Semi-Quantitative Group Testing

Amin Emad and Olgica Milenkovic
University of Illinois, Urbana-Champaign, IL
E-mail: {emad2,milenkov}@illinois.edu

Abstract—We consider a novel group testing procedure, termed *semi-quantitative group testing*, motivated by a class of problems arising in genome sequence processing. Semi-quantitative group testing (SQGT) is a non-binary pooling scheme that may be viewed as a combination of an adder model followed by a quantizer. For the new testing scheme we define the capacity and evaluate the capacity for some special choices of parameters using information theoretic methods. We also define a new class of disjunct codes suitable for SQGT, termed *SQ-disjunct* codes. We also provide both explicit and probabilistic code construction methods for SQGT with simple decoding algorithms.

I. INTRODUCTION

Group testing (GT) is a pooling scheme for identifying a number of subjects with some particular characteristic – called "positives" – among a large pool of subjects. The idea behind GT is that if the number of positives is much smaller than the number of subjects, one can reduce the number of experiments by testing adequately chosen groups of subjects rather than testing each subject individually. In its full generality, GT may be viewed as the problem of inferring the state of a system from a superposition of the state vectors of a subset of the system's elements. As such, it has found many applications in communication theory, signal processing, computer science, mathematics, biology, etc. (for example see [1]-[3]).

Although many models have been considered for combinatorial GT, two main models include the original model considered by Dorfman [4] (henceforth, conventional GT) and the adder model (also known as the adder channel or quantitative GT) [1]. In the former case, the result of a test is an indicator determining if there exist at least one positive in the test (equal to 0 if no positive in the test, and 1 otherwise), while in the latter case, the result of a test specifies the exact number of positives in that test. Motivated by applications in genome sequence processing, we propose a novel non-adaptive test model termed semi-quantitative group testing (SQGT). This model accounts for the fact that in most applications a test is not precise enough to exactly determine the number of positives, but it is more informative than a simple indicator of the presence of at least one positive. In other words, schemes in which results are obtained using a test device with *limited* precision may be modeled as instances of SQGT[1].

We also allow for the possibility of having different amounts of sample material for different test subjects, which results in non-binary test matrices. Although binary testing is required for some applications – such as coin weighing – in other

[1]One may view the SQGT scheme as a generalization of thresholded group testing to multiple thresholds and zero gaps [5].

applications, such as conflict resolution in multiple access channel (MAC) and genotyping, non-binary tests may be used to further reduce the number of tests. In the former example, different non-binary values in a test correspond to different power levels of the users, while in the latter example, they correspond to different amounts of genetic material of different subjects. The reason that non-binary tests are extremely important is that in applications like genotyping, tests are very expensive so that one may be inclined to reduce the number of tests at the expense of extracting more genetic material. While there exists information theoretic analysis that is applicable to the non-binary test matrices [10, Ch. 6], to the best of the authors' knowledge, the only attempts of non-binary code construction relevant to group testing is limited to a handful of papers, including [6] and [7], where constructions are considered for an *adder* MAC channel.

For the new and versatile model of SQGT with Q-ary test results and q-ary test sample sizes, $Q, q \geqslant 2$, we define the concept of capacity. Furthermore, we define a new generalization of the family of *disjunct* codes, first introduced in [8], called "SQ-disjunct" codes. Similar to the family of disjunct codes, this new code family may be decoded using a simple, low-complexity algorithm. We conclude our exposition with a probabilistic method for code construction, of use in applications where the physics of the experiments prohibits structured codes.

The paper is organized as follows. Section II describes the SQGT model, while Section III introduces the capacity of SQGT. In Section IV, we define SQ-disjunct codes and present some simple properties of these codes. In Section V, we describe a number of constructions for SQGT codes.

II. SEMI-QUANTITATIVE GROUP TESTING

Throughout this paper, we adopt the following notation. Bold-face upper-case and bold-face lower-case letters denote matrices and vectors, respectively. Calligraphic letters are used to denote sets. Asymptotic symbols such as $\sim$ and $o(\cdot)$ are used in a standard manner. For an integer k, we define $[k] := \{0, 1, \cdots, k-1\}$.

Let N denote the number of test subjects, and let m denote the number of positives. Also, let u denote an upper bound on the number of positives (i.e. $m \leqslant u$). Let S_i denote the ith subject, $i \in \{1, 2, \cdots, N\}$, and let $S_{i_j} = D_j$ be the jth positive, $j \in \{1, 2, \cdots, m\}$. Furthermore, let $\mathcal{D}$ denote the set of positives, so that $|\mathcal{D}| = m$. We assign to each subject a unique q-ary vector of length n, termed the "signature" or the codeword of the subject. Each coordinate of the signature

corresponds to a test. If $\mathbf{x}_i \in [q]^n$ denotes the signature of the ith subject, then the kth coordinate of $\mathbf{x}_i$ may be viewed as the "amount" of S_i (i.e. sample size, concentration, etc.) used in the kth test. Note that the symbol 0 indicates that S_i is not in the test. For convenience, we refer to the collection of codewords arranged column-wise as the *test matrix* or *code*.

The result of each test is an integer from the set $[Q]$. Each test outcome depends on the number of positives and their sample amount in the test through Q thresholds, η_l ($l \in \{1, 2, \cdots, Q\}$). More precisely, the outcome of the kth test, y_k, equals

$$y_k = r \quad \text{if} \quad \eta_r \leqslant \sum_{j=1}^{m} x_{k,i_j} < \eta_{r+1}, \tag{1}$$

where x_{k,i_j} is the kth coordinate of $\mathbf{x}_{i_j}$, and $\eta_0 = 0$. Based on the definition, it is clear that SQGT may be viewed as a concatenation of an adder channel and a decimator (quantizer). Also, if $q = Q = 2$ and $\eta_1 = 1$, the SQGT model reduces to conventional GT. Furthermore, if $Q - 1 = m(q - 1)$ and $\forall r \in [Q]$, $\eta_r = r$, then the SQGT reduces to the adder channel, with a possibly non-binary test matrix. Note that in this model, we assume that $\eta_Q > (q - 1)u$.

Of special interest is SQGT with a uniform quantizer - i.e. SQGT with equidistant thresholds. In this case, $\eta_r = r\eta$, where $r \in [Q + 1]$, and the following definition may be used to simplify (1).

Definition 1: The "SQ-sum" of $s \geqslant 1$ codewords $\mathbf{x}_j \in [q]^n$, $1 \leqslant j \leqslant s$, denoted by $\mathbf{y} = \circledast_{j=1}^{s} \mathbf{x}_j = \mathbf{x}_1 \circledast \mathbf{x}_2 \circledast \cdots \circledast \mathbf{x}_s$, is a vector of length n with its ith coordinate equal to

$$y_i = \left\lfloor \frac{x_{i,1} + x_{i,2} + \cdots + x_{i,s}}{\eta} \right\rfloor. \tag{2}$$

Here, "+" stands for real-valued addition.

Using this definition, (1) reduces to

$$\mathbf{y} = \circledast_{j=1}^{m} \mathbf{x}_{i_j}. \tag{3}$$

where $\mathbf{x}_{i_j}$ is the signature of the jth positive.

III. CAPACITY OF SQGT

It is well-known that group testing may be viewed as a special instance of a multiple access channel (MAC) (e.g. see [10]). Using this connection, Malyutov [9], D'yachkov [10], and Atia et al. [11] derived information theoretic necessary and sufficient conditions on the required number of tests for conventional GT. It is tedious, yet straightforward, to show that the model described in [9], [10], [11] may also be used to evaluate SQGT schemes. The main difference in the analysis of GT and SQGT arises due to the different forms of the mutual information used to express the necessary and sufficient conditions. We therefore focus on characterizing the mutual informations arising in the SQGT framework. Our notation follows the setup of [11].

Let the sample amount of each subject in each test be chosen in an i.i.d manner from a q-ary alphabet, according to a distribution P_T. Also, let $\mathcal{D}_1^{\{i\}}$ and $\mathcal{D}_2^{\{i\}}$ be disjoint
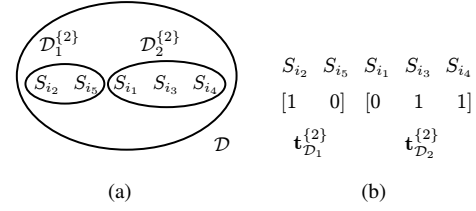

Fig. 1. One choice of $(\mathcal{D}_1^{\{2\}}, \mathcal{D}_2^{\{2\}})$ and their corresponding $\mathbf{t}_{\mathcal{D}_1}^{\{2\}}$ and $\mathbf{t}_{\mathcal{D}_2}^{\{2\}}$ in a binary test design for $m = 5$.

partitions of the set of positives, $\mathcal{D}$, such that $|\mathcal{D}_1^{\{i\}}| = i$ and $|\mathcal{D}_2^{\{i\}}| = m - i$; we denote by $\mathcal{A}_{\mathcal{D}}^{\{i\}}$ the set of all possible pairs $(\mathcal{D}_1^{\{i\}}, \mathcal{D}_2^{\{i\}})$. For a single test, we define y as the test result, and $\mathbf{t}_{\mathcal{D}_j}^{\{i\}}$ (where $j = 1, 2$) as a vector of size $1 \times |\mathcal{D}_j^{\{i\}}|$, with its kth entry equal to the sample amount of the kth positive of $\mathcal{D}_j^{\{i\}}$ in the test. Fig. 1 shows a choice of $(\mathcal{D}_1^{\{2\}}, \mathcal{D}_2^{\{2\}})$ and their corresponding vectors $\mathbf{t}_{\mathcal{D}_1}^{\{2\}}$ and $\mathbf{t}_{\mathcal{D}_2}^{\{2\}}$ for the case where $m = 5$ and $q = 2$.

By following the same steps as in [11], it can be shown that for any fixed m, if the number of tests n satisfies

$$n > \max_{i : (\mathcal{D}_1^{\{i\}}, \mathcal{D}_2^{\{i\}}) \in \mathcal{A}_{\mathcal{D}}^{\{i\}}} \frac{\log \binom{N-m}{i} \binom{m}{i}}{I(\mathbf{t}_{\mathcal{D}_1}^{\{i\}}; \mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)} \quad i = 1, 2, \cdots, m, \tag{4}$$

then the average probability of error asymptotically approaches zero[2]. In this equation, $I(\mathbf{t}_{\mathcal{D}_1}^{\{i\}}; \mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)$ stands for the mutual information between $\mathbf{t}_{\mathcal{D}_1}^{\{i\}}$ and $(\mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)$. Note that since the sample amounts of the subjects are chosen independently and identically, the value of $I(\mathbf{t}_{\mathcal{D}_1}^{\{i\}}; \mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)$ does not depend on the specific choice of $(\mathcal{D}_1^{\{i\}}, \mathcal{D}_2^{\{i\}})$. Similarly, a necessary condition for zero average error probability for SQGT is

$$n \geqslant \max_{i : (\mathcal{D}_1^{\{i\}}, \mathcal{D}_2^{\{i\}}) \in \mathcal{A}_{\mathcal{D}}^{\{i\}}} \frac{\log \binom{N-m+i}{i}}{I(\mathbf{t}_{\mathcal{D}_1}^{\{i\}}; \mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)} \quad i = 1, 2, \cdots, m. \tag{5}$$

Definition 2: [Asymptotic capacity of SQGT channel] Using (4) and (5), we define the *asymptotic capacity* of the channel corresponding to the SQGT scheme (henceforth, SQGT channel) as

$$C = \sup_{P_T, \boldsymbol{\eta}} \alpha(m, P_T, \boldsymbol{\eta}), \tag{6}$$

where $\alpha(m, P_T, \boldsymbol{\eta}) = \min_{i=1,2,\cdots,m} \frac{I(\mathbf{t}_{\mathcal{D}_1}^{\{i\}}; \mathbf{t}_{\mathcal{D}_2}^{\{i\}}, y)}{i}$, and where $\boldsymbol{\eta}$ is a vector of length Q with η_k its kth entry.

In certain applications, $\boldsymbol{\eta}$ may be determined a priori by the resolution of the test equipment. In such applications, the only design parameter to optimize is P_T. On the other hand, if one is able to control the thresholds, $\boldsymbol{\eta}$ becomes a design parameter and clearly exhibits a strong influence on the capacity of the test scheme.

Define the rate of a group test as $R = \frac{\log N}{n}$. The next theorem clarifies the use of the term "capacity" in Definition 2.

Theorem 1: For SQGT, $C = \sup_{P_T, \boldsymbol{\eta}} I(\mathbf{t}_{\mathcal{D}_1}^{\{m\}}; \mathbf{t}_{\mathcal{D}_2}^{\{m\}}, y)/m$, and all rates bellow capacity are achievable. In other words,

---

[2] A sufficient condition for zero error probability can also be found based on the exponential asymptotics of the average error probability. Such calculations are presented in [10] for the particular case of conventional GT.
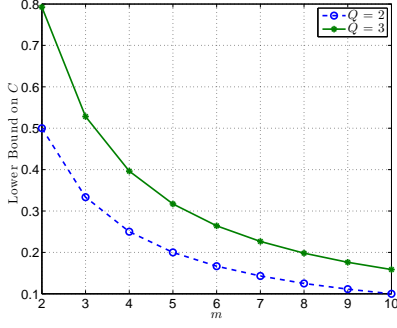
Fig. 2. Numerically obtained lower bound for SQGT with $q = 3$ for different values of m.

m	P_T	quantizer
2	[0.33 0.34 0.33]	$\{0, 1\}\{2\}\{3, 4\}$
3	[0.43 0.46 0.11]	$\{0, 1\}\{2\}\{3, 4, 5, 6\}$
4	[0.18 0.64 0.18]	$\{0, 1, 2, 3\}\{4\}\{5, 6, 7, 8\}$
5	[0.15 0.70 0.15]	$\{0, 1, 2, 3, 4\}\{5\}\{6, 7, 8, 9, 10\}$
6	[0.46 0.15 0.39]	$\{0, 1, 2, 3, 4\}\{5, 6\}\{7, 8, \cdots, 12\}$
7	[0.34 0.25 0.41]	$\{0, 1, \cdots, 6\}\{7, 8\}\{9, 10, \cdots, 14\}$
8	[0.10 0.80 0.10]	$\{0, 1, \cdots, 7\}\{8\}\{9, 10, \cdots, 16\}$
9	[0.09 0.82 0.09]	$\{0, 1, \cdots, 8\}\{9\}\{10, 11, \cdots, 18\}$
10	[0.58 0.28 0.14]	$\{0, 1, \cdots, 4\}\{5, 6\}\{7, 8, \cdots, 20\}$

for every rate $R < C$, there exists a test design for which the average probability of error converges to zero. Conversely, any test design with zero achieving average probability of error must asymptotically satisfy $R < C$.

Proof: Follows from simple modifications of arguments in [9], [10]. ∎

The mutual information $I(\mathbf{t}_{\mathcal{D}_1}^{\{m\}}; \mathbf{t}_{\mathcal{D}_2}^{\{m\}}, y)$ in this theorem may be evaluated as follows. Let W_1 denote the l_1-norm of $\mathbf{t}_{\mathcal{D}_1}^{\{m\}}$. Then,

$$I_{\text{SQ}}(\mathbf{t}_{\mathcal{D}_1}^{\{m\}}; \mathbf{t}_{\mathcal{D}_2}^{\{m\}}, y) = H(y|\mathbf{t}_{\mathcal{D}_2}^{\{m\}}) - H(y|\mathbf{t}_{\mathcal{D}_1}^{\{m\}}, \mathbf{t}_{\mathcal{D}_2}^{\{m\}}) = H(y).$$

On the other hand $\forall l \in [Q]$,

$$P(y = l) = P(\eta_l \leqslant W_1 < \eta_{l+1}) = \sum_{w_1 = \eta_l}^{\eta_{l+1} - 1} P_{W_1}(w_1) \quad (7)$$

where $P_{W_1}(w_1)$ is the probability mass function (PMF) of W_1 and can be found using

$$P_{W_1}(w_1) = P_T(t_1) * P_T(t_2) * \cdots * P_T(t_m), \quad (8)$$

where "$*$" denotes convolution. Note that when $q = 2$,

$$P(y = l) = \sum_{j = \eta_l}^{\eta_{l+1} - 1} \binom{m}{j} p^j (1-p)^{m-j} \quad (9)$$

where p is the probability that a subject is present in a test.

Due to the complicated expression for the mutual information for an arbitrary distribution, a closed-form expression for the test capacity cannot be obtained. We therefore evaluated (6) numerically using a simple search procedure that allows us to quickly determine a lower bound on the capacity. Fig. 2 shows the obtained lower bound on the capacity when $q = 3$, and $Q = 2$ or $Q = 3$. Table I shows one set of probability distributions and thresholds achieving this bound for $Q = 3$.

Table I reveals an interesting property of the quantizers found through numerical search: there exists at least one quantization region that consists of one or two elements only. What this finding implies is that in order to reduce the number of tests as much as possible, a sufficient amount of qualitative information has to be preserved. For example, by having a quantizer that assigns the value v only to inputs of value v, allows for resolving a large amount of uncertainty. Furthermore, the most informative input, left unaltered after

quantization, corresponds to a statistical average of the input symbols, reminiscent to the centroid of a quantization region. These findings will be discussed in more detail in the full version of the paper.

IV. GENERALIZED DISJUNCT AND SEPARABLE CODES FOR SQGT

Disjunct codes were first introduced in [8] for efficient zero-error group testing reconstruction. In what follows, we define a new family of disjunct codes suitable for SQGT that shares many of the properties of binary disjunct codes.

Definition 3: The *syndrome* of a set of vectors $\{\mathbf{x}_i\}$, $i \in \{1, 2, \cdots, s\}$, such that $\mathbf{x}_i \in [q]^n$, is a vector $\mathbf{y} \in [Q]^n$ equal to $\mathbf{y} = \circledast_{j=1}^{s} \mathbf{x}_j$.

Definition 4: A set of codewords $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s\}$ with syndrome $\mathbf{y}_{\mathcal{X}}$ is said to be *included* in another set of codewords $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_t\}$ with syndrome $\mathbf{y}_{\mathcal{Z}}$, if $\forall i \in \{1, 2, \cdots, n\}$, $y_{\mathcal{X}_i} \leqslant y_{\mathcal{Z}_i}$. We denote this inclusion property by $\mathcal{X} \lhd \mathcal{Z}$ or equivalently $\mathbf{y}_{\mathcal{X}} \lhd \mathbf{y}_{\mathcal{Z}}$.

Remark 1: By this definition, it can be easily verified that if $\mathcal{X} \subseteq \mathcal{Z}$, then $\mathcal{X} \lhd \mathcal{Z}$.

Note that for $q = 2$, this definition is equivalent to the definition of inclusion for conventional GT, defined in [8].

Definition 5: A code is called a $[q; Q; \boldsymbol{\eta}; u]$-SQ-disjunct code of length n and size N if $\forall s, t \leqslant u$ and for any sets of q-ary codewords $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s\}$ and $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_t\}$, $\mathcal{X} \lhd \mathcal{Z}$ implies $\mathcal{X} \subseteq \mathcal{Z}$.

Henceforth, we focus on the case where the thresholds are equidistant. We call such codes $[q; Q; \eta; u]$-SQ-disjunct codes.

Proposition 1: A code is $[q; Q; \eta; u]$-SQ-disjunct if and only if no codeword is included in the set of u other codewords.

Proof: It is easy to verify that if a code is $[q; Q; \eta; u]$-SQ-disjunct, then no codeword is included in the set of u other codewords.

Conversely, let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s\}$ and $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_t\}$ be two sets of codewords where $s, t \leqslant u$. From the assumption that no codeword is included in the set of u other codewords, one can conclude that no codeword is included in the set of t other codewords when $t \leqslant u$. If $\mathcal{X} \lhd \mathcal{Z}$

but $\mathcal{X} \nsubseteq \mathcal{Z}$, then there exists a codeword $\mathbf{x}_j$ ($j \in \{1, 2, \cdots, s\}$) such that $\{\mathbf{x}_j\} \nsubseteq \mathcal{Z}$. But since $\{\mathbf{x}_j\} \lhd \mathcal{X} \lhd \mathcal{Z}$, then $\{\mathbf{x}_j\} \lhd \mathcal{Z}$, which contradicts the assumption that no codeword is included in t other codewords. ■

Remark 2: From Proposition 1, one can conclude that a code is $[q; Q; \eta; u]$-SQ-disjunct if and only if for any set of $u + 1$ codewords, $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{u+1}\}$, there exists a unique coordinate $k(i)$ in each codeword $\mathbf{x}_i$, for which

$$\left\lfloor \frac{x_{k(i),i}}{\eta} \right\rfloor > \left\lfloor \frac{\sum_{j=1, j \neq i}^{u+1} x_{k(i),j}}{\eta} \right\rfloor. \tag{10}$$

By unique coordinate, we mean that $k(i) \neq k(j)$, if $i \neq j$. Consequently, a necessary condition for the existence of a $[q; Q; \eta; u]$-SQ-disjunct code is that $q - 1 \geqslant \eta$. As a result, there exist no binary $[2; Q; \eta; u]$-SQ-disjunct code when $\eta > 1$.

Proposition 2: Any code generated by multiplying a conventional binary u-disjunct code by $q - 1$, where $q - 1 \geqslant \eta_1$, is a $[q; Q; \boldsymbol{\eta}; u]$-SQ-disjunct code. As a result, the rate of the best $[q; Q; \boldsymbol{\eta}; u]$-SQ-disjunct code is at least as large as the rate of the best binary u-disjunct code with the same size and length.

Our interest in SQ-disjunct codes lies in their simple decoding procedures, of complexity $O(nN)$. However, one can construct codes for SQGT using other GT codes, such as binary u-separable codes for conventional GT [8] or codes designed for the adder channel [12]. It can be shown that any of these two family of codes can be multiplied by η to form a code for SQGT.

Remark 3: The constructions described in this section reveal the following, and highly intuitive fact: the number of individuals that may be successfully examined with Q-ary SQGT may be as large as the number of individuals that may be tested under the adder channel model, provided that one is allowed to pool different amounts of sample material in each test. In other words, the rate of adder and SGQT channels may be the same, despite the loss of information induced by the quantizer, provided that the alphabet size of the latter scheme is sufficiently larger than the alphabet size of the former scheme.

V. CODE CONSTRUCTION FOR SQGT

In what follows, we discuss two approaches for constructing SQGT codes. For simplicity, we focus on SQGT codes with equidistant thresholds. The first approach relies on classical combinatorial methods, while the second approach relies on probabilistic methods. The second approach is of special interest for applications such as genotyping, where one cannot arbitrarily choose the test matrices. The tests are usually determined by the physics of the experiment, and only certain statistical properties of the tests are known. In this scenario, "structure" is to be seen as probabilistic trait. We show that one way to approach this problem is to characterize the number of tests that ensures that *almost all members of a code family possess a given trait and act as SQGT codes*.

A. Combinatorial Construction

Fix a binary u-disjunct code matrix $\mathbf{C}_b$ of dimensions $n_b \times N_b$, with code-length n_b and N_b codewords. Let $K = \left\lfloor \log_u \left(\left(\frac{q-1}{\eta} \right) (u - 1) + 1 \right) \right\rfloor$; construct a code of length $n = n_b$ and size $N = KN_b$ by concatenating K matrices, $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_K]$, where $\mathbf{C}_j = \left(\sum_{i=0}^{j-1} u^i \eta \right) \mathbf{C}_b$, $1 \leqslant j \leqslant K$.

Theorem 2: Let the concatenated code $\mathbf{C}$ be as described above. The code is capable of uniquely identifying up to u positives.

Proof: The proof is based on exhibiting a decoding procedure and showing that the procedure allows for distinguishing between any two different sets of positives. The decoder is described below.

Let $\mathbf{y}$ be the Q-ary vector of test outcomes, or equivalently, the syndrome of the positives. For a rational vector $\mathbf{z}$, let $\lfloor \mathbf{z} \rfloor$ and $\langle \mathbf{z} \rangle$ denote the vector of integer parts of $\mathbf{z}$ and fractional parts of $\mathbf{z}$, respectively. If $u = 1$, decoding reduces to finding the column of $\mathbf{C}$ equal to $\eta \mathbf{y}$. If $u > 1$, decoding proceeds as follows.

Step 1: Set $\mathbf{y}'_K = \mathbf{y}$ and form vectors $\mathbf{y}_j$, $1 \leqslant j \leqslant K$, using the rules:

$$\mathbf{y}_j = \left(\frac{u^j - 1}{u - 1} \right) \left\lfloor \left(\frac{u - 1}{u^j - 1} \right) \mathbf{y}'_j \right\rfloor, \tag{11}$$

and

$$\mathbf{y}'_{j-1} = \left(\frac{u^j - 1}{u - 1} \right) \left\langle \left(\frac{u - 1}{u^j - 1} \right) \mathbf{y}'_j \right\rangle. \tag{12}$$

Step 2: Identify the positives as follows: if the syndrome of a column of $\mathbf{C}_j$ is included in $\mathbf{y}_j$, declare the subject corresponding to that column positive. Declare the subject negative otherwise.

The result is obviously true for $u = 1$. Therefore, we focus on the case $u > 1$. First, using induction, one can prove that each $\mathbf{y}_j$, $1 \leqslant j \leqslant K$, is the syndrome of a subset of columns of $\mathbf{C}_j$ corresponding to positives. Let $\mathbf{C}'_j = [\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_j]$, where $1 \leqslant j \leqslant K$. Since the non-zero entries of $\mathbf{C}$ are multiples of η, $\eta \mathbf{y}$ is the sum of columns of $\mathbf{C}$ corresponding to a subset of positives. Also, the maximum value of the entries of $\mathbf{C}'_{K-1}$ equals $\eta \frac{u^{K-1} - 1}{u - 1}$. Since there are at most u positives, the maximum value of their sum does not exceed $\eta \frac{u^K - u}{u - 1}$. This bound is strictly smaller than $\eta \frac{u^K - 1}{u - 1}$, the minimum non-zero entry of $\mathbf{C}_K$. As a result, $\mathbf{y}_K$ is the syndrome of the positives with signatures in $\mathbf{C}_K$, and $\mathbf{y}'_{K-1}$ is the syndrome of positives with signatures in $\mathbf{C}'_{K-1}$. Similarly, it can be shown that $\forall j, 1 \leqslant j \leqslant K - 1$, $\mathbf{y}_j$ is the syndrome of the positives with signature in $\mathbf{C}_j$, and $\mathbf{y}'_{j-1}$ is the syndrome of the positives with signatures in $\mathbf{C}'_{j-1}$.

From Proposition 2, we know that each $\mathbf{C}_j$ is a $[q; Q; \eta; u]$-SQ-disjunct code. Consequently using step 2, one can uniquely identify the positives with signatures from $\mathbf{C}_j$. ■

Remark 4: The method described above can be used with any binary separable code for conventional GT or adder channel to generate a SQGT code.

Remark 5: All the constructions described in this paper are able to identify up to u positives in a pool of N subjects when $u \ll N$. However, when $0 \leqslant u \leqslant N$, one can construct non-binary codes with length n and asymptotic size of $N \sim (\lfloor \log \lfloor \frac{q-1}{\eta} \rfloor \rfloor + \log n/2)n$ (see the full version of this paper).

B. Probabilistic Construction

We consider the following problem: find a critical rate such that *any* randomly generated q-ary code with rate less than the critical rate is a $[q; Q; \eta; u]$-SQ-disjunct code *with probability close to one*. Based on the critical rate, which depends on the statistical properties of the process used to generate the codes, one can identify the smallest number of tests required to ensure that any code in the family may be used for SGGT.

Theorem 3: Let $R_{\text{critical}} = \frac{\log \gamma}{u+1} + \frac{\log(\epsilon u!)}{n(u+1)}$, $\forall \epsilon > 0$, where $\gamma = q^{(u+1)}/A$, $A = \eta \binom{(I-1)\eta+u}{u+1} - \eta \binom{\eta+u}{u+1} + (q-I\eta)\binom{I\eta+u-1}{u}$, and $I = \lfloor \frac{q-1}{\eta} \rfloor$. Any q-ary code of length n and size N with rate asymptotically satisfying $R \leqslant R_{\text{critical}}$ is a $[q; Q; \eta; u]$-SQ-disjunct code with probability at least $1 - \epsilon$.

Proof: Let $\mathcal{C}$ be a code of length n and size N, and let $\mathcal{M}$ be a set of $u+1$ codewords of $\mathcal{C}$. There are $L = \binom{N}{u+1}$ different ways to choose $\mathcal{M}$. For the ith choice of $\mathcal{M}$, we define E_i as the event that the syndrome of at least one of the codewords in $\mathcal{M}$ is included in the syndrome of the other u codewords. Suppose that $P(E_i) \leqslant p'$ for all i; using the union bound, $P\left(\bigcup_{i=1}^{L} E_i\right) \leqslant L p'$. Therefore, if

$$p' \leqslant \frac{\epsilon}{L}, \tag{13}$$

then $P\left(\bigcap_{i=1}^{L} \bar{E}_i\right) \geqslant 1 - \epsilon$, where $\bar{E}_i$ is the complement of the event E_i. In other words, $\mathcal{C}$ is a $[q; Q; \eta; u]$-SQ-disjunct code with probability at least $1 - \epsilon$.

From the definition of E_i, one has $P(E_i) = \frac{a}{(u+1)!\binom{q^n}{u+1}}$, where a is the number of q-ary matrices of size $n \times (u+1)$ that do not satisfy (10) and have distinct columns; also, $(u+1)!\binom{q^n}{u+1}$ is the total number of $n \times (u+1)$ matrices with distinct columns. In order to find an upper bound on a, we use the fact that a matrix that satisfies (10) has distinct columns. Consequently,

$$(u+1)!\binom{q^n}{u+1} - a = q^{n(u+1)} - b \tag{14}$$

where $q^{n(u+1)}$ is the number of q-ary matrices with (possibly) repeated columns, and b is the number of such matrices that satisfy (10). It can be easily seen that,

$$b \leqslant (u+1)c \tag{15}$$

where c is the number of q-ary matrices that do not contain a row, $\mathbf{x}$, satisfying $\lfloor \frac{x_1}{\eta} \rfloor > \lfloor \frac{\sum_{j=2}^{u+1} x_j}{\eta} \rfloor$. On the other hand, $c = A^n$ where A is the number of "acceptable" q-ary rows of length $u+1$. Let $\mathbf{x} \in [q]^{u+1}$ denote an acceptable row. If A_i, $i \in \{1, 2, \cdots, \lfloor \frac{q-1}{\eta} \rfloor\}$, denotes the number of acceptable rows with the first entry x_1 from $\{i\eta, i\eta+1, \cdots, (i+1)\eta-1\}$, then $A = \sum_i A_i$. Let $I = \lfloor \frac{q-1}{\eta} \rfloor$. If $i < I$, there are η choices for x_1; if $i = I$, we have $(q - I\eta)$ choices for x_1. The number of

ways to choose the rest of the entries (denoted by B_i) is

$$B_i = \sum_{k=0}^{i\eta-1} \binom{k+u-1}{u-1} = \binom{i\eta+u-1}{u} \tag{16}$$

where $\binom{k+u-1}{u-1}$ counts the number of non-negative integer solutions to $\sum_{j=2}^{u+1} x_j = k$. Consequently,

$$A = \eta \sum_{i=1}^{I-1} \binom{i\eta+u-1}{u} + (q-I\eta)\binom{I\eta+u-1}{u} \tag{17}$$

$$= \eta \binom{(I-1)\eta+u}{u+1} - \eta \binom{\eta+u}{u+1} + (q-I\eta)\binom{I\eta+u-1}{u}.$$

Using these results,

$$P(E_i) \leqslant 1 - \frac{q^{n(u+1)} - (u+1)A^n}{(u+1)!\binom{q^n}{u+1}} \leqslant \frac{A^n}{u!\binom{q^n}{u+1}}. \tag{18}$$

Note that the second inequality does not loosen the bound significantly since $1 - \frac{q^{n(u+1)} - (u+1)A^n}{(u+1)!\binom{q^n}{u+1}} \sim \frac{A^n}{u!\binom{q^n}{u+1}}$ as $n \to \infty$.

As $n, N \to \infty$, $p' \sim \frac{(u+1)A^n}{q^{n(u+1)}}$ and $L \sim \frac{N^{u+1}}{(u+1)!}$. Consequently, (13) asymptotically simplifies to

$$N^{u+1} \leqslant \gamma^n \epsilon u! \Rightarrow R \leqslant R_{\text{critical}} = \frac{\log \gamma}{u+1} + \frac{\log(\epsilon u!)}{n(u+1)}$$

where $\gamma = q^{(u+1)}/A$. This completes the proof. ∎

REFERENCES

[1] D. -Z. Du and F. K. Hwang, *Combinatorial Group Testing and Its Applications*, NJ: World Scientific, 2000.
[2] D. -Z. Du and F. K. Hwang, *Pooling Designs and Nonadaptive Group Tesging*, NJ: World Scientific, 2006.
[3] P. Z. Fan, M. Darnell, and B. Honary, "Superimposed Codes for the Multiaccess Binary Adder Channel," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1178-1182, 1995.
[4] R. Dorfman, "The Detection of positive Members of Large Populations," *Ann. Math. Stat.*, vol. 14, pp. 436-440, 1943.
[5] P. Damaschke, "Threshold Group Testing," *General Theory of Information Transfer and Combinatorics* in: LNCS, vol. 4123, pp. 707-718, 2006.
[6] D. Jevtić, "On Families of Sets of Integral Vectors Whose Representatives Form Sum-distinct Sets," *SIAM J. Discrete Math.*, vol. 8, no. 4, pp. 652-660, Nov. 1995.
[7] J. Cheng and Y. Watanabe, "T-User Code with Arbitrary Code Length for Multiple-Access Adder Channel," *IEICE Trans. Fundamentals*, vol. E82-A, no. 10, pp. 2011-2016, Oct. 1999.
[8] W. Kautz and R. Singleton, "Nonrandom Binary Superimposed Codes," *IEEE Trans. Inf. Theory*, vol. 10, no. 4, pp. 363-377, Oct. 1964.
[9] M. Malyutov, "Screening Designs for Non-Symmetric Response Functions," *Mat. Zametki*, vol. 27, pp. 109-127, 1980.
[10] A. G. D'yachkov, "Lectures on Designing Screening Experiments," *Lecture Note Series 10*, Feb. 2004.
[11] G. Atia and V. Saligrama, "Boolean Compressed Sensing and Noisy Group Testing," arXiv:0907.1061v4, 2010.
[12] B. Lindström, "Determining Subsets by Unramified Experiments," *A survey of statistical designs and linear models*, North-Holland, Amsterdam, 1975.
[13] D. Deng, D. R. Stinson, and R. Wei, "The Lovász Local Lemma and Its Applications to Some Combinatorial Arrays," *Designs, Codes and Cryptography*, vol. 32, pp. 121-134, May 2004.