arXiv:2304.01365v1 [cs.IT] 3 Apr 2023

通过非自适应半定量群体检测发现突发阳性样本

Yun-Han Li[a]、Ryan Gabrys[b]、Jin Sima[a]、Ilan Shomorony[a] 和 Olgica Milenkovic[a]a.电气与计算机工程系，伊利诺伊大学厄巴纳-香槟分校，美国 电子邮箱：{yunhanl2,jsima,ilans,milenkov}@illinois.eedub.美国海军信息战中心，圣迭戈 电子邮箱：ryan.gabrys@gmail.com

摘要——受病原体引起的疾病检测所启发，我们研究一种新的非自适应群组测试问题，其中：（1）阳性个体以突发形式出现，这反映了受感染的检测对象往往呈簇状分布；（2）检测结果来自半定量测量，能够粗略提供任一检测组内阳性个体数量的信息。我们的模型推广了先前关于利用经典群体检测检测单个突发阳性样本 **[1]** 以及半定量群组测试（SQGT） **[2]**的研究。具体而言，我们研究已知突发长度 ℓ ，且半定量检测对检测组内阳性个体数量给出可能不均匀估计的情形。这些估计表示包含阳性个体总数（精确值）的量化区间的索引，阈值可任意设定为 η_1、.....、 η_s。有趣的是，我们证明，突发识别所需的最少检测次数本质上仅取决于最大阈值 η_s。在此背景下，我们的主要结果是一种阶数量级最优检测方案，可用大约 $\frac{\ell}{2\eta_s}+\log_{s+1}(n)$次测量恢复任意长度为$\ell$ 的突发。这表明，在处理突发时，较高的饱和水平 η_s 比精细量化的信息更为重要。我们还针对相关建模假设和特殊的阈值选择给出了结果。

I. 引言

群体检测（GT）是一种用于从较大的实体集合（称为检测对象）中识别相对较小的标记元素子集（称为阳性个体）的方案。其核心思路是将检测对象分组到精心设计的子池中，并对每个子池中的对象进行联合检测，从而减少检测次数，相较于单独检测所需的次数更少。Dorfman [3] 最早提出了由两轮检测组成的群体检测方案，用于查找患有性传播疾病的个体。他的方案也首次采用了自适应检测，即利用某一轮检测的测量结果来指导后续轮次的检测选择。与自适应检测不同，非自适应群体检测要求所有检测同时设计并实施。自 Dorfman 的研究以来，群组测试在许多不同方向上得到了扩展和推广，并已广泛应用于搜索系统、实验与电路设计以及计算生物学。感兴趣的读者可参阅 [4], [5]，了解全面综述。

在 [1], 中，Colbourn 研究了一种专门的群体检测（GT）技术，用于在包含 n 个元素的有序列表中识别一个长度为 $\leq \ell$ 的连续阳性样本突发。对于非自适应技术，Colbourn 证明，达到最优阶所需的测量次数为 $\ell+\log(n)$。后续研究致力于改进该方案的某些方面 [6]–[8], ，并扩展其研究结果，以纳入新的自适应协议 [9], ，并将该方法推广至处理多个突发 [10]。

然而，在许多现实场景中，例如通过定量聚合酶链式反应（定量聚合酶链式反应，qPCR）检测病毒病原体感染，检测结果是实值，通常限定在某个区间内，例如 $[10,45]$。测量值称为 C_t （循环阈值）值，它能传达个体感染的可能性信息。例如，接近 40 的 C_t 值强烈表明受检者为阴性，而低于 20 的值则强烈表明该个体具有很强的传染性。因此，可以使用经过精心选择的一组 s 阈值$\underline{\eta}$ = （η_1, ..., η_s）对 C_t 值进行量化，使每个量化区间都能估计样本池中的病毒载量，进而估计样本池中的阳性个体数量。这类群组检测方法称为半定量群组检测（SQGT） [2], [11]。此外，在对大规模人群进行检测时，如果共同居住的个体在检测顺序中自然相邻（例如家庭成员、宿舍室友等），则适合采用突发阳性模型；与经典群组检测方法相比，这种模型可以显著节省资源 [12] [1]。

鉴于额外的定量信息以及关于阳性个体连续排序的假设，我们很容易设想对突发型阳性个体进行 SQGT，将C_t 数值量化到不同的量化区间，以表示病毒载量水平，或估算人群中的感染者数量 [13], [14]。本文首次研究了在这种设置下可实现的群组测量数量削减。具体而言，我们探讨了两种新的突发型 SQGT 模型 [2], ：一种模型中，突发长度已知且固定为 ℓ （以下称为固定长度突发模型， $B(n,\ell,\underline{\eta})$）；另一种模型中，突发长度已知且上限为 ℓ （以下称为有界长度突发模型， $B(n,\leq\ell,\underline{\eta})$）。

我们的主要贡献包括
1）对于$B(n,\ell,\underline{\eta})$设置，给出阶最优构造（即构造性的下界和上界相差一个常数因子 2），其量化阈值为 $\underline{\eta}$，且满足$\ell=\Omega(\eta_s\log_2(\eta_s))$. 2）对于 $B(n,\leq\ell,\underline{\eta})$ 设置，给出阶最优构造（即下界和上界相差一个常数因子 4），其 SQGT 阈值为 $\underline{\eta}=(1,\ldots,s)$，对应于所谓的饱和模型 [13], [15]。

这里有两点重要说明。半定量测量可显著减少所需的测试次数，以用于

$B(n,\le\ell,\underline{\eta})$设置（改进幅度与阈值数量 s呈线性关系）。有些出人意料的是，对于 $B(n,\ell,\underline{\eta})$设置，测试次数基本由最大阈值 η_s 决定，而非由阈值总数 s决定。这些发现可能会对实际定量聚合酶链式反应方案中采用的测试计划和量化方案产生有趣的影响。

本文组织如下。第二节介绍符号并给出正式的问题表述。第三节给出下界结果，第四节则包含本文的主要结果，即在突发长度固定且有上界的情况下，SQGT 突发识别模型数量的上界。

II. 问题表述

我们首先介绍相关符号，以及基于 SQGT 测量的定长和有界长度单突发识别问题。

令 $h_{\mathbf{M}}$、$w_{\mathbf{M}}$、$\mathbf{M}(*,i)$、$\mathbf{M}(j,*)$分别表示矩阵 $\mathbf{M}^{h_{\mathbf{M}}\times w_{\mathbf{M}}}$ 的行数（高度）、列数（宽度）、第 i 列和第 j 行。行索引取自 $[0,h_{\mathbf{M}}-1]$，而列索引限定在 $[0,h_{\mathbf{M}}-1]$ 内。此外，$\mathcal{R}(\mathbf{M})$、$\mathbf{M}^c$、$\mathbf{M}^\infty$ 分别表示由 M 通过反转列顺序得到的矩阵（因此$\mathcal{R}(\mathbf{M})(*,i)=\mathbf{M}(*,w_{\mathbf{M}}-i-1)$）、M 的 c 重水平拼接（即 $[\mathbf{M}\ldots,\mathbf{M}]$，其中包含 c 个组成矩阵），以及矩阵 M 的水平拼接，其中 $w_{\mathbf{M}\infty}$ 在构造过程中取指定值。最后，我们用 $\mathbf{M}(i,j)$ 表示 M 中第 i 行、第 j 列的元素。

单个突发阳性样本问题需要引入以下概念。

突发段：一个突发段表示为二进制 $n\times 1$ 列向量$\boldsymbol{b}$，由起始位置和结束位置 $h_{\boldsymbol{b}}\le t_{\boldsymbol{b}}$ 定义，它们决定突发段的长度 $\ell_{\boldsymbol{b}}=t_{\boldsymbol{b}}-h_{\boldsymbol{b}}+1$。它由连续阳性样本组成：

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

当 $\ell_{\boldsymbol{b}}$ 固定时，$\boldsymbol{b}^i$ 表示起始位置为 $h_{\boldsymbol{b}}=i$ 的突发段，而两个突发段 $\boldsymbol{b}^i$、$\boldsymbol{b}^j$ 之间的距离定义为它们起始位置 $|i-j|$ 之差。

SQGT 测量：一次 SQGT 测量由一个 $1\times n$ 二进制向量 $\boldsymbol{m}$ 描述，使得

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

以及一组整数值量化阈值

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

使得 SQGT 测量结果 e qual

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

定义 2.1：当 $\underline{\eta}=(1,2,\ldots,s)$ 时，我们将这种专门的 SQGT 方案称为饱和 SQGT 模型。

正确的突发检测：对于任意隐藏突发 $\boldsymbol{b}$，检测算法生成的估计值 $\hat{}$（记为 $\boldsymbol{b}$）应等于$\boldsymbol{b}$。

定义 2.2：$B(n,\ell,\underline{\eta})$ 和 $B(n,\le\ell,\underline{\eta})$ 分别表示突发长度为 $=\ell$ 和 $\le\ell$、具有 n 个测试元素和 SQGT量化阈值 $\underline{\eta}$ 的定长与有界长度突发问题。

具有 m 次测量、针对 n 个元素的非自适应 SQGT 测试方案，可用一个 $m\times n$ 二元测量矩阵 M 表示，其中每一行对应一次 SQGT 测量。当且仅当满足以下条件时，我们称 M 解决了 $B(n,\ell,\underline{\eta})$（或 $B(n,\le\ell,\underline{\eta})$）问题：

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

在所有非自适应 SQGT 方案中，满足此要求所需的最少测量次数记为 $m^*_{B(n,\ell,\underline{\eta})}$ 和 $m^*_{B(n,\le\ell,\underline{\eta})}$。

我们的构造将使用格雷码及其推广形式。若 $\mathbf{G}_{s,\ h}\in\{0,\ldots,s\}^{h\times s^h}$ 满足以下两个条件，则称其表示一个长度为 h 的 s 元格雷码：

1）任意两个相邻列恰好有一个位置不同，且差值的绝对值为一。

2）$\mathbf{G}_{s,\ h}$ 包含所有可能的 s^h 码字，且每个恰好出现一次。

例 2.1：下列矩阵的列构成了一个长度为二的三进制格雷码：

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

事实 2.1：格雷码 $\mathbf{G}_{s,\ h}$ 可通过以下步骤构造：先根据下述规则递归构造成对格雷码矩阵 $\mathbf{P}_{s,\ h}:=[\mathbf{G}_{s,\ h},\mathcal{R}(\mathbf{G}_{s,\ h})]$，然后从右侧移除一半的列：

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

这里，$\otimes$ 表示克罗内克积，而 $\mathbf{1}^a$ 是一个元素均为 1 的行向量。

例 2.2：下列矩阵 $\mathbf{P}_{3,2}$ 根据式（1）递归构造而成。正如所述，其左半部分等于 $\mathbf{G}_{3,2}$，并已在例 2.1 中展示：

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

我们还利用二进制格雷码的以下性质。

事实 2.2：$\mathbf{G}_{2,h}((i,*))$ 对所有 i（除 $i=0$ 外）都包含 2^{i-1} 段连续的1；其中 $i=0$ 只包含一段连续的1。因此，该矩阵的各行中共包含 $\sum_{i=1}^{h-1}2^{i-1}+1=2^{h-1}$ 段连续的1。下面的例子展示了 $\mathbf{G}_{2,3}$，中共有 $2^{3-1}=4$ 段连续的1。

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

III. 下界

我们首先给出 $m^*_{B(n,\ell,\eta)}$ 和 $m^*_{B(n,\le\ell,\underline{\eta})}$ 两种情形下所需的最少测量次数的下界。证明主要借鉴了 [1] 中的思路。

定理 1：我们有

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

用于 $m^*_{B(n,\le\ell,\underline{\eta})}$ 的证明方法与用于 $m^*_{B(n,\ell,\underline{\eta})}$ 的方法相似；因此，我们仅给出 $m^*_{B(n,\ell,\underline{\eta})}$ 的证明。我们通过分别证明右侧的各个界，并用最大值将它们合并，来证明第一个界。

1) 界 $\log_{s+1}(n-\ell+1)$ 可由简单的计数论证得出：头部位置总数为 $n-\ell+1$，每次测量有 $s+1$ 种可能结果。

2) 界 $\lceil \frac{\ell}{2\eta_s} \rceil$：我们证明，即使只要求区分前 $\ell+1$ 个突发（即满足 $0 \le i \le \ell$ 的突发 $\boldsymbol{b}^i$），仍需要 $\lceil \frac{\ell}{2\eta_s} \rceil$ 次测量。对于任意测量 $\boldsymbol{m}$，令 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 分别表示 $\boldsymbol{m}$ 中由 ℓ 比特组成的第一块和第二块。只有 $\boldsymbol{m}_1$ 中最后的 η_s 个非零比特以及 $\boldsymbol{m}_2$ 中最前面的 η_s 个非零比特是相关的。为简单起见，我们仅给出 $\boldsymbol{m}_2$ 情形的证明。令 $\ell \le i_1 < i_2 < \ldots < 2\ell$ 为包含在 $\boldsymbol{m}_2$ 中的元素。由于 $\ell_{\boldsymbol{b}} = \ell$ 且 $h_{\boldsymbol{b}} \le \ell$，如果 i_j 被包含在突发 $\boldsymbol{b}$ 中，那么 i_1、……、i_{j-1} 也必须包含在其中。因此，如果 $j > \eta_s$，注意到 η_s 是最大的阈值，就可以从 $\boldsymbol{m}_2$ 中移除 i_j，而不改变结果。由此，只保留 $\boldsymbol{m}_2$ 中最前面的 η_s 个非零比特，也能得到相同结果。因此，只需考虑满足 $\sum_{j=0}^{2\ell-1} \boldsymbol{m}(j) \le 2\eta_s$ 的 $\boldsymbol{m}$。令 $\mathrm{M}^{h_{\mathrm{M}} \times 2\ell}$ 为测量矩阵在前 2ℓ 列上的限制。假设 $h_{\mathrm{M}} < \frac{\ell}{2\eta_s}$；则 $\sum_{i=0}^{h_{\mathrm{M}}-1} \sum_{j=0}^{2\ell-1} \mathrm{M}(i,j) \le 2\eta_s h_{\mathrm{M}} < \ell$。这意味着存在某个 $0 \le j < \ell$，使得 $\mathrm{M}(*,j) = \mathrm{M}(*,j+\ell) = \mathbf{0}^{h_{\mathrm{M}} \times 1}$。于是 $\underline{\eta}\left(\mathrm{M}\boldsymbol{b}^{i+1}\right) = \underline{\eta}\left(\mathrm{M}\boldsymbol{b}^i + \mathrm{M}(*,j+\ell) - \mathrm{M}(*,j)\right) = \underline{\eta}\left(\mathrm{M}\boldsymbol{b}^i\right)$。因此 $m^*_{B(n,\ell,\underline{\eta})} \ge \lceil \frac{\ell}{2\eta_s} \rceil$。

IV. 主要结果

在第 IV-A 节中，我们针对两种不同情形，描述了 $B(n,\ell,\underline{\eta})$ 问题的一种阶最优测量矩阵构造：一是具有 $\ell = \Omega(\eta_s \log(\eta_s))$ 的一般 SQGT 阈值情形，二是具有 $\ell \le \eta_s = s$ 的饱和模型。值得注意的是，对于第一种情形，$m^*_{B(n,\ell,\underline{\eta})}$ 基本上只取决于最大阈值 η_s。换言之，只要 $\ell = \Omega(\eta_s \log_2(n))$ 的常数足够大，与采用单个最大阈值 η_s 的阈值群组测试 (TGT) 相比，使用多个阈值 (SQGT) 并无益处。在第 IV-B 节中，我们描述了针对 $B(n,\le\ell,\underline{\eta})$ 问题和饱和模型的一种阶最优方案（近似常数为 4）。

A. $B(n,\ell,\underline{\eta})$ 模型

由于这种情况下突发长度固定，因此只需确定突发段起始位置 $h_{\boldsymbol{b}} \in [0, n-\ell]$ $\boldsymbol{b}$。粗略地说，近乎最优的构造遵循一种两步草图与细化过程。第一部分称为通用草图方案，使用测量矩阵 K（定理 2）来区分相隔 $> \ell+1$ 个位置的突发。第二部分称为一般细化方案，使用测量矩阵 R（定理 3）来区分相隔 $< 2\ell$ 个位置的突发。将这两个测量矩阵叠放即可得到定理 4 中的结果。

定理 2：对于 $B(n,\ell,\underline{\eta})$，第 IV-A1 节所述的测量矩阵 K 可区分距离为 $> \ell+1$ 的所有突发，仅需 $\lceil \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) \rceil$ 次测量。

定理 3：对于参数为 $\eta_s = 2^{h-1}+2$ and $\ell = c2^h+1$ 的 $B(n,\ell,\underline{\eta})$，其中 $c,h \in \mathbb{N}$ and $c > 2(h+1)$，第 IV-A2 节所述的测量矩阵 R 能以约 $\frac{\ell}{2\eta_s}$ 次测量区分距离为 $< 2\ell$ 的所有突发。该方案仅取决于最大阈值 η_s。

定理 4：结合定理 2 的通用草图矩阵和定理 3 的通用细化矩阵，可得测量矩阵 $[\mathrm{K}^\intercal, \mathrm{R}^\intercal]^\intercal$，它可用于求解 $B(n,\ell,\underline{\eta})$，所需测量次数约为 $\frac{\ell}{2\eta_s} + \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)$。该测量次数至多是定理 1 所述下界对应测量次数的两倍。

备注 4.1：注意，定理 3 中的方案只使用最大的阈值 η_s。因此，如果我们在定理 2 的通用草图方案中只使用 η_s，则得到的测量矩阵高度约为 $\frac{\ell}{2\eta_s} + \log_2\left(\frac{n-\ell+1}{\ell}\right)$，且只依赖一个阈值 η_s。当 $\ell = \Omega(\eta_s \log_2(n))$ 且常数足够大时，$\frac{\ell}{2\eta_s} + \log_{s+1}\left(\frac{n-\ell+1}{\ell}\right) = \Omega\left(\log_2(n)\right) = \frac{\ell}{2\eta_s} + \log_2\left(\frac{n-\ell+1}{\ell}\right)$。因此，在这一参数范围内，使用多个阈值并无益处。

1）定理 2 的证明：我们先介绍一些相关记号。设 $\mathrm{K}^{h_{\mathbf{K}} \times n}$ 为测量矩阵。如果 $\mathrm{O} = \left[\underline{\eta}\left(\mathrm{K}\boldsymbol{b}^0\right) \ldots \underline{\eta}\left(\mathrm{K}\boldsymbol{b}^{n-\ell+1}\right)\right]$ 表示使用测量矩阵 K 对所有长度为 ℓ 的突发 $\boldsymbol{b}^i$ 得到的结果集合，则称 $\mathrm{K}^{h_{\mathbf{K}} \times n}$ 产生结果矩阵 $\mathrm{O}^{h_{\mathbf{K}} \times n-\ell+1}$。

接下来，令 $\vec{B}_\ell(x) := \mathbf{0}^{\ell-x}\mathbf{1}^x$ 和 $B_\ell(x) := \mathbf{1}^x\mathbf{0}^{\ell-x}$，对所有 $x \in \{0,\ldots,\ell\}$ 均成立。另令 $\vec{B}_\ell(\mathbf{0})^i, B_\ell(\mathbf{0})^i$ 表示 i 个 $\vec{B}_\ell(x)$ 和 $B_\ell(x)$ 的水平拼接。注意，$\lceil \log_{s+1}\left(\frac{n-\ell+1}{l}\right) \rceil$（几乎）达到计数界 $\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)$。关键思路是先构造 K，使其满足 $w_{\mathbf{K}} \ge n$，从而结果矩阵满足

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

然后将其截断为 n 列。根据 O 的定义，当且仅当 $\mathrm{G}_{s+1,\ h_{\mathbf{K}}}$ 的所有列各不相同时，K 才能识别距离为 $\ge \ell+1$ 的所有突发；这一点成立，是因为格雷码恰好包含每个向量 $\{0,\ldots,s\}^{h_{\mathbf{K}}}$ 一次。后续推导需要以下引理。

引理 4.1：以下测量

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

会得到结果 $\left([0,\ldots,s,s,\ldots,0] \otimes \mathbf{1}^{c\ell+1}\right)^\infty$。

证明：$c=1$ 这种情形很容易证明，以下例子对此作了说明。对于 $\underline{\eta} = (1,2,4)$ 和 $\ell = 6$, $\boldsymbol{m}(1)$

等于

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

对于 $c>1$，以及任意长度为 ℓ 的行向量 $\boldsymbol{x}$，对于所有 $\boldsymbol{b}^i$，$\underline{\eta}\left(\boldsymbol{x}^c\boldsymbol{b}^i\right)$ 均保持不变，其中 $i\in[0,(c-1)\ell]$。

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

■

现在我们可以给出我们的构造了。令 $\mathbf{M}[i]$ 为按以下方式递归构造的测量矩阵：

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

注意，$\frac{(\ell+1)(s+1)^{i-1}-1}{\ell}$可能不是整数。因此，我们先关注特殊情形 $s=\ell$（因此 $\frac{(\ell+1)(s+1)^{i-1}-1}{\ell}=\frac{(\ell+1)^i-1}{\ell}\in\mathbb{N}$），然后通过对论证稍作修改，将结果推广到 $s<\ell$。

引理4.2：对于 $s=\ell$，$\mathbf{M}[i]^\infty$ 得到结果矩阵 $\left(\mathrm{P}_{\ell+1,\ i}\otimes\mathbf{1}^{\ell+1}\right)^\infty$。其中 $\mathrm{P}_{\ell+1,\ i}$ 是 $\ell+1$ 元、长度为 i 的配对格雷码矩阵。

证明：使用归纳法进行证明。

1) 对于 $i=1$：根据引理 4.1，$\mathbf{M}[i]^\infty=\boldsymbol{m}(1)^\infty$ 得到结果矩阵 $\left([0,\dots,\ell,\ell,\dots,0]\otimes\mathbf{1}^{\ell+1}\right)^\infty=\left(\mathbf{P}_{\ell+1,1}\otimes\mathbf{1}^{\ell+1}\right)^\infty$.

2) 对于 $i>1$：假设该结论对 $i-1$ 成立。那么

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

得到结果矩阵

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

■

根据引理4.2，$\mathbf{M}\ [h_{\mathbf{K}}]$ 截去右侧的 $(\ell+1)^{h_{\mathbf{K}}+1}+\ell-1$ 列后，得到结果矩阵 $\mathrm{G}_{\ell+1,h_{\mathbf{K}}}\otimes\mathbf{1}^{\ell+1}$，因此能够区分所有距离不超过 $>\ell+1$ 的突发。不难证明，单次测量 $\mathbf{0}^\ell\left(\mathbf{1}^{\ell+1}\mathbf{0}^{\ell+1}\right)^\infty$ 得到结果矩阵 $(0,\dots,\ell,\ell,\dots,0)^\infty$。因此，我们得到以下定理。

定理5：对于具有 ℓ 阈值$\underline{\eta}=(1,\dots,\ell)$的饱和 SQGT 模型，右侧截取至$n$ 列的 $\begin{bmatrix}\mathbf{M}\ [\lceil\log_{\ell+1}\ n-\ell+1\ -1\rceil]^\infty\\ \mathbf{0}^\ell\left(\mathbf{1}^{\ell+1}\mathbf{0}^{\ell+1}\right)^\infty\end{bmatrix}$()可用作具有 $\lceil\log_{\ell+1}(n-\ell+1)\rceil$ 次测量的$B(n,\ell,\underline{\eta})$模型的测试矩阵。

对于 $s<\ell$ 这种情况，需要对式 (3) 中给出的递归进行一些修改。修改方法是在递归的每个阶段，从 $\mathbf{M}'[i]$ 的左侧、右侧或两侧截去一定数量的列 i：

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

其中 $\alpha=\frac{w_{\mathbf{M}'[i-1]}}{2}-1$，$\mathbf{M}'_r[i-1]$、$\mathbf{M}'_l[i-1]$、$\mathbf{M}'_{lr}[i-1]$ 分别表示从右侧、左侧和两侧截去 $\alpha\bmod\ell$ 列后的 $\mathbf{M}'[i-1]$。

采用与上述证明类似的方法，并进行一些简单但繁琐的计算，可以证明从右侧截取至 n 列的 $\mathbf{M}'\left[\lceil\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)\rceil\right]$ 可用作 K。因此，K 可以通过 $\lceil\log_{s+1}\left(\frac{n-\ell+1}{\ell}\right)\rceil$ 次测量区分所有距离为$>\ell+1$ 的突发。

2) 定理 3 的证明：现在我们将注意力集中在一般细化方案上。令 B 表示 B 中各列向左循环移位一个位置后得到的矩阵，使得 $\mathrm{B}(*,i)=\mathrm{B}(*,i+1\bmod\ell)$。我们需要以下引理。

引理 4.3：假设二进制矩阵 $\mathrm{B}^{h_{\mathrm{B}}\times\ell}$ 满足以下三个条件：

1) 所有列及其二进制补 $\{\mathrm{B}(*,i)\ \bar{}\mathrm{B}(*,i)\}_{i=0}^{\ell-1}$ 均互不相同。

2) 第一列是零向量，$\mathrm{B}(*,0)=\mathbf{0}^{h_{\mathrm{B}}\times1}$。

3) B − B 的每一行都有 η_s-1 个元素等于 -1。

那么，以下测量矩阵可用于区分距离为 $<2\ell$ 的所有突发 $\boldsymbol{b}\neq\boldsymbol{b}'$：

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

其中，R^- 表示B −B的“负”部分（通过将1设为0、将 -1设为1得到），而R^+ 表示B−B的正部（通过将 -1设为0得到）；最后一列则由 $\mathbf{0}^{h_{\mathrm{B}}\times1}$ 变为 $\mathbf{1}^{h_{\mathrm{B}}\times1}$（注意，修改前B − B的最后一列为$\mathrm{B}(*,0)-\mathrm{B}(*,\ell-1)=-\mathrm{B}(*,\ell-1)$，这意味着正部为零）。

证明：由于R是R^- 和R^+的重复水平拼接，只需证明

$$\forall0\le i\neq j<2\ell,\ \underline{\eta}\left(\mathbf{R}\boldsymbol{b}^i\right)\neq\underline{\eta}\left(\mathbf{R}\boldsymbol{b}^j\right).\tag{5}$$

特别地，我们证明t at h

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

注意，$\mathrm{R}\boldsymbol{b}^i$ 的每个元素要么是 η_s-1，要么是 η_s。根据条件 1，所有 $\mathrm{R}\boldsymbol{b}^i$ 都不相同。因此，所有 $\underline{\eta}\left(\mathrm{R}\boldsymbol{b}^i\right)$ 也都不相同。所以，利用 $h_{\mathbf{B}}$ 次测量，R 可以区分距离为 $<2\ell$的所有突发；只有最大的阈值 η_s 是相关的。接下来我们证明 (6)。

For $0\le i<\ell$,

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

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

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

等式(1)= 由条件 2 得出。最后，根据条件 3，以及我们将 R^{+} 的最后一列从$\mathbf{0}^{h_{\mathbf{B}}\times 1}$ 改为 $\mathbf{1}^{h_{\mathbf{B}}\times 1}$这一事实，$\mathrm{R}\boldsymbol{b}^{0}=($ $\eta_s-1)\,\mathbf{1}^{h_{\mathbf{B}}\times 1}$ 和 $\mathrm{R}\boldsymbol{b}^{\ell}-\mathrm{R}\boldsymbol{b}^{0}=\mathbf{1}^{h_{\mathbf{B}}\times 1}\Rightarrow \mathrm{R}\boldsymbol{b}^{\ell}=\eta_s\mathbf{1}^{h_{\mathbf{B}}\times 1}$。 ■

接下来要构造一个矩阵 B，使其满足引理 4.3 中的条件 1–3，且 $h_{\mathbf{B}}$ 大致等于 $\frac{\ell}{2\eta_s}$。令 $\mathrm{G}_{2\ h}^{h\times 2^h}$，为长度为 h 的二进制格雷码的码矩阵，使得 $2(h+1)<h_{\mathbf{B}}$。我们构造 $\bar{}\mathrm{G}_{2\ h\ i}^{h_{\mathbf{B}}\times 2^h}$，如下：

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

然后，B 按如下方式构造：

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

示例 4.1：矩阵 B 使用 $\mathrm{G}_{2,2}$ 和 for$h_{\mathbf{B}}=7$构造如下：

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

- 条件 1：我们展示一个两步过程，根据每列的内容恢复其索引，从而证明所有列互不相同。思路是先使用 $\mathbf{1}^{1\times 2^h}$ 恢复 i （$\mathrm{G}_{2,h,i}$ 的索引），然后利用格雷码中接下来的 h 位定位确切的列。为此，在第一步中，我们需要增加一个约束 $h_{\mathbf{B}}\geq 2(h+1)$。简而言之，我们可以通过查看每列中长度为 $h+1$ 的零突发之后的第一个 1 来恢复 i。此外，由于这一约束，B 中每列的零都多于一。因此，所有 $\mathrm{B}(*,i)$ 和 $\bar{}\mathrm{B}(*,i)$ 都必须互不相同。

- 条件 2：此条件很容易验证。
- 条件 3：根据 (IV-A2) 中的构造，每一行 $\mathrm{B}(i)$ 都是

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

在左侧额外添加一个 0。在示例 4.1 中，我们有 $\boldsymbol{v}=[111100110110]$。按照此构造，B − B 每一行中 -1s 的数量等于 $\mathrm{B}(i,*)$ 中连续的1的段数，这又等于 $\mathbf{1}^{1\times 2^h}$ 中连续的1的段数与格雷码 $\mathrm{G}_{2,\ h}(i,*)$ 的每一行中连续的1的段数之和。根据事实 2.2，每个 $\mathrm{B}(i,*)$ 中连续的1的总段数等于

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

最后，我们设定 $\underline{\eta}s=2^{h-1}+2$ $\frac{\ell}{h_{\mathbf{B}}}=\frac{h_{\mathbf{B}}2^h+1}{h_{\mathbf{B}}}>2^h=$。然后，$2\underline{\eta}s-4$ $h_{\mathbf{B}}>2\ h+1$ 在满足 ( ) 这一轻微限制的条件下。

B. 适用于 $B(n,\leq \ell,\underline{\eta})$ 的饱和 SQGT 模型

对于有界长度突发问题 $B(n,\leq \ell,\underline{\eta})$，需要同时恢复 $h_{\boldsymbol{b}}$ 和 $t_{\boldsymbol{b}}$，才能确定该突发。接下来，我们介绍一种针对饱和 SQGT 模型下的 $B(n,\leq \ell,\underline{\eta})$ 的阶最优方案。首先，对于 $\eta_s=s\geq l$，我们介绍一种最优方案，称为整数码。然后，对于 $\eta_s=s<l$，通过改进 [1], 中的突发式 GT 方案，我们得到一种阶最优方案（误差不超过 4 倍常数因子）。

定理 6：对于 $B(n,\leq \ell,\underline{\eta})$ 及满足 $\underline{\eta}=(0,\ldots,s)$ 和 $s\geq \ell$ 的饱和 SQGT 模型，存在一个阶数最优方案 N，使用 $\lceil\log_2(n)\rceil+1$ 次测量即可求解 $B(n,\leq \ell,\underline{\eta})$。

证明：矩阵 N 由一个 $\lceil\log_2(n)\rceil\times n$ 索引矩阵和一个 $\mathbf{1}^{1\times n}$ 纵向拼接而成。注意，索引矩阵的第 i 列是 i 的二进制表示。

例 4.2：对于 n=8，我们有

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

由于 $\ell\leq s$, $\underline{\eta}(\mathrm{N}\boldsymbol{b})=\mathrm{N}\boldsymbol{b}$。随后，我们将结果向量视为整数 k 的二进制表示，其值为

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

我们可以根据对应于 $\mathbf{1}^{1\times n}$ 的结果推导出 $\ell_{\boldsymbol{b}}$。因此，N 使用 $\lceil\log_2(n)\rceil+1$ 次测量即可恢复突发 $\boldsymbol{b}$。 ■

定理 7：对于饱和 SQGT 模型下的 $B(n,\leq \ell,\underline{\eta})$，其中 $\underline{\eta}=(0,\ldots,s)$ 且 $s<\ell$，存在一个阶最优方案（在常数因子 4 以内）C，可用 $\leq\frac{2\ell}{s}+2\log_2(n)$ 次 $+3$ 测量求解 $B(n,\leq \ell,\underline{\eta})$。

证明：矩阵 C 如下：

$$\mathbf{C}=\left[\mathbf{C}_1^{\intercal}\quad \mathbf{C}_2^{\intercal}\quad \mathbf{C}_3^{\intercal}\right]^{\intercal},$$

其中，C_1 是定理 3.2 中的第 1 阶段矩阵，出自 [1]。根据 [1], 中所述的论证，C_1 可以区分距离在 $\geq 2(\ell-2)$ 以内的所有突发 $\boldsymbol{b}$。接下来，我们使用从右侧截取至 n 列的 $\mathrm{C}_2:=\left[\mathrm{I}_{\lceil\frac{2\ell}{s}\rceil}\otimes\mathbf{1}^{s}\right]^{\infty}$。与 C_2 对应的结果向量是一个（按循环意义）由非零元素构成的单段，其形式为 $o_h,s,\ldots,s,o_t$，其中 $o_h,o_t\in\{1,\ldots,s\}$。于是，

1) 对于 $\ell_{\boldsymbol{b}}>s$：由于 $\frac{\ell_{\boldsymbol{b}}}{s}>1$和 $\lceil\frac{2\ell}{s}\rceil\geq\lceil\frac{2\ell_{\boldsymbol{b}}}{s}\rceil\geq\lceil\frac{\ell_{\boldsymbol{b}}}{s}\rceil+1$，$h\neq t$。因此，可以使用 o_h，o_t 来恢复 $h_{\boldsymbol{b}}$ 模2ℓ，$t_{\boldsymbol{b}}$ 模 2ℓ。因此，C_2 能够区分所有满足 $\ell_{\boldsymbol{b}}>s$ 且距离为 $<2\ell$ 的突发。
2) 对于 $\ell_{\boldsymbol{b}}\leq s$：根据 C_2, 的测量结果，可以恢复 $\ell_{\boldsymbol{b}}$。如果 $\ell_{\boldsymbol{b}}\leq s$，那么根据定理 6，$\mathrm{C}_3:=\mathrm{N}$ 能够区分所有满足 $\ell_{\boldsymbol{b}}\leq s$的突发。

因此，使用 $h_{\mathbf{C}_1}+h_{\mathbf{C}_2}+h_{\mathbf{C}_3}\leq\log_2(n)+\lceil\frac{2\ell}{s}\rceil+\lceil\log_2(n)\rceil+1\leq\frac{2\ell}{s}+2\log_2(n)+3$ 次测量，C 可用于求解 $B(n,\leq \ell,\eta)$；测量次数至多为定理 1 所给下界的 4 倍。 ■

参考文献

[1] C. J. Colbourn，“连续阳性样本的群组测试”，Annals of Combinatorics，第3卷，第1期，第37–41页，1999年。[2] A. Emad 和 O. Milenkovic，“半定量群组测试”，IEEE Transactions on Information Theory，第60卷，第8期，第4614–4636页，2014年。

[3] R. Dorfman，“大规模人群中缺陷成员的检测”，The Annals of mathematical statistics，第14卷，第4期，第436–440页，1943年。[4] M. Aldridge、O. Johnson、J. Scarlett 等，“群组测试：信息论视角”，Foundations and Trends® in Communications and Information Theory，第15卷，第3–4期，第196–392页，2019年。[5] D. Du、F. K. Hwang 和 F. Hwang，《组合群组测试及其应用》。World Scientific，2000年，第12卷。[6] Y.-L. Lin、C. Ward 和 S. Skiena，“用于信号定位搜索的合成序列设计”，Algorithmica，第67卷，第368–383页，2013年。[7] M. Müller 和 M. Ji mbo，“连续阳性可检测矩阵与连续阳性样本的群组测试”，Discrete mathematics，第279卷，第1–3期，第369–381页，2004年。[8] T. V. Bui、M. Cheraghchi 和 T. D. Nguyen，“连续阳性样本的非自适应群组测试的改进算法”，载于2021年IEEE信息论国际研讨会（ISIT）。IEEE，2021年，第1961–1966页。[9] J. S.-T. Juan 和 G. J. Chang，“连续阳性样本的自适应群组测试”，Discrete mathematics，第308卷，第7期，第1124–1129页，2008年。[10] T. V. Bui、Y. M. Chee、J. Scarlett 等，“含阳性块的群组测试”，载于2022年IEEE信息论国际研讨会（ISIT）。IEEE，2022年，第1082–1087页。[11] A. Emad 和 O. Milenkovic，“具有非均匀阈值的半定量群组测试的码构造与译码算法”，IEEE Transactions on Information Theory，第62卷，第4期，第1674–1687页，2016年。[12] Y.-J. Lin、C.-H. Yu、T.-H. Liu、C.-S. Chang 和 W.-T. Chen，“正相关样本降低混样检测成本”，IEEE Transactions on Network Science and Engineering，第8卷，第3期，第2170–2182页，2021年。[13] R. Gabrys、S. Pattabiraman、V. Rana、J. Ribeiro、M. Cheraghchi、V. Gu- ruswami 和 O. Milenkovic，“Ac-dc：新冠肺炎群组测试的扩增曲线诊断”，arXiv预印本 arXiv:2011.05223，2020年。[14] M. Cheraghchi、R. Gabrys 和 O. Milenkovic，“最多两轮的半定量群组测试”，载于2021年IEEE信息论国际研讨会（ISIT）。IEEE，2021年，第1973–1978页。[15] A. G. Dyachkov 和 V. V. Rykov，“叠加码理论综述”，Problems of Control and Information Theory，第12卷，第4期，第1–13页，1983年。