Finding a Burst of Positives via Nonadaptive Semiquantitative Group Testing

Yun-Han Li*, Ryan Gabrys†, Jin Sima*, Ilan Shomorony* and Olgica Milenkovic*
*Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, USA
Email: {yunhanl2,jsima,ilans,milenkov}@illinois.edu
†Naval Information Warfare Center, San Diego
Email: ryan.gabrys@gmail.com

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206886

***Abstract*—Motivated by testing for pathogenic diseases we consider a new nonadaptive group testing problem for which: (1) positives occur within a burst, capturing the fact that infected test subjects often come in clusters, and (2) that the test outcomes arise from semiquantitative measurements that provide coarse information about the number of positives in any tested group. Our model generalizes prior work on detecting a single burst with classical group testing [1] to the setting of semiquantitative group testing (SQGT) [2]. Specifically, we study the setting where the burst-length ℓ is known and the semiquantitative tests provide potentially nonuniform estimates on the number of positives in a test group. The estimates represent the index of a quantization bin containing the (exact) total number of positives, for arbitrary thresholds $\eta_1, \ldots, \eta_s$. Interestingly, we show that the minimum number of tests needed for burst identification is essentially only a function of the largest threshold η_s. In this context, our main result is an order-optimal test scheme that can recover any burst of length ℓ using roughly $\lfloor \frac{\ell}{2\eta_s} \rfloor + \log_{s+1}(n)$ measurements. This suggests that a large saturation level η_s is more important than finely quantized information when dealing with bursts. We also provide results for related modeling assumptions and specialized choices of thresholds.**

I. Introduction

Group testing (GT) is a protocol for identifying relatively small subsets of marked elements, referred to as *positives*, within a larger collection of entities termed test subjects. The gist of the approach is to group subjects into carefully selected subpools and test the subjects in each subpool jointly so as to reduce the number of tests compared to that needed for individual testing. The first GT scheme comprising two stages of testing was described by Dorfman [3] in the context of finding individuals with venereal diseases. His scheme also represents the first instance of *adaptive testing*, where measurements from one round of testing can be used to inform the test selections in subsequent rounds. Unlike adaptive testing, *nonadaptive* GT requires that all tests be designed and conducted simultaneously. Since Dorfman's work, GT has been extended and generalized in many different directions and has found numerous applications in search systems, experimental and circuit design and computational biology. For comprehensive surveys, the interested reader is referred to [4], [5].

In [1], Colbourn considered a specialized GT technique for identifying one single burst of consecutive positives of length $\leq \ell$ within an ordered list of n elements. For nonadaptive techniques, Colbourn showed that the order-optimal number of measurements equals $\ell + \log(n)$. Follow-up works focused on improving some aspects of the scheme [6]–[8], extending the results to include new adaptive protocols [9], and generalizing the approach to handle multiple bursts [10].

However, in many real-life scenarios, such as testing for infections with viral pathogens based on quantitative PCR (quantitative polymerase chain reaction, qPCR), the outcomes are real-valued and usually confined to an interval such as $[10, 45]$. A measurement is known as the C_t (cycle threshold) value and it conveys information about how likely an individual is to be infected. For example, a C_t value close to 40 is highly indicative of a negative subject, while a value below 20 is a strong sign that the individual is highly virulent. One can therefore quantize the C_t values using a carefully selected collection of s thresholds $\underline{\eta} = (\eta_1, \ldots, \eta_s)$ so that each quantization bin provides an estimate of the viral load in the pool and, consequently, an estimate of the number of positives in the pool. This type of GT approach is known as semiquantitative GT (SQGT) [2], [11]. Furthermore, whenever testing is done on large populations in which individuals that cohabitate are naturally adjacent in the order used for testing [12] (for example, families, dorm-mates etc.), bursty positive models are appropriate and can result in significant savings compared to classical GT approaches [1].

Given the additional quantitative information and the assumption regarding consecutive orderings of positives, one can easily envision performing SQGT for bursty positives that quantizes the C_t values into quantization bins that indicate the level of the viral load, or an estimate of the number of infected individuals in the population [13], [14]. Here, for the first time, we study the reduction in the number of group measurements achievable in such a nonadaptive setting. In particular, we investigate two new bursty SQGT models [2], one in which the length of the burst is known and fixed to ℓ (henceforth referred to as the fixed-length burst model, $B(n, \ell, \underline{\eta})$); and another, in which the length of the burst is known to be upper-bounded by ℓ (henceforth referred to as the bounded-length burst model, $B(n, \leq \ell, \underline{\eta})$).

Our main contributions include

1) Order-optimal constructions (i.e., constructive lower and upper bounds that differ by a constant factor of 2) for the $B(n, \ell, \underline{\eta})$ setting with quantization thresholds $\underline{\eta}$ for which $\ell = \Omega\left(\eta_s \log_2(\eta_s)\right)$.
2) Order-optimal constructions (i.e., lower and upper bounds that differ by a constant factor of 4) for the nonadaptive $B(n, \leq \ell, \underline{\eta})$ setting with SQGT thresholds $\underline{\eta} = (1, \ldots, s)$ corresponding to the so-called *saturation model* [13], [15].

Two important comments are in place. Semiquantitative measure-

ments significantly decrease the number of tests needed for the $B(n, \leq \ell, \underline{\eta})$ setting (the improvement is linear in the number of thresholds s). Somewhat surprisingly, for the $B(n, \ell, \underline{\eta})$ setting the number of tests is basically determined by the value of the largest threshold η_s rather than by the total number of thresholds s. These findings may have interesting consequences for test schedules and quantization schemes used for practical quantitative PCR protocols.

The paper is organized as follows. Section II introduces the notation and provides the formal problem formulation. Section III contains the results for the lower bounds, while Section IV contains the main results of the work, pertaining to upper bounds on the number of SQGT burst identification models for a fixed and upper-bounded length of the burst.

II. Problem Formulation

We start by introducing the relevant notation as well as the fixed-length and bounded-length single burst identification problems under SQGT measurements.

Let $h_{\mathbf{M}}, w_{\mathbf{M}}, \mathbf{M}(*, i), \mathbf{M}(j, *)$ denote the number of rows (height), number of columns (width), i-th column and j-th row of the matrix $\mathbf{M}^{h_{\mathbf{M}} \times w_{\mathbf{M}}}$, respectively. Our row indices lie in $[0, h_{\mathbf{M}} - 1]$, while the column indices are confined to $[0, w_{\mathbf{M}} - 1]$. In addition, $\mathcal{R}(\mathbf{M}), \mathbf{M}^c, \mathbf{M}^\infty$ are used to denote a matrix obtained from $\mathbf{M}$ by reversing the column order (so that $\mathcal{R}(\mathbf{M})(*, i) = \mathbf{M}(*, w_{\mathbf{M}} - i - 1)$, a c-fold horizontal concatenation of $\mathbf{M}$ (i.e, $[\mathbf{M}, \ldots, \mathbf{M}]$ with c constituent matrices), and an infinitely horizontal concatenation of matrices $\mathbf{M}$ such that $w_{\mathbf{M}\infty} = \infty$. Finally, we use $\mathbf{M}(i, j)$ to denote the entry in $\mathbf{M}$ in row i and column j.

The single burst of positives problem requires introducing the following notions.

Bursts: A burst is denoted by a binary $n \times 1$ column vector $\boldsymbol{b}$, and is specified by a head and tail $h_{\boldsymbol{b}} \leq t_{\boldsymbol{b}}$, which dictate its length $\ell_{\boldsymbol{b}} = t_{\boldsymbol{b}} - h_{\boldsymbol{b}} + 1$. It comprises consecutive positives:

$$\boldsymbol{b}(i) = \begin{cases} 0, & 0 \leq i < h_{\boldsymbol{b}}, \\ 1, & h_{\boldsymbol{b}} \leq i \leq t_{\boldsymbol{b}}, \\ 0, & t_{\boldsymbol{b}} < i \leq n. \end{cases}$$

When $\ell_{\boldsymbol{b}}$ is fixed, $\boldsymbol{b}^i$ denotes the burst with $h_{\boldsymbol{b}} = i$, and the distance between two burst $\boldsymbol{b}^i, \boldsymbol{b}^j$ is defined as the difference of their head position $|i - j|$.

SQGT measurements: An SQGT measurement is described by a $1 \times n$ binary vector $\boldsymbol{m}$ such that

$$\boldsymbol{m}(i) = \begin{cases} 1, & i\text{th element is included in the test}, \\ 0, & \text{otherwise}, \end{cases}$$

and a set of integer-valued quantized thresholds

$$\underline{\eta} = (\eta_1, \ldots, \eta_s) \text{ with } 0 < \eta_1 < \ldots < \eta_s \leq n,$$

such that the SQGT measurement outcomes equal

$$\underline{\eta}(\boldsymbol{mb}) = \begin{cases} 0, & 0 \leq \boldsymbol{mb} < \eta_1, \\ i, & \eta_i \leq \boldsymbol{mb} < \eta_{i+1}, \\ s, & \eta_s \leq \boldsymbol{mb} \leq n. \end{cases}$$

Definition 2.1: When $\underline{\eta} = (1, 2, \ldots, s)$, we refer to this specialized SQGT scheme as the *saturation* SQGT model.

Correct burst detection: for any hidden burst $\boldsymbol{b}$, the estimate $\hat{\boldsymbol{b}}$ generated by the detection algorithm should equal $\boldsymbol{b}$.

Definition 2.2: $B(n, \ell, \underline{\eta})$ and $B(n, \leq \ell, \underline{\eta})$ are used to denote the fixed-length and bounded-length burst problem with burst-lengths $= \ell$ and $\leq \ell$, respectively, and with n test elements and SQGT quantized thresholds $\underline{\eta}$.

A nonadaptive SQGT testing scheme with m measurements on n elements is represented by a $m \times n$ binary measurement matrix $\mathbf{M}$ with each row corresponding to a single SQGT measurement. We say $\mathbf{M}$ solves the $B(n, \ell, \underline{\eta})$ (or the $B(n, \leq \ell, \underline{\eta})$) problem if and only if

$$\forall \boldsymbol{b} \neq \boldsymbol{b}' \text{ allowed by the } B(n, \ell, \underline{\eta}) \ (B(n, \leq \ell, \underline{\eta}) \ ) \text{ problem, one has } \underline{\eta}(\mathbf{M}\boldsymbol{b}) \neq \underline{\eta}(\mathbf{M}\boldsymbol{b}').$$

The smallest possible number of measurements possible to meet this requirement, among all nonadaptive SQGT schemes is denoted by $m^*_{B(n, \ell, \underline{\eta})}$ and $m^*_{B(n, \leq \ell, \underline{\eta})}$.

Our constructions will make use of Gray codes (also used in [1]) and generalizations thereof. However, The ways for using Gray codes are different. The way of [1] can handle multiple lengths but only leverage a single threshold. On the other hand, our way (Theorem 2) can only handle a single length but leverage all thresholds. Our construction leverage Fact 2.1 and Fact 2.2. To learn more about Gray code, see [16]. We say that $\mathbf{G}_{s,h} \in \{0, \ldots, s\}^{h \times s^h}$ represents an s-ary Gray code with length h if it satisfies the following two conditions:

1) Any two consecutive columns differ in exactly one position, and the difference has magnitude one.
2) $\mathbf{G}_{s,h}$ includes all possible s^h codewords exactly once.

Example 2.1: The following matrix has columns that constitute a 3-ary Gray code of length two:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 2 \end{bmatrix}.$$

Fact 2.1: The Gray code $\mathbf{G}_{s,h}$ can be constructed by first recursively constructing paired Gray code matrices $\mathbf{P}_{s,h} := [\mathbf{G}_{s,h}, \mathcal{R}(\mathbf{G}_{s,h})]$ using the rule below and then removing half of the columns from the right side:

$$\begin{cases} \mathbf{P}_{s,1} = [0, \ldots, s-1, s-1, \ldots, 0], \\ \mathbf{P}_{s,i} = \begin{bmatrix} \mathbf{P}_{s,1} \otimes \mathbf{1}^{s^{i-1}} \\ \mathbf{P}^s_{s,i-1\cdot} \end{bmatrix} \end{cases} \tag{1}$$

Here, $\otimes$ stands for the Kronecker product while $\mathbf{1}^a$ is a row vector of 1s.

Example 2.2: The following matrix $\mathbf{P}_{3,2}$ is constructed recursively using (1). The left half, as claimed, equals $\mathbf{G}_{3,2}$ and was illustrated in Example 2.1:

$$\begin{bmatrix} 000 & 111 & 222 & 222 & 111 & 000 \\ 012 & 210 & 012 & 210 & 012 & 210 \end{bmatrix}.$$

We also make use of the following property of binary Gray codes.

Fact 2.2: $\mathbf{G}_{2,h}((i, *))$ contains 2^{i-1} runs of 1s for all i except $i = 0$, which contains only one run of 1s. Consequently, the matrix contains a total of $\sum_{i=1}^{h-1} 2^{i-1} + 1 = 2^{h-1}$ runs of 1s

within its rows. This is illustrated by the following example for $\mathbf{G}_{2,3}$, with a total number of $2^{3-1}=4$ runs of 1s.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & [1 & 1 & 1 & 1] \\ 0 & 0 & [1 & 1 & 1 & 1] & 0 & 0 \\ 0 & [1 & 1] & 0 & 0 & [1 & 1] & 0 \end{bmatrix}.$$

III. Lower Bounds

We first provide lower bounds for the smallest number of measurements needed for the $m^*_{B(n,\ell,\underline{\eta})}$ and $m^*_{B(n,\leq\ell,\underline{\eta})}$ settings. The proofs mostly use ideas from [1].

Theorem 1: We have

$$\begin{cases} m^*_{B(n,\ell,\underline{\eta})} \geq \max\left(\log_{s+1}(n-\ell+1), \lceil \frac{\ell}{2\eta_s} \rceil\right), \\ m^*_{B(n,\leq\ell,\underline{\eta})} \geq \max\left(\log_2(n), \lceil \frac{\ell}{s} \rceil\right). \end{cases}$$

The proof technique used for $m^*_{B(n,\leq\ell,\underline{\eta})}$ is similar to that for $m^*_{B(n,\ell,\underline{\eta})}$; hence, we only provide the proof for $m^*_{B(n,\ell,\underline{\eta})}$. We prove the first bound by establishing each of the bounds on the right-hand side separately and combining them via maximization.

1) The bound $\log_{s+1}(n-\ell+1)$ follows from a simple counting argument: there are a total of $n-\ell+1$ different head positions and a total of $s+1$ possible outcomes for each measurement.
2) The bound $\lceil \frac{\ell}{2\eta_s} \rceil$: we show that even if we only require to discriminate among the first $\ell+1$ bursts (i.e., bursts $\boldsymbol{b}^i$ with $0 \leq i \leq \ell$), we still need $\lceil \frac{\ell}{2\eta_s} \rceil$ measurements. For any measurement $\boldsymbol{m}$, let $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ denote the first and second block of ℓ bits of $\boldsymbol{m}$. Only the last η_s nonzero bits in $\boldsymbol{m}_1$ and the first η_s nonzero bits in $\boldsymbol{m}_2$ are relevant. For simplicity, we only provide a proof for the $\boldsymbol{m}_2$ case. Let $\ell \leq i_1 < i_2 < \ldots < 2\ell$ be the elements included in $\boldsymbol{m}_2$. Since $\ell_{\boldsymbol{b}} = \ell$ and $h_{\boldsymbol{b}} \leq \ell$, if i_j is included in the burst $\boldsymbol{b}$ then $i_1, \ldots, i_{j-1}$ must also be included. Therefore, if $j > \eta_s$, by observing that η_s is the largest threshold, one can remove i_j from $\boldsymbol{m}_2$ without changing the outcome. Hence one can only retain the first η_s nonzero bits in $\boldsymbol{m}_2$ and still arrive at the same outcome. As a result, it suffices to only consider those $\boldsymbol{m}$ for which $\sum_{j=0}^{2\ell-1} \boldsymbol{m}(j) \leq 2\eta_s$. Let $\mathbf{M}^{h_{\mathbf{M}} \times 2\ell}$ be our measurement matrix restricted to the first 2ℓ columns. Suppose that $h_{\mathbf{M}} < \frac{\ell}{2\eta_s}$; then $\sum_{i=0}^{h_{\mathbf{M}}-1} \sum_{j=0}^{2\ell-1} \mathbf{M}(i,j) \leq 2\eta_s h_{\mathbf{M}} < \ell$. This implies that there exists a $0 \leq j < \ell$ such that $\mathbf{M}(*,j) = \mathbf{M}(*,j+\ell) = \mathbf{0}^{h_{\mathbf{M}} \times 1}$. Then $\underline{\eta}\left(\mathbf{M}\boldsymbol{b}^{i+1}\right) = \underline{\eta}\left(\mathbf{M}\boldsymbol{b}^i + \mathbf{M}(*,j+\ell) - \mathbf{M}(*,j)\right) = \underline{\eta}\left(\mathbf{M}\boldsymbol{b}^i\right)$. Therefore $m^*_{B(n,\ell,\underline{\eta})} \geq \lceil \frac{\ell}{2\eta_s} \rceil$.

IV. Main Results

In Section IV-A, we describe an order-optimal construction of measurement matrices for the $B(n,\ell,\underline{\eta})$ problem pertaining to two different cases, the case of general SQGT thresholds with $\ell = \Omega(\eta_s \log(\eta_s))$ and the saturation model with $\ell \leq \eta_s = s$. It is interesting to note that for the first case, $m^*_{B(n,\ell,\underline{\eta})}$ basically depends only on the largest threshold η_s. In other words, as long as $\ell = \Omega(\eta_s \log_2(n))$ with a sufficiently large constant, there is no benefit of using multiple thresholds (SQGT) compared to threshold group testing (TGT) with the single biggest threshold η_s. In Section IV-B, we describe an order-optimal scheme (within an approximation constant 4) for the $B(n,\leq\ell,\underline{\eta})$ problem and the saturation model.

A. The $B(n,\ell,\underline{\eta})$ Model

Since for this case the burst length is fixed, one only needs to locate the position of the head $h_{\boldsymbol{b}} \in [0, n-\ell]$ of the burst $\boldsymbol{b}$. Vaguely speaking, the near-optimal construction follows a two-step sketch-and-refine procedure, also used in classical binary burst GT detection [1]. The first part, referred to as the *General Sketch Scheme*, uses a measurement matrix $\mathbf{K}$ (Theorem 2) that produces different outcomes for all potential bursts whose starting locations are separated by $> \ell+1$ positions. The second part, referred to as the *General Refinement Scheme*, uses a measurement matrix $\mathbf{R}$ (Theorem 3) that produces different outcomes for all potential bursts whose starting locations are separated by $< 2\ell$ positions. Stacking the two measurement matrices leads to the result reported in Theorem 4.

Theorem 2: For $B(n,\ell,\underline{\eta})$, the measurement matrix $\mathbf{K}$ described in Section IV-A1 produces different outcomes for all potential bursts whose starting locations are at distances $> \ell+1$ using $\lceil \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) \rceil$ measurements.

Theorem 3: For $B(n,\ell,\underline{\eta})$ with parameters $\eta_s = 2^{h-1}+2$ and $\ell = c2^h + 1$, where $c, h \in \mathbb{N}$ and $c > 2(h+1)$, the measurement matrix $\mathbf{R}$ described in Section IV-A2 produces different outcomes for all potential bursts whose starting locations are at distances $< 2\ell$ using $\leq \frac{\ell}{2\eta_s} + 1$ measurements. The construction depends only on the largest threshold η_s.

Theorem 4: Combining the General Sketch matrix of Theorem 2 and the General Refinement matrix of Theorem 3, leads to the measurement matrix $[\mathbf{K}^\intercal, \mathbf{R}^\intercal]^\intercal$ which can be used to solve $B(n,\ell,\underline{\eta})$ using $\leq \frac{\ell}{2\eta_s} + \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) + 2$ measurements whenever $\ell > 4\eta_s \log_2(4\eta_s)$. This number of measurements is at most twice the number of measurements from the lower bounds reported in Theorem 1.

Remark 4.1: Note that the scheme from Theorem 3 only uses the largest threshold η_s. Therefore, if we only make use of η_s in the General Sketch Scheme of Theorem 2, the resulting measurement matrix has height roughly $\frac{\ell}{2\eta_s} + \log_2\left(\frac{n-\ell+1}{\ell}\right)$ and depends on one threshold, η_s. When $\ell = \Omega(\eta_s \log_2(n))$ with a sufficiently large constant, $\frac{\ell}{2\eta_s} + \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) = \Omega\left(\log_2(n)\right) = \frac{\ell}{2\eta_s} + \log_2\left(\frac{n-\ell+1}{\ell}\right)$. Therefore, in this parameter regime, there is no benefit from using multiple thresholds.

1) Proof of Theorem 2: We start with some relevant notation. Let $\mathbf{K}^{h_{\mathbf{K}} \times n}$ be the measurement matrix. We say that $\mathbf{K}^{h_{\mathbf{K}} \times n}$ results in the outcome matrix $\mathbf{O}^{h_{\mathbf{K}} \times n-\ell+1}$ if $\mathbf{O} = \left[\underline{\eta}\left(\mathbf{K}\boldsymbol{b}^0\right) \ldots \underline{\eta}\left(\mathbf{K}\boldsymbol{b}^{n-\ell+1}\right)\right]$ represents the collection of outcomes for all length-ℓ bursts $\boldsymbol{b}^i$ when using the measurement matrix $\mathbf{K}$.

Next, let $\vec{B}_\ell(x) := \mathbf{0}^{\ell-x}\mathbf{1}^x$ and $\overleftarrow{B}_\ell(x) := \mathbf{1}^x\mathbf{0}^{\ell-x}$, for all $x \in \{0,\ldots,\ell\}$. Also, let $\vec{B}_\ell(0)^i, \overleftarrow{B}_\ell(0)^i$ stand for the horizontal concatenation of i copies of $\vec{B}_\ell(x)$ and $\overleftarrow{B}_\ell(x)$. Observe that $\lceil \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) \rceil$ (almost) matches the counting bound $\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)$. The key idea is to first construct $\mathbf{K}$ with $w_{\mathbf{K}} \geq n$ such that the outcome matrix satisfies

$$\mathbf{O} = \mathbf{G}_{s+1,h_{\mathbf{K}}} \otimes \mathbf{1}^{\ell+1}, \tag{2}$$

and then truncate it to n columns. By the definition of $\mathbf{O}$, $\mathbf{K}$ can identify all bursts at distance $\geq \ell+1$ if and only if all columns

of $\mathbf{G}_{s+1,h_{\mathbf{K}}}$ are different; that this is true follows from the fact that Gray codes include all vectors $\{0,\ldots,s\}^{h_{\mathbf{K}}}$ exactly once. We need the following lemma for our subsequent derivations.

Lemma 4.1: The following measurement

$$\boldsymbol{m}(c) := \Big[\vec{B}_\ell(0)^c\, 0\vec{B}_\ell(\eta_1)^c \ldots 0\vec{B}_\ell(\eta_s)^c \\ 1\overleftarrow{B}_\ell(\ell)^c\, 1\overleftarrow{B}_\ell(\eta_s-1)^c \ldots 1\overleftarrow{B}_\ell(\eta_1-1)^c\, 0\Big]^\infty$$

results in the outcome $\left([0,\ldots,s,s,\ldots,0]\otimes\mathbf{1}^{c\ell+1}\right)^\infty$.

Proof: The case $c=1$ can be proved easily and is illustrated by the following example. For $\underline{\eta}=(1,2,4)$ and $\ell=6$, $\boldsymbol{m}(1)$ equals

$$(000000\ 0000001\ 0000011\ 0001111 \\ 1111111\ 1111000\ 1100000\ 1000000\ 0)$$

For $c>1$, and any length-ℓ row-vector $\boldsymbol{x}$, $\underline{\eta}\left(\boldsymbol{x}^c\boldsymbol{b}^i\right)$ remains unchanged for all $\boldsymbol{b}^i$, where $i\in[0,(c-1)\ell]$.

$$\forall i,\ \underline{\eta}\left(\boldsymbol{x}^c\boldsymbol{b}^{i+1}\right)=\underline{\eta}\left(\boldsymbol{x}^c\boldsymbol{b}^i+\boldsymbol{x}^c(*,i+\ell)-\boldsymbol{x}^c(*,i)\right)=\underline{\eta}\left(\boldsymbol{x}^c\boldsymbol{b}^i\right)$$

■

We are now ready to present our construction. Let $\mathbf{M}[i]$ be the measurement matrix recursively constructed as follows:

$$\begin{cases}\mathbf{M}[1]=\boldsymbol{m}(1),\\ \mathbf{M}[i]=\begin{bmatrix}\boldsymbol{m}\left(\frac{(\ell+1)(s+1)^{i-1}-1}{\ell}\right)\\ \mathbf{M}[i-1]^{s+1}\end{bmatrix}.\end{cases} \tag{3}$$

Note that $\frac{(\ell+1)(s+1)^{i-1}-1}{\ell}$ may not be an integer. We therefore first focus on the special case $s=\ell$ (therefore $\frac{(\ell+1)(s+1)^{i-1}-1}{\ell}=\frac{(\ell+1)^i-1}{\ell}\in\mathbb{N}$) to illustrate the idea and then generalize the result for $s<\ell$ through slight modifications of the argument in Theorem 6.

Lemma 4.2: For $s=\ell$, $\mathbf{M}[i]^\infty$ results in the outcome matrix $\left(\mathbf{P}_{\ell+1,i}\otimes\mathbf{1}^{\ell+1}\right)^\infty$. Where $\mathbf{P}_{\ell+1,i}$ is the $\ell+1$-ary length-i paired gray code matrix.

Proof: The proof is by induction.

1) For $i=1$: by Lemma 4.1, $\mathbf{M}[i]^\infty=\boldsymbol{m}(1)^\infty$ results in the outcome matrix $\left([0,\ldots,\ell,\ell,\ldots,0]\otimes\mathbf{1}^{\ell+1}\right)^\infty=\left(\mathbf{P}_{\ell+1,1}\otimes\mathbf{1}^{\ell+1}\right)^\infty$.
2) For $i>1$: Suppose that the claim holds for $i-1$. Then

$$\mathbf{M}[i]^\infty=\begin{bmatrix}\boldsymbol{m}\left(\frac{(\ell+1)^i-1}{\ell}\right)\\ \mathbf{M}[i-1]^{\ell+1}\end{bmatrix}^\infty$$

results in the outcome matrix

$$\begin{bmatrix}[0,\ldots,\ell,\ell,\ldots,0]\otimes\mathbf{1}^{(\ell+1)^i}\\ \mathbf{P}^{\ell+1}_{\ell+1,i-1}\otimes\mathbf{1}^{\ell+1}\end{bmatrix}^\infty=\left(\mathbf{P}_{\ell+1,i}\otimes\mathbf{1}^{\ell+1}\right)^\infty.$$

■

By Lemma 4.2, $\mathbf{M}[h_{\mathbf{K}}]$ truncated to $(\ell+1)^{h_{\mathbf{K}}+1}+\ell-1$ columns from the right results in the outcome matrix $\mathbf{G}_{\ell+1,h_{\mathbf{K}}}\otimes\mathbf{1}^{\ell+1}$, and consequently produces different outcomes for all potential bursts whose starting locations are within distance $>\ell+1$. It is not hard to show that a single measurement $\mathbf{0}^\ell\left(\mathbf{1}^{\ell+1}\mathbf{0}^{\ell+1}\right)^\infty$ results in the outcome matrix $(0,\ldots,\ell,\ell,\ldots,0)^\infty$. Hence we have the following theorem.

Theorem 5: For the saturation SQGT model with ℓ thresholds $\underline{\eta}=(1,\ldots,\ell)$, $\begin{bmatrix}\mathbf{M}\left[\lceil\log_{\ell+1}(n-\ell+1)-1\rceil\right]^\infty\\ \mathbf{0}^\ell\left(\mathbf{1}^{\ell+1}\mathbf{0}^{\ell+1}\right)^\infty\end{bmatrix}$ truncated to n columns on the right can be used as the test matrix for the $B(n,\ell,\underline{\eta})$ model with $\lceil\log_{\ell+1}(n-\ell+1)\rceil$ measurements.

For the case $s<\ell$, some modifications in the recursion given by (3) are required. The modification involves truncating a certain number of columns from the left, right, or both sides of $\mathbf{M}'[i]$ at each stage of recursion i:

$$\begin{cases}\mathbf{M}'[1]=\boldsymbol{m}(1),\\ \mathbf{M}'[i]=\begin{bmatrix}\mathbf{0}^{\alpha \bmod \ell} & \boldsymbol{m}\left(\lfloor\frac{\alpha}{\ell}\rfloor\right) & \mathbf{0}^{\alpha \bmod \ell}\\ \mathbf{M}'_r[i-1] & \mathbf{M}'_{lr}[i-1]^{s-1} & \mathbf{M}'_l[i-1]\end{bmatrix},\end{cases}$$

where $\alpha=\frac{w_{\mathbf{M}'[i-1]}}{2}-1$, and $\mathbf{M}'_r[i-1],\mathbf{M}'_l[i-1],\mathbf{M}'_{lr}[i-1]$ denotes $\mathbf{M}'[i-1]$ truncate $\alpha \bmod \ell$ columns from the right, left, and both sides, respectively.

By using a similar proof as the one described above and some simple but tedious calculations, one can show that $\mathbf{M}'\left[\lceil\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)\rceil\right]$ truncated to n columns from the right can be used as $\mathbf{K}$. Hence we have the following result.

Theorem 6: For $s<\ell$, $\mathbf{M}'\left[\lceil\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)\rceil\right]$ truncated to n columns from the right produces different outcomes for all potential bursts whose starting locations are at distance $>\ell+1$ using $\lceil\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)\rceil$ measurements.

2) Proof of Theorem 3: We now focus our attention on the General Refinement Scheme. Let $\overleftarrow{\mathbf{B}}$ to denote the cyclic shift of columns in $\mathbf{B}$ one position to the left so that $\overleftarrow{\mathbf{B}}(*,i)=\mathbf{B}(*,i+1 \bmod \ell)$. We need the following lemma.

Lemma 4.3: Suppose that a binary matrix $\mathbf{B}^{h_{\mathbf{B}}\times\ell}$ satisfies the following three conditions:

1) All columns and their binary complement $\{\mathbf{B}(*,i),\bar{\mathbf{B}}(*,i)\}_{i=0}^{\ell-1}$ are distinct.
2) The first column is the zero vector, $\mathbf{B}(*,0)=\mathbf{0}^{h_{\mathbf{B}}\times1}$.
3) Each row of $\overleftarrow{\mathbf{B}}-\mathbf{B}$ has η_s-1 elements equal to -1.

Then the following measurement matrix produces different outcomes for all potential bursts $\boldsymbol{b}\neq\boldsymbol{b}'$ whose starting locations are at distance $<2\ell$:

$$\mathbf{R}:=\left[\mathbf{R}^-\mathbf{R}^+\mathbf{R}^-\mathbf{R}^+\ldots\right], \tag{4}$$

where $\mathbf{R}^-$ denotes the "negative" part of $\overleftarrow{\mathbf{B}}-\mathbf{B}$ (obtained by setting 1s to 0s and -1s to 1s), while $\mathbf{R}^+$ denotes the positive part of $\overleftarrow{\mathbf{B}}-\mathbf{B}$ (obtained by setting -1s to 0s), and the last column is changes from $\mathbf{0}^{h_{\mathbf{B}}\times1}$ to $\mathbf{1}^{h_{\mathbf{B}}\times1}$ (note that the last column of $\overleftarrow{\mathbf{B}}-\mathbf{B}$ before the modification is $\mathbf{B}(*,0)-\mathbf{B}(*,\ell-1)=-\mathbf{B}(*,\ell-1)$, which implies that the positive part is zero).

Proof: Since $\mathbf{R}$ is a repeated horizontal concatenation of $\mathbf{R}^-$ and $\mathbf{R}^+$, it suffices to show that

$$\forall 0\le i\neq j<2\ell,\ \underline{\eta}\left(\mathbf{R}\boldsymbol{b}^i\right)\neq\underline{\eta}\left(\mathbf{R}\boldsymbol{b}^j\right). \tag{5}$$

In particular, we prove that

$$\mathbf{R}\boldsymbol{b}^i=\begin{cases}(\eta_s-1)\,\mathbf{1}^{h_{\mathbf{B}}\times1}+\mathbf{B}(*,i) & 0\le i<\ell,\\ \eta_s\mathbf{1}^{h_{\mathbf{B}}\times1}-\mathbf{B}(*,i) & \ell\le i<2\ell.\end{cases} \tag{6}$$

Note that each entry of $\mathbf{R}\boldsymbol{b}^i$ is either η_s-1 or η_s. By condition 1, all $\mathbf{R}\boldsymbol{b}^i$ are different. Consequently, all $\underline{\eta}\left(\mathbf{R}\boldsymbol{b}^i\right)$ are different as well. Therefore, $\mathbf{R}$ produces different outcomes for all potential bursts whose starting locations are at distance $<2\ell$ using $h_{\mathbf{B}}$

measurements; only the largest threshold η_s is relevant. Next we prove (6).

For $0 \le i < \ell$,

$$\begin{aligned}\mathbf{R}\boldsymbol{b}^i - \mathbf{R}\boldsymbol{b}^0 &= \sum_{j=1}^{i} \mathbf{R}\boldsymbol{b}^j - \mathbf{R}\boldsymbol{b}^{j-1} = \sum_{j=1}^{i} \mathbf{R}\left(\boldsymbol{b}^j - \boldsymbol{b}^{j-1}\right) \\ &= \sum_{j=1}^{i} \mathbf{R}^+(*, j-1) - \mathbf{R}^-(*, j-1) \\ &= \sum_{j=1}^{i} \mathbf{R}(*, j-1) \overset{(a)}{=} \mathbf{B}(*, i).\end{aligned}$$

For $\ell \le i < 2\ell$,

$$\begin{aligned}\mathbf{R}\boldsymbol{b}^i - \mathbf{R}\boldsymbol{b}^\ell &= \sum_{j=\ell+1}^{i} \mathbf{R}\boldsymbol{b}^j - \mathbf{R}\boldsymbol{b}^{j-1} = \sum_{j=\ell+1}^{i} \mathbf{R}\left(\boldsymbol{b}^j - \boldsymbol{b}^{j-1}\right) \\ &= \sum_{j=\ell+1}^{i} \mathbf{R}^-(*, j-1) - \mathbf{R}^+(*, j-1) \\ &= -\sum_{j=1}^{i-\ell} \mathbf{R}(*, j-1) \overset{(a)}{=} -\mathbf{B}(*, i).\end{aligned}$$

The equalities $\overset{(a)}{=}$ follows from Condition 2. Finally, by Condition 3 and the fact that we changed the last column of $\mathbf{R}^+$ from $\mathbf{0}^{h_\mathbf{B}\times 1}$ to $\mathbf{1}^{h_\mathbf{B}\times 1}$, $\mathbf{R}\boldsymbol{b}^0 = (\eta_s - 1)\,\mathbf{1}^{h_\mathbf{B}\times 1}$ and $\mathbf{R}\boldsymbol{b}^\ell - \mathbf{R}\boldsymbol{b}^0 = \mathbf{1}^{h_\mathbf{B}\times 1} \Rightarrow \mathbf{R}\boldsymbol{b}^\ell = \eta_s \mathbf{1}^{h_\mathbf{B}\times 1}$. ∎

It remains to construct a matrix $\mathbf{B}$ that satisfies Conditions 1-3 in Lemma 4.3 with $h_\mathbf{B}$ roughly equal to $\frac{\ell}{2\eta_s}$. Let $\mathbf{G}_{2,h}^{h\times 2^h}$ be the code matrix of a binary Gray code of length h such that $2(h+1) < h_\mathbf{B}$. We construct $\bar{\mathbf{G}}_{2,h,i}^{h_\mathbf{B}\times 2^h}$ as

$$\bar{\mathbf{G}}_{2,h,i}^\intercal = \left[\mathbf{0}^{2^h\times i} \quad \mathbf{1}^{2^h\times 1} \quad \mathbf{G}_{2,h}^\intercal \quad \mathbf{0}^{(2^h\times h_\mathbf{B}-i-h-1)}\right]^\intercal.$$

Then, $\mathbf{B}$ is constructed as follows:

$$\mathbf{B} = \left[\mathbf{0}^{h_\mathbf{B}\times 1} \quad \bar{\mathbf{G}}_{2,h,h_\mathbf{B}-1} \quad \dots \quad \bar{\mathbf{G}}_{2,h,0}\right].$$

Example 4.1: The matrix $\mathbf{B}$ is constructed using $\mathbf{G}_{2,2}$ and for $h_\mathbf{B} = 7$:

$$\begin{bmatrix}
0 & 0011 & 0110 & & & & & 1111 \\
0 & 0110 & & & & & 1111 & 0011 \\
0 & & & & & 1111 & 0011 & 0110 \\
0 & & & & 1111 & 0011 & 0110 & \\
0 & & & 1111 & 0011 & 0110 & & \\
0 & & 1111 & 0011 & 0110 & & & \\
0 & 1111 & 0011 & 0110 & & & &
\end{bmatrix}$$

- Condition 1: we demonstrate a 2-step procedure for recovering the index of each column based on its content which establishes that all columns are different. The idea is first to use $\mathbf{1}^{1\times 2^h}$ to recover i (the index of $\bar{\mathbf{G}}_{2,h,i}$) and then use the following h bits from the Gray code to locate the exact column. To do so, in the first step, we need an additional constraint $h_\mathbf{B} \ge 2(h+1)$. In a nutshell, we can recover i by looking at the first 1 after a length-$h+1$ burst of 0 in each column. Moreover, by this constraint, each column in $\mathbf{B}$ has more zeros than ones. Consequently, all $\mathbf{B}(*, i)$ and $\bar{\mathbf{B}}(*, i)$ must be distinct.
- Condition 2: this condition is easy to verify.
- Condition 3: by the construction from (IV-A2), each row $\mathbf{B}(i)$ is a horizontal cyclic shift of

$$\boldsymbol{v} := \left[\mathbf{1}^{2^h}, \mathbf{G}_{2,h}(0, *), \dots, \mathbf{G}_{2,h}(h-1, *)\right]$$

with an additional 0 appended at the left. In Example 4.1, we have $\boldsymbol{v} = [111100110110]$. By this construction, the number of -1s in each row of $\bar{\mathbf{B}} - \mathbf{B}$ equals the number of runs of 1s in $\mathbf{B}(i, *)$, which equals the sum of the number of runs of 1s in $\mathbf{1}^{1\times 2^h}$ and each row of the Gray code $\mathbf{G}_{2,h}(i, *)$. By Fact 2.2, the total number of runs of 1s in each $\mathbf{B}(i, *)$ equals

$$\sum_{i=1}^{h-1} 2^{i-1} + 2 = 2^{h-1} + 1.$$

Finally, we set $\eta_s = 2^{h-1} + 2$. Then, $\frac{\ell}{h_\mathbf{B}} = \frac{h_\mathbf{B} 2^h + 1}{h_\mathbf{B}} > 2^h = 2\eta_s - 4$ with the minor restriction that $h_\mathbf{B} > 2(h+1)$.

B. The Saturation SQGT Model for $B(n, \le \ell, \underline{\eta})$

For the bounded-length burst problem $B(n, \le \ell, \underline{\eta})$, one needs to recover both $h_{\boldsymbol{b}}$ and $t_{\boldsymbol{b}}$ in order recover the burst. We describe next an order-optimal scheme for $B(n, \le \ell, \underline{\eta})$ restricted to the saturation SQGT model. First, for $\eta_s = s \ge l$, we describe an optimal scheme termed the *Integer code*. Then, for $\eta_s = s < l$, by adapting the bursty GT scheme from [1], we arrive at an order-optimal scheme (within a constant factor of 4) in Theorem 8 with the proof left in full version.

Theorem 7: For $B(n, \le \ell, \underline{\eta})$ and the saturation SQGT model for which $\underline{\eta} = (0, \dots, s)$ and $s \ge \ell$, there exists and order-optimal scheme $\mathbf{N}$ that solves $B(n, \le \ell, \underline{\eta})$ using $\lceil \log_2(n) \rceil + 1$ measurements.

Proof: The matrix $\mathbf{N}$ is a vertical concatenation of a $\lceil \log_2(n) \rceil \times n$ Index matrix and an $\mathbf{1}^{1\times n}$. Note that the ith column of the Index matrix is the binary representation of i.

Example 4.2: For n=8, we have

$$\mathbf{N} = \begin{bmatrix}
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}.$$

Since $\ell \le s$, $\underline{\eta}\left(\mathbf{N}\boldsymbol{b}\right) = \mathbf{N}\boldsymbol{b}$. We then treat the outcome vector as a binary representation of an integer k equal to

$$\begin{aligned}k :&= \sum_{i=0}^{\lceil \log_2(n) \rceil} 2^i \left(\mathbf{N}\boldsymbol{b}\right)(i, 0) = \sum_{i=0}^{\lceil \log_2(n) \rceil} 2^i \sum_{j=h_{\boldsymbol{b}}}^{t_{\boldsymbol{b}}} \mathbf{N}(i, j) \\ &= \sum_{j=h_{\boldsymbol{b}}}^{t_{\boldsymbol{b}}} \sum_{i=0}^{\lceil \log_2(n) \rceil} 2^i \mathbf{N}(i, j) = \sum_{j=h_{\boldsymbol{b}}}^{t_{\boldsymbol{b}}} j = \frac{h_{\boldsymbol{b}} + t_{\boldsymbol{b}}}{2} \ell_{\boldsymbol{b}}.\end{aligned}$$

We can deduce $\ell_{\boldsymbol{b}}$ from the outcome corresponding to $\mathbf{1}^{1\times n}$. Hence, $\mathbf{N}$ recover the burst $\boldsymbol{b}$ using $\lceil \log_2(n) \rceil + 1$ measurements. ∎

Theorem 8: For $B(n, \le \ell, \underline{\eta})$ under the saturation SQGT model with $\underline{\eta} = (0, \dots, s)$ and $s < \ell$, there exists an order-optimal scheme (within a constant factor 4) $\mathbf{C}$ that solves $B(n, \le \ell, \underline{\eta})$ using $\le \frac{2\ell}{s} + 2\log_2(n) + 3$ measurements. When $s = 1$, this recovers the original bound for GT from [1].

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) under grants CCF 2107344 and CCF 2046991.

REFERENCES

[1] C. J. Colbourn, "Group testing for consecutive positives," *Annals of Combinatorics*, vol. 3, no. 1, pp. 37–41, 1999.

[2] A. Emad and O. Milenkovic, "Semiquantitative group testing," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4614–4636, 2014.

[3] R. Dorfman, "The detection of defective members of large populations," *The Annals of mathematical statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[4] M. Aldridge, O. Johnson, J. Scarlett *et al.*, "Group testing: an information theory perspective," *Foundations and Trends® in Communications and Information Theory*, vol. 15, no. 3-4, pp. 196–392, 2019.

[5] D. Du, F. K. Hwang, and F. Hwang, *Combinatorial group testing and its applications*. World Scientific, 2000, vol. 12.

[6] Y.-L. Lin, C. Ward, and S. Skiena, "Synthetic sequence design for signal location search," *Algorithmica*, vol. 67, pp. 368–383, 2013.

[7] M. Müller and M. Jimbo, "Consecutive positive detectable matrices and group testing for consecutive positives," *Discrete mathematics*, vol. 279, no. 1-3, pp. 369–381, 2004.

[8] T. V. Bui, M. Cheraghchi, and T. D. Nguyen, "Improved algorithms for non-adaptive group testing with consecutive positives," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1961–1966.

[9] J. S.-T. Juan and G. J. Chang, "Adaptive group testing for consecutive positives," *Discrete mathematics*, vol. 308, no. 7, pp. 1124–1129, 2008.

[10] T. V. Bui, Y. M. Chee, J. Scarlett *et al.*, "Group testing with blocks of positives," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1082–1087.

[11] A. Emad and O. Milenkovic, "Code construction and decoding algorithms for semi-quantitative group testing with nonuniform thresholds," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1674–1687, 2016.

[12] Y.-J. Lin, C.-H. Yu, T.-H. Liu, C.-S. Chang, and W.-T. Chen, "Positively correlated samples save pooled testing costs," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2170–2182, 2021.

[13] R. Gabrys, S. Pattabiraman, V. Rana, J. Ribeiro, M. Cheraghchi, V. Guruswami, and O. Milenkovic, "Ac-dc: Amplification curve diagnostics for covid-19 group testing," *arXiv preprint arXiv:2011.05223*, 2020.

[14] M. Cheraghchi, R. Gabrys, and O. Milenkovic, "Semiquantitative group testing in at most two rounds," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1973–1978.

[15] A. G. Dyachkov and V. V. Rykov, "A survey of superimposed code theory," *Problems of Control and Information Theory*, vol. 12, no. 4, pp. 1–13, 1983.

[16] C. Savage, "A survey of combinatorial gray codes," *SIAM review*, vol. 39, no. 4, pp. 605–629, 1997.