of which is given in (5.19) for \( i = 2 \) and in (5.24) for \( i = 3 \)) prevent no more than \( (\frac{G+i}{i})(\frac{G+k}{k}) \) choices for \( x_{i,1} \) for each \( y_{k,j}^{(t)} \), where \( 1 \le k \le i-1, 1 \le j \le 2k-1, \) and \( 1 \le t \le k \) (since \( |A_{i,1}(x_{i,1})| = \left\lceil \frac{G}{i} \right\rceil + 1 \le \frac{G+i}{i} \)), and \( |A_{k,j}(x_{k,j})| = \left\lceil \frac{G-1}{k} \right\rceil + 1 \le \frac{G+k}{k} \). By summing over all choices it follows that at most

(5.32) \[ \left( \frac{G+i}{i} \right) \sum_{k=1}^{i-1} (2k-1) \left( \frac{G+k}{k} \right) \le (G + i) \left( \frac{G+i-1}{i} \right) \sum_{k=1}^{i-1} (2k-1) = (G + i) \left( \frac{G+i-1}{i} \right) (i-1)^2 \]

\( i \)-th power residues cannot be chosen for \( x_{i,1} \). Since there are \( P-1 \) \( i \)-th power residues, we thus require

(5.33) \[ P - 1 > (G + i)(G + i - 1)(i - 1)^2 \]

for each level \( i \). Note that since the expression on the right-hand side of the inequality (5.33) is an increasing function of positive \( i \), each subsequent level poses a lower bound on \( P \) that subsumes all previous bounds. It is thus sufficient to have \( P - 1 > (G + r)(G + r - 1)(r - 1)^2 \), as given in the statement of the lemma.

Consider \( x_{i,2} \) and \( x_{i,3} \) as distinct \( i \)-th power residues \( \mod P \) that satisfy \( x_{i,2} + x_{i,3} \equiv 2x_{i,1} \mod P \) for a previously chosen \( x_{i,1} \). We require that \( x_{i,2} \) and \( x_{i,3} \) give rise to sets \( B_{i,2}(x_{i,2}) \) and \( B_{i,3}(x_{i,3}) \) that are disjoint and that are disjoint from each of \( B_{k,j}(x_{k,j}) \) for \( 1 \le k < i, 1 \le j \le 2k-1 \) and from \( B_{i,1}(x_{i,1}) \). By construction, if the sets \( B_{i,1}(x_{i,1}), B_{i,2}(x_{i,2}), \) and \( B_{i,3}(x_{i,3}) \) are disjoint, then so are sets \( A_{i,1}(x_{i,1}), A_{i,2}(x_{i,2}), \) and \( A_{i,3}(x_{i,3}) \). Constraints based on the previously encountered \( y_{j,k}^{(t)} \) for \( 1 \le k < i, 1 \le j \le 2k-1, 1 \le t \le k \) prevent at most \( \left( \frac{G+i-1}{i} \right)(\frac{G+k}{k}) \) choices for each of \( x_{i,2} \) and \( x_{i,3} \), for each \( y_{j,k}^{(t)} \) (since \( |A_{i,2}(x_{i,2})| = |A_{i,3}(x_{i,3})| = \left\lceil \frac{G-1}{i} \right\rceil + 1 \le \frac{G+i-1}{i} \)
and \(|A_{k,j}(x_{k,j})| = \left\lfloor \frac{G-1+i}{k} \right\rfloor + 1 \leq \frac{G+i}{k}\). Combined with the restriction based on the disjointness with \(B_{i,1}(x_{i,1})\) and the requirement that \(B_{i,2}(x_{i,2})\) and \(B_{i,3}(x_{i,3})\) be nonintersecting, it follows that

\[
W_i > 2\left(\frac{G+i-1}{i}\right) \left[\sum_{k=1}^{i-1}(2k-1)k\left(\frac{G+i}{k}\right) + \left(\frac{G+i}{i}\right)\right] + \left(\frac{G+i-1}{i}\right)^2
\]

is sufficient for the pair \((x_{i,2}, x_{i,3})\) to exist.

Likewise, for \(x_{i,2}\) and \(x_{i,2l+1}\) to be distinct \(i\)th power residues mod \(P\) that satisfy \(x_{i,2} + x_{i,2l+1} \equiv 2^ix_{i,1} \mod P\), that give rise to disjoint sets \(B_{i,2l}(x_{i,2})\) and \(B_{i,2l+1}(x_{i,2l+1})\) and that are also disjoint from all previously constructed set \(B_{k,j}(x_{k,j})\), we require

\[
W_i > 2\left(\frac{G+i-1}{i}\right) \left[\sum_{k=1}^{i-1}(2k-1)k\left(\frac{G+i}{k}\right) + (2i-1)\left(\frac{G+i}{i}\right)\right] + \left(\frac{G+i-1}{i}\right)^2
\]

for the pair \((x_{i,2l}, x_{i,2l+1})\) to exist. Note that (5.35) subsumes (5.34). Since at each level \(i\) we construct \(i-1\) pairs \(x_{i,2l}\) and \(x_{i,2l+1}\), and since the right-hand side of (5.35) is an increasing function of \(i\), it is sufficient to upper bound the expression in (5.35) for \(l = i-1\),

\[
W_i > 2\left(\frac{G+i-1}{i}\right) \left[\sum_{k=1}^{i-1}(2k-1)k\left(\frac{G+i}{k}\right) + (2i-1)\left(\frac{G+i}{i}\right)\right] + \left(\frac{G+i-1}{i}\right)^2
\]

\[
\iff W_i > 2\left(\frac{G+i-1}{i}\right) \left[(i-1)^2(G+i) + \frac{2i-3}{i}(G+i)\right] + \left(\frac{G+i-1}{i}\right)^2
\]

\[
\iff W_i > (G+i)(G+i-1)\left(\frac{G+i}{i}\right)\left(\frac{i}{2} - \frac{3}{4} + \frac{1}{6}\right).
\]

Some simplification yields

\[
W_i > (G+i)(G+i-1)\frac{2i^2-4i^2+6i-5}{i}^2
\]

as a sufficient condition for the disjoint sets \(B_{i,j}(x_{i,j})\) to exist that are also disjoint from all sets \(B_{k,j}(x_{k,j})\) for \(k < i\).

Further simplifying the last inequality, it is sufficient that

\[
W_i > 2i(G+i)(G+i-1)
\]

to make these sets disjoint. We have thus demonstrated that with the appropriate lower bounds on \(P\) and \(W_i\)’s, it is possible to construct disjoint sets \(B_{i,j}(x_{i,j})\).

Note that all residues mod \(P\) can be expressed as a sum of a subset of elements of \(V_i = \bigcup_{j=1}^{2i-1} A_{i,j}(x_{i,j})\) by Lemma 5.3 for each \(i, 1 \leq i \leq r\). Also note that \(|V_i|\) scales as \(\log_2(P)\), since \(|A_{i,j}(x_{i,j})| = \left\lfloor \frac{G-1+1}{i} \right\rfloor + 1\). For \(F_i = \bigcup_{j=1}^{2i-1} B_{i,j}(x_{i,j})\), \(|F_i|\) also scales as \(\log_2(P)\), since \(|B_{i,j}(x_{i,j})| = i \left\lfloor \frac{G-1}{i} \right\rfloor + 1\).

We now discuss how large prime \(P\) needs to be so that the conditions of Lemma 5.4 hold. Namely, we require

\[
P - 1 > (r - 1)^2(G + r)(G + r - 1)
\]

and

\[
W_i > 2i(G+i)(G+i-1)\quad \text{for } 2 \leq i \leq r.
\]

Using Lemma 5.2 it follows that it is sufficient that

\[
P > 4r^3(G+r)(G+r-1) + r^2 \sqrt{P} + 6r^2\quad \text{for } r \geq 2,
\]
for (5.40) to hold. Moreover, if (5.41) holds, it implies (5.39). (For $r = 1$, the requirement is $P > 1$.) The expression (5.41) certainly holds as $P \to \infty$, and for the finite values of $P$ we (loosely) have that

$$
P > 2 \times 10^2 \quad \text{for } r = 1,
$$

$$
P > 4 \times 10^3 \quad \text{for } r = 2,
$$

$$
P > 2 \times 10^4 \quad \text{for } r = 3,
$$

$$
P > 6 \times 10^4 \quad \text{for } r = 4,
$$

$$
P > 2 \times 10^5 \quad \text{for } r = 5.
$$

(5.42)

For a given large enough integer $n$, we now show that there exists a prime number $P$ that satisfies (5.41) (which holds for $P$ large enough) and for which $\text{lcm}(2, 3, \ldots, r)|(P - 1)$ such that $P$ lies in an interval that is linear in $n$. Since the elements of $M = \bigcup_{i=1}^r F_i \cup \{0\}$ are to be reserved for the indices of bins of zeros of the prefix in the transformed domain, we also require that $P - n > |M|$, since the total number of bins of zeros to be used is at most $n$ (from the original string) + $|M|$ (from the prefix), and each bin receives a distinct index. Since $F_i = \bigcup_{j=1}^{2^{i-1}} B_{i,j}(x_{i,j})$ and $|B_{i,j}(x_{i,j})| = i \left(\lfloor \frac{G_i - i}{i} \rfloor + 1\right)$, whereby $i \left(\frac{G_i - i}{i}\right) \leq |B_{i,j}(x_{i,j})| \leq i \left(\frac{G_i + i}{i}\right)$, it follows that

$$
|M| \leq \sum_{i=1}^r (2i - 1)(G + i) + 1 \leq (G + r) \sum_{i=1}^r (2i - 1) = r^2(G + r) + 1
$$

(5.43) and

$$
|M| \geq \sum_{i=1}^r (2i - 1)(G - i) + 1 \geq (G - r) \sum_{i=1}^r (2i - 1) = r^2(G - r) + 1.
$$

Equation (5.43) yields a sufficient requirement on how large $P$ needs to be:

$$
P > n + r^2(\log_2(P) + r) + 1.
$$

(5.45)

By (5.43) and (5.44), the total number of bins $|M|$ reserved for the prefix scales as $\Theta(r^2 \log_2(P))$, and since the length of the prefix $l_n$ is $2|M| + 1$, it also scales as $\Theta(r^2 \log_2(P))$.

To express this logarithmic scaling in terms of $n$, for given integers $n$ and $r$ ($n$ is typically large and $r$ is small), we essentially need to show that there exists a prime $P$ for which $k = \text{lcm}(2, 3, \ldots, r)|(P - 1)$ and $P \in (c_1 n, c_2 n)$ (here $c_1$ and $c_2$ are positive numbers that do not depend on $n$) and such that $P$ satisfies (5.41) and (5.45).

For the asymptotic regime as $n \to \infty$ we recall the prime number theorem for arithmetic progressions [10], which states that

$$
\pi(n, k, 1) \sim \frac{1}{\phi(k)} \frac{n}{\log(n)},
$$

(5.46)

where $\pi(n, k, 1)$ denotes the number of primes $\leq n$ that are congruent to 1 mod $k$, and $\phi(k)$ is the Euler function and represents the number of integers $\leq k$ that are relatively prime with $k$. As $n \to \infty$, we may let $c_1 := 2$ and $c_2 := 4$, so that

$$
\frac{\pi(4n, k, 1)}{\pi(2n, k, 1)} \sim 2,
$$

(5.47)
and thus there exists a prime $P$, $k|(P - 1)$, in an interval that is linear in $n$. Clearly, as $n \to \infty$, such $P$ also satisfies (5.41) and (5.45).

For finite (but possibly very large) values of $n$ and certain small $r$, we appeal to results by Ramare and Rumely [7]. The number-theoretic function $\theta(x; k, l)$ is usually defined as

\[(5.48) \quad \theta(x; k, l) = \sum_{\text{prime } p, p \equiv l \mod k, p \leq x} \ln p.\]

To show that there exists a prime $P$ in the interval $(c_1n, c_2n)$ for which $k = \text{lcm}(2, 3, \ldots, r)|(P - 1)$ it is sufficient to have

\[(5.49) \quad \theta(c_2n; k, 1) > \theta(c_1n; k, 1),\]

where $k = \text{lcm}(2, 3, \ldots, r)$.

Theorem 2 in [7] states that $|\theta(x; k, 1) - \frac{x}{\phi(k)}| \leq 2.072\sqrt{x}$ for all $x \leq 10^{10}$ for $k$ given in Table I of [7]. For larger $x$, Theorem 1 in [7] provides the bounds of the type

\[(5.50) \quad (1 - \varepsilon) \frac{x}{\phi(k)} \leq \theta(x; k, 1) \leq (1 + \varepsilon) \frac{x}{\phi(k)}\]

for $k$ given in Table I of [7] and $\varepsilon$ also given in Table I of [7] for various $x$. Here $\phi(k)$ is the Euler function and denotes the number of integers $\leq k$ that are relatively prime with $k$.

For $c_2n \leq 10^{10}$, using

\[(5.51) \quad \theta(c_1n; k, 1) < \frac{c_1n}{\phi(k)} + 2.072\sqrt{c_1n}\]

and

\[(5.52) \quad \theta(c_2n; k, 1) > \frac{c_2n}{\phi(k)} - 2.072\sqrt{c_2n},\]

it is thus sufficient to have

\[(5.53) \quad 2.072\phi(k) < \sqrt{n}(\sqrt{c_2} - \sqrt{c_1})\]

for $\theta(c_2n; k, 1) > \theta(c_1n; k, 1)$ to hold.

For $c_1n \leq 10^{10}$, using

\[(5.54) \quad \theta(c_1n; k, 1) < (1 + \varepsilon) \frac{c_1n}{\phi(k)}\]

and

\[(5.55) \quad \theta(c_2n; k, 1) > (1 - \varepsilon) \frac{c_2n}{\phi(k)}\]

after some simplification, it is sufficient to have

\[(5.56) \quad (1 + \varepsilon)c_1 < (1 - \varepsilon)c_2\]

for $\theta(c_2n; k, 1) > \theta(c_1n; k, 1)$ to hold.

Expressing $P \in (c_1n, c_2n)$ in terms of $c_1n$ and $c_2n$, it is sufficient that

\[(5.57) \quad (c_1 - 1)n > r^2(\log_2 n + \log_2 c_2 + r) + 1\]
for (5.45) to hold. Likewise, for \( r \geq 2 \), it is sufficient that

\[
(5.58) \quad c_1 n > 4 r^3 (\log_2 n + \log_2 c_2 + r) (\log_2 n + \log_2 c_2 + r - 1) + r^2 (6 + \sqrt{c_2 n})
\]

for (5.41) to hold.

Parameters \( c_1 \) and \( c_2 \) can be chosen as a function of \( r \) to make (5.53) (or (5.56)), (5.57), and (5.58) hold. We consider now some suitable choices for \( c_1 \) and \( c_2 \) for small values of \( r \) and provide explicit bounds on the length of the prefix \( l_n \) in terms of \( n \).

- \( r = 1 \): The condition (5.57) reduces to \((c_1 - 1)n > \log_2 n + \log_2 c_2 + 2\). For \( c_2 n < 10^{10} \), the condition (5.53) reduces to \( \sqrt{n} (\sqrt{c_2} - \sqrt{c_1}) > 2.072 \). We may let \( c_2 = 4 \) and \( c_1 = 2 \) for \( 12 < n < 10^{10}/4 \) to ensure that there exists a prime in the interval \((2n, 4n)\) which satisfies (5.57).

The condition (5.56) applies to \( c_1 n > 10^{10} \), and so we may let \( c_1 = 4 \) for \( n > 10^{10}/4 \). Since all \( \varepsilon \) entries for \( k = 1 \) in Table I of [7] are \( \ll 1/9 \), we may let \( c_2 = 5 \) to make the condition (5.57) hold.

Since \( |M| \leq \lfloor (\log_2 P) + 2 \rfloor \leq (\log_2 n + \log_2 c_2 + 2) \) (from (5.43)), and \( |M| \geq \lceil \log_2 P \rceil \geq (\log_2 n + \log_2 c_1 - 2) + 1 \) (from (5.44)) it follows that \( \log_2 n \leq |M| \leq (\log_2 n + 4) \) for \( 12 < n < 10^{10}/4 \) and \( \log_2 n + 1 \leq |M| \leq (\log_2 n + 5) \) for \( n > 10^{10}/4 \).

Thus we have, for \( n \geq 10 \), \( 4 (\log_2 n + 7) + 1 \leq |M| \leq 4 (\log_2 n + 14) + 1 \). Consequently, by \( l_n = 2 |M| + 1 \), it follows that \( 8 \log_2 n + 59 \leq l_n \leq 8 \log_2 n + 115 \).

- \( r = 3 \): The conditions (5.57) and (5.58) reduce to \((c_1 - 1)n > 9 (\log_2 n + \log_2 c_2 + 3) + 1 \) and \( c_1 n > 4 \cdot 27 (\log_2 n + \log_2 c_2 + 3) (\log_2 n + \log_2 c + 2) + 9 (6 + \sqrt{c_2 n}) \).

For \( c_2 n < 10^{10} \), the condition (5.53) is now \( \sqrt{n} (\sqrt{c_2} - \sqrt{c_1}) > 2.072 \times 2 \). We may let \( c_1 = 2^{12} \) and \( c_2 = 2^{13} \) to satisfy the required conditions (5.53), (5.57), and (5.58) for \( 10 \leq n \leq 10^{10}/2^{11} = 1/2 \times 5^{10} \).

For \( n \geq 1/2 \times 5^{10} \), we may let \( c_1 = 2^{11} \) and \( c_2 = 2^{12} \) to satisfy the required conditions (5.56) (since all \( \varepsilon \) entries in Table I of [7] are \( \ll 1/3 \), (5.57), and (5.58). Thus for \( n \geq 10 \), \( 9 (\log_2 n + 8) + 1 \leq |M| \leq 9 (\log_2 n + 17) + 1 \) and, consequently, \( 18 \log_2 n + 147 \leq l_n \leq 18 \log_2 n + 309 \).

- \( r = 4 \): The conditions (5.57) and (5.58) reduce to \((c_1 - 1)n > 16 (\log_2 n + \log_2 c_2 + 4) + 1 \) and \( c_1 n > 4 \cdot 64 (\log_2 n + \log_2 c_2 + 4) (\log_2 n + \log_2 c_2 + 3) + 16 (6 + \sqrt{c_2 n}) \).

For \( c_2 n < 10^{10} \), the condition (5.53) is \( \sqrt{n} (\sqrt{c_2} - \sqrt{c_1}) > 2.072 \times 4 \). We may let \( c_1 = 2^{13} \) and \( c_2 = 2^{14} \) to satisfy the required conditions (5.53), (5.57), and (5.58) for \( 16 \leq n \leq 10^{10}/2^{14} = 1/16 \times 5^{10} \).

For \( n \geq 1/16 \times 5^{10} \), it suffices to let \( c_1 = 2^{14} \) and \( c_2 = 2^{15} \) to ensure that (5.53), (5.57), and (5.58) are satisfied.
Thus for \( n \geq 16 \), we have \( 16(\log_2 n + 8) + 1 \leq |M| \leq 16(\log_2 n + 19) + 1 \) and, consequently, \( 32 \log_2 n + 259 \leq l_n \leq 32 \log_2 n + 611 \).

- \( r = 5 \): The conditions (5.57) and (5.58) reduce to \((c_1 - 1)n > 25(\log_2 n + \log_2 c_2 + 5) + 1 \) and \( c_1 n > 4 \cdot 125(\log_2 n + \log_2 c_2 + 5)(\log_2 n + \log_2 c_2 + 4) + 25(6 + \sqrt{c_2 n}) \).

For \( c_2 n < 10^{10} \), the condition (5.53) is \( \sqrt{n(\sqrt{c_2} - \sqrt{c_1})} > 2.072 \times 16 \). We may let \( c_1 = 2^{14} \) and \( c_2 = 2^{15} \) to satisfy the required conditions (5.53), (5.57), and (5.58) for \( 19 \leq n \leq 10^{10} / 2^{15} = 1/32 \times 5^{10} \).

For \( n \geq 1/32 \times 5^{10} \) it suffices to let \( c_1 = 2^{15} \) and \( c_2 = 2^{10} \) to ensure that (5.53), (5.57), and (5.58) are satisfied.

Thus for \( n \geq 19 \), we have \( 25(\log_2 n + 8) + 1 \leq |M| \leq 25(\log_2 n + 21) + 1 \) and, consequently, \( 50 \log_2 n + 403 \leq l_n \leq 50 \log_2 n + 1053 \).

5.2. Prefixing algorithm. Let \( r \) denote the target synchronization error correction capability. The goal of this section is to provide an explicit prefixing scheme which, based on the string \( s \) of length \( n \), produces a fixed length prefix \( p_s \) of length \( l_n \), where \( p_s \) is a function of \( s \), such that the string \( t_s = [p_s \ s] \) after the transformation \( T_{l_n+n} \) given in (2.1) satisfies first \( r \) congruency constraints of the type previously described in (4.12), which were shown to be sufficient to provide immunity to \( r \) repetition errors. Using a judiciously chosen prefix, we will show that this will be possible for \( l_n = |p_s| = \Theta(\log n) \).

We select as \( p_s \) that preimage with the property that in the concatenation \([p_s \ s]\) the last bit of \( p_s \) is the complement of the first bit of \( s \). This property ensures that no bin of zeros in the transformed domain spans the boundary separating the substrings corresponding to the transformed prefix and the transformed original string.

For a given repetition error correction capability \( r \) and the original string length \( n \), let \( P \) be a prime number with the property that \( k = \text{lcm}(2, 3, \ldots, r) \mid (P - 1) \) and such that \( P \) lies in the interval that scales linearly with \( n \), namely, that \( P \in (c_1 n, c_2 n) \) for \( 1 < c_1 < c_2 \), where \( c_1, c_2 \) possibly depend on \( r \) but not on \( n \) and are chosen such that (5.53) (or (5.56) for appropriate \( k \) and \( n \)), (5.57), and (5.58) hold. The existence of such \( P \) was discussed in the previous section. Let \( R_P \) be the set of all residues mod \( P \). Recall that \( M = \cup_{i=1}^{r} F_i \cup \{0\} \) denotes the set of indices of bins of zeros reserved for the prefix, where \( F_i = \cup_{j=1}^{i-1} B_{i,j}(x_{i,j}) \), where \( B_{i,j}(x_{i,j}) \) are given in (5.9), and are constructed such that all sets \( B_{i,j}(x_{i,j}) \) for \( 1 \leq i \leq r, 1 \leq j \leq 2i - 1 \) are nonintersecting. The existence of disjoint sets \( B_{i,j}(x_{i,j}) \) for such \( P \) was proved in Lemma 5.4. Let \( L = |M| \). Let \( N \) denote the total number of bins of zeros of \( \tilde{s} \), where \( \tilde{s} = sT_n \). By construction, \( N \leq n \).

\[
\begin{align*}
    a'_1 &= \sum_{i=L+1}^{L+N} b_i \tilde{f}_i \quad \text{mod } P, \\
    a'_2 &= \sum_{i=L+1}^{L+N} b_i \tilde{f}_i^2 \quad \text{mod } P, \\
    &\vdots \\
    a'_r &= \sum_{i=L+1}^{L+N} b_i \tilde{f}_i^r \quad \text{mod } P,
\end{align*}
\]  

(5.59)

where \( b_i \) is the size of the \( i \)th bin of zeros in \( \tilde{t}_s \) (obtained by transforming \( t_s \) using (2.1)) and \( f_i \) in (5.59) are chosen in increasing order from the set \( R_P \setminus M \). Since \( N \leq n \), and since, by condition (5.57), \( n \leq P - L \), the set \( R_P \setminus M \) is large enough to accommodate such \( f_i \)’s.

We may think of \( a'_1 \) through \( a'_r \) as the contribution of the original string to the overall congruency value of \( \tilde{t}_s \), since the \( i \)th bin of zeros for \( L + 1 \leq i \leq L + N \) is precisely the \( j \)th bin of zeros in \( \tilde{s} \) for \( j = i - L \), since no run spans both \( p_s \) and \( s \) by the choice of \( p_s \).
Since not all strings in the original code may have the same number of bins of zeros in the transformed domain, we may view the unused elements of the set \( R_P \setminus M \) as corresponding to “virtual” bins of size zero. Since these bins are not altered during the transmission that causes \( r \) or fewer repetitions, the locations of repetitions can be uniquely determined as shown in the proof of Lemmas 4.1 and 4.3.

We now show that it is always possible to achieve

\[
\begin{aligned}
a_1 &\equiv \sum_{i=1}^{L+N} b_i f_i \mod P, \\
a_2 &\equiv \sum_{i=1}^{L+N} b_i f_i^2 \mod P, \\
&\vdots \\
a_r &\equiv \sum_{i=1}^{L+N} b_i f_i^r \mod P
\end{aligned}
\]

(5.60)

for arbitrary but fixed values \( a_1 \) through \( a_r \) irrespective of the values \( a'_1 \) through \( a'_r \), where \( b_i \) is either 0 or 1 for \( 1 \leq i \leq L - 1 \), and where \( f_L = 0 \).

Before describing the encoding method that achieves (5.60), we state the following convenient result.

**Lemma 5.5.** Suppose \( P \) is a prime number such that \( i \mid (P - 1) \). Suppose the equation \( x^t \equiv a \mod P \) has a solution, \( 1 \leq a \leq P - 1 \). Then the equation \( x^t \equiv a \mod P \) has \( i \) distinct solutions [1], and we may call them \( x_1 \) through \( x_i \). The sum \( \sum_{k=1}^{i} x_k^j \equiv 0 \mod P \) for \( 1 \leq j \leq i - 1 \).

**Proof.** Let us consider the equation \( x^t \equiv a \mod P \). Using Vieta’s formulas and Newton’s identities over \( GF(P) \) it follows that \( \sum_{k=1}^{i} x_k^j \equiv 0 \mod P \) for \( 1 \leq j \leq i - 1 \). \( \square \)

The encoding procedure is recursive and proceeds as follows. Let \( l \) be the \( l \)th level of recursion for \( l = 1 \) to \( l = r \). The \( l \)th level ensures that the \( l \)th congruency constraint in (5.60) is satisfied without altering previous \( l - 1 \) levels. At each level \( l \), starting with \( l = 1 \) and while \( l \leq r \), do the following.

1. Select a subset \( T_l \) of \( F_l = \bigcup_{j=1}^{l-1} B_{l,j}(x_{l,j}) \) such that \( \sum_{k \in T_l} k!^l \equiv a_l - a'_l - \sum_{i=1}^{l-1} d_i \mod P \), and such that if an element \( y, y^l = z \mod P \) of \( B_{l,j}(x_{l,j}) \) is selected, then so are all other \( l - 1 \) \( l \)th roots of \( z \) (which are also elements of \( B_{l,j}(x_{l,j}) \) by construction). For \( l = 1 \), \( \sum_{k \in T_l} k \equiv a_1 - a'_1 \mod P \).
2. Let \( d_{i,j} \equiv \sum_{k \in T_l} k^j \mod P \) for \( l + 1 \leq j \leq r \).
3. For each \( i \), \( 1 \leq i \leq |F_l| \), for which \( f_i \in T_l \), we set \( b_i = 1 \), and for each \( i \) for which \( f_i \in (F_l \setminus T_l) \), we set \( b_i = 0 \).
4. Proceed to level \( l + 1 \).

After the level \( r \) is completed, let \( b_L = \sum_{i=1}^{r} (|F_i| - |T_i|) \). The purpose of this bin with zero-weighting is to ensure that the overall string \( t_s \) has the same length irrespective of the structure of the starting string \( s \): Since \( |\bigcup_{i=1}^{r} F_i| = |M| - 1 \), and each bin of zeros is bordered by a single “1,” the total prefix length in the transformed domain is \( 2(|M| - 1) + 2 = 2|M| \), and thus the prefix length \( l_n \) in the original domain is always \( 2|M| + 1 \).

The existence of \( T_i, T_i \subseteq F_l \) in step 1 follows from the lemmas in section 2. In particular, recall that each residue mod \( P \) can be expressed as a sum of a subset \( L_l \) of \( \bigcup_{j=1}^{l-1} A_{l,j}(x_{l,j}) \) by Lemma 5.3. We then let \( T_l \) consist of all \( l \)th power roots of elements in \( L_l \). By construction, \( T_l \) is the union of appropriate subsets of sets \( B_{l,j}(x_{l,j}) \), whose \( l \)th powers are precisely the elements of \( L_l \), and these subsets are disjoint by construction.

Recall that the sets \( B_{l,j}(x_{l,j}) \) are constructed such that if an \( l \)th power root of a residue \( y \) belongs to \( B_{l,j}(x_{l,j}) \), then all \( l \) power roots of \( y \) also belong to \( B_{l,j}(x_{l,j}) \).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
Then, by Lemma 5.5, the contribution to each congruency sum for levels 1 through \( l - 1 \) of the elements of \( F_l \) is zero. Hence, once the target congruency value is reached for a particular level, it will not be altered by establishing congruencies at subsequent levels. As a result, since each string \( \tilde{t}_s \) satisfies congruency constraints given in (4.12), the resulting set of strings is immune to \( r \) repetitions while incurring asymptotically negligible redundancy.

6. Summary and concluding remarks. In this paper we discussed the problem of constructing repetition error correcting codes (subsets of binary strings) and the problem of guaranteeing immunity to repetition errors of a collection of binary strings. We presented explicit number-theoretic constructions and provided results on the cardinalities of these constructions. We provided a generalization of a generating function calculation of Sloane [9] 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 [4]. The latter construction was then used to develop a technique for prefixing a collection of binary strings for guaranteed immunity to repetition errors. The presented prefixing scheme relies on introducing a carefully chosen prefix for each original binary string such that the resulting strings (each consisting of the prefix and one of the original strings) belong to the set previously shown to be immune to repetition errors. The prefix length is constructed to be only logarithmic in the size of the original collection.

Acknowledgment. The authors would like to thank the anonymous reviewers for the careful reading and for suggestions that improved the manuscript.

REFERENCES