NEW GROUP TESTING PARADIGMS:  
FROM PRACTICE TO THEORY

BY

AMIN EMAD

DISSERTATION
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:
Associate Professor Olgica Milenkovic, Chair
Professor Yoram Bresler
Professor Pierre Moulin
Associate Professor Angelia Nedich
Assistant Professor Maxim Raginsky
Assistant Professor Yihong Wu
We propose a novel group testing framework, termed semi-quantitative group testing, motivated by a class of problems arising in genome screening experiments in addition to other applications such as interpretable rule learning for decision making. Semi-quantitative group testing (SQGT) is a (possibly) non-binary pooling scheme that may be viewed as a concatenation of an adder channel and an integer-valued quantizer. In its full generality, SQGT may be viewed as a unifying framework for group testing, in the sense that most group testing models are special instances of SQGT. For the new testing scheme, we define the notion of SQ-disjunct and SQ-separable test matrices, representing generalizations of classical disjunct and separable matrices. We describe combinatorial and probabilistic constructions for such matrices without considering any restriction on the thresholds of the SQGT model (i.e. SQGT with arbitrary thresholds). Then, we focus on the important special case in which the thresholds are equidistant, and construct SQ-disjunct and SQ-separable matrices for this model. While for most of the constructions described in this dissertation, it is assumed that the number of defectives is much smaller than total number of test subjects, we also consider the case in which there is no restriction on the number of defectives and they may be as large as the total number of subjects. For the constructed matrices, we describe a number of efficient decoding algorithms based on algebraic methods and message passing on graphical models. Finally, we introduce the novel probabilistic group testing framework of Poisson group testing, applicable to dynamic testing with diminishing relative rates of defectives. For this new model, we consider both nonadaptive and adaptive testing schemes and develop lower bounds and tight constructive upper bounds on the number of required tests.
To my wife, for her love and support.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis advisor Prof. O. Milenkovic for her support throughout the course of my Ph.D. studies. Her guidance and help has been crucial to the completion of this dissertation.

I have also been very fortunate to collaborate with Prof. P. Moulin, Prof. V. Veeravalli, and Prof. J. Ma, and to benefit from their insightful comments and suggestions during many discussions. I would also like to thank my other collaborators Dr. E. Soljanin and Dr. C. Nuzman at Alcatel-Lucent Bell Labs, and Dr. K. Varshney and Dr. D. Malioutov at IBM Thomas J. Watson Research Center.

My sincere thanks also go to members of my doctoral committee: Prof. Y. Bresler, Prof. P. Moulin, Prof. A. Nedich, Prof. M. Raginsky, and Prof. Y. Wu who have provided much great advice and help through the years.

I would like to thank my parents who have always been a major source of support, inspiration, and encouragement to me during my whole life. Last but not least, I would like to express my gratitude to my wife, Julia, who has always been there for me during my course of Ph.D. studies, without whom I could not achieve this goal.
# TABLE OF CONTENTS

**CHAPTER 1  INTRODUCTION** ................................. 1  
1.1 Challenges in genotyping, and motivation for semi-quantitative group testing ................................. 4  
1.2 Challenges in dynamic testing, and motivation for Poisson group testing ................................. 7  

**CHAPTER 2  THE SQGT MODEL** ................................. 10  

**CHAPTER 3  GENERALIZED DISJUNCT AND SEPARABLE MATRICES FOR SQGT** ................................. 14  
3.1 SQ-disjunct matrices ................................. 14  
3.2 SQ-separable matrices ................................. 17  

**CHAPTER 4  TEST MATRIX CONSTRUCTIONS AND DECODING ALGORITHMS FOR SQGT WITH ARBITRARY THRESHOLDS** ................................. 19  
4.1 Construction of $q$-ary SQGT test matrices using binary disjunct/separable matrices for CGT ................................. 22  
4.2 SQ-separable matrices using quantized $B_h$ sequences ................................. 23  
4.3 SQ-separable matrices using SQLO$_s$ sequences ................................. 33  
4.4 SQ-separable matrices using SQLO$_t$ sequences ................................. 42  
4.5 Construction of binary SQ-separable matrices ................................. 46  

**CHAPTER 5  TEST MATRIX CONSTRUCTIONS AND DECODING ALGORITHMS FOR SQGT WITH EQUIDISTANT THRESHOLDS** ................................. 60  
5.1 Construction of $q$-ary SQ-disjunct matrices ................................. 62  
5.2 Construction of $q$-ary SQ-separable matrices ................................. 67  
5.3 Construction of SQ-separable matrices for arbitrary number of defectives ................................. 69  
5.4 Belief propagation decoders for SQGT ................................. 75  

**CHAPTER 6  POISSON GROUP TESTING FOR THE CGT MODEL** ................................. 85  
6.1 Introduction ................................. 85  
6.2 Problem setup ................................. 88
6.3 Nonadaptive methods for Poisson PGT . . . . . . . . . . . . . . . . . . . . . . . . 90
6.4 Semi-adaptive methods for Poisson PGT . . . . . . . . . . . . . . . . . . . . . . . 113
6.5 Summary of the results and discussion . . . . . . . . . . . . . . . . . . . . . . . . 119

CHAPTER 7 CONCLUSIONS AND FUTURE WORK . . . . . . . . 123

REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
CHAPTER 1

INTRODUCTION

Group testing (GT) is a general term for a family of test schemes designed to identify a number of subjects with some particular characteristic – called defectives (or positives) – among a large pool of subjects. The idea behind GT is that if the number of defectives is much smaller than the number of subjects, one can reduce the number of experiments required for identifying the defectives by testing properly chosen subgroups 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 the superposition of the state vectors of a subset of the system’s elements. As such, GT has found many applications in communication theory [1, 2, 3, 4], signal processing [5, 6, 7], computer science [8, 9, 10], and mathematics [11]. Some examples of these applications include error-correcting coding [3, 12, 13], identifying users accessing a multiple access channel (MAC) [14, 15], reconstructing sparse signals from low-dimensional projections [5, 6], and many others.

The group testing literature examines two partially overlapping categories of problems, based on the way the number of defectives is modeled: probabilistic GT and combinatorial GT. In the former case, a probability distribution is considered for the number of defectives, and the goal is to minimize the expected number of tests (see for example [16, 17, 18, 19]). In the latter case, the number of defectives (or at least an upper bound on the number of defectives) is known in advance [7].

Another way to distinguish between different GT schemes is through the way the tests are performed. In nonadaptive group testing all the tests are designed in advance. In other words, the tests are designed in one pass,
and the outcome of one test does not affect the design of another test. On the other hand, in sequential (adaptive) group testing, the result of one test may be used to govern the design of other tests, leading to more efficient pooling schemes (see [7] and references therein). Although, in general, sequential GT requires fewer tests, in most practical applications nonadaptive GT is preferred since it allows one to perform all tests simultaneously. This reduces the overall time required for testing. In what follows, we focus on combinatorial, nonadaptive GT.

Many different models have been considered for combinatorial GT. In the original setting described by Dorfman [16] (henceforth, conventional GT or CGT) the result of a test indicates if there exists at least one defective in the test; hence, the test output equals 0 if there are no defectives in the test, and 1 otherwise. Another important model is the additive model [7], also known as quantitative GT (QGT). In this model, the result of a test equals the exact number of defectives in that test. In the threshold group testing (TGT) model [20], if the number of defectives in a test is smaller than a fixed lower threshold, the test outcome is negative (or equal to 0); if the number of defectives is larger than a fixed upper threshold, the test outcome is positive (or equal to 1); and if the number of defectives is between the lower and upper threshold, the test result is arbitrary (either equal to 0 or 1). The difference between the upper and lower thresholds is called the gap. In yet another model introduced in [21], a threshold is fixed beforehand and the test output corresponds to an additive model output whenever the number of defectives does not exceed the threshold. If the number of defectives exceeds the threshold, the output of the test is some value outside the range of the sub-thresholded additive model output.

In all these models, each subject is assigned a unique binary vector (codeword) of length equal to the total number of tests. Each coordinate of a subject’s codeword corresponds to a test and equals 1 if the subject is present in the test, and equals 0 otherwise. Since in nonadaptive GT all the tests are designed in parallel, it is convenient to group all the codewords into a matrix (code) termed the test matrix (test code). The test matrix is a binary matrix of size $m \times n$, where $m$ is the number of tests and $n$ is the number of subjects. The design of efficient test matrices has been a topic of interest for many years: for a comprehensive survey of such matrices, see [7], [22], and [23]. The two main families of test matrices were originally designed for CGT.
by Kautz and Singleton [24]. The first family is known as *disjunct matrices* (or zero-false-drop matrices), while the second family is usually referred to as *separable matrices* (or uniquely decipherable matrices). Disjunct matrices satisfy an *inclusion* constraint: a $d$-disjunct matrix has the property that no column is included in – or is covered by – the component-wise Boolean ORs of any other $\leq d$ number of columns. This property enables disjunct matrices to uniquely identify up to $d$ defectives and also endows them with an efficient decoding algorithm. Separability is a weaker notion than disjunctness as it only requires the component-wise Boolean ORs of any two distinct sets of $\leq d$ columns to be different.

Despite the significant interest the subject has garnered in computer science, coding and combinatorial theory, and despite the analysis of many diverse extensions of the underlying problem, group testing has still not seen widespread use in medical sciences and biology. Two notable exceptions were the early use of group testing for DNA sequence analysis [22] and the very recent work on group testing for genotyping and biosensing [25, 26, 27]. The reason behind this practical failure of group testing in life sciences is that most analytical models do not capture the full complexity of bioengineering systems. Model simplifications are necessarily introduced in order to derive closed-form expressions on the smallest number of tests required to perform the experiments or to guarantee test matrix constructions with provable performance guarantees, thereby neglecting the fact that in practical applications such simplifications may not be appropriate. For example, one would be inclined to accept a number of tests higher than those predicted to be theoretically optimal for a coarse model if there is evidence that the scheme is suitable for practical implementation.

This work represents the first step in developing a novel framework for group testing that caters to the unique needs of the emerging field of genotyping through high-throughput sequencing.\(^3\)

\(^3\)Although this work was motivated by applications in genotyping, the model, results, and test matrix constructions are applicable to a wide variety of applications in biology, communication theory, signal processing, etc. One example of such applications is interpretable rule learning for decision making, which we will briefly discuss in the last chapter.
1.1 Challenges in genotyping, and motivation for semi-quantitative group testing

Genotyping is an emerging field in systems biology concerned with determining genetic variations in the traits of individuals. At the core of every genotyping method is DNA sequencing – determining the genetic blueprint of an individual – and a comparative analysis of the sequences obtained from different individuals. Comparative studies of the DNA makeup play an indispensable role in medical genetics, the goals of which are to efficiently determine “outliers” in the genetic code that may lead to devastating disorders or illnesses [25].

One of the most important applications of genotyping is detecting the carriers of a particular genetic disorder. Since the human genome consists of pairs of chromosomes, and paired chromosomes contain genes with matching functionalities, a human who has inherited a mutated gene may not display the symptoms of the genetic disease. In this situation, the individual has a normal (unmutated) copy of a gene, which prohibits the disease from being expressed. Although the carrier does not display disease symptoms, the offspring of two carriers may have the disease. While affected individuals can be diagnosed based on their symptoms, a carrier can only be identified via DNA screening.

In the screening process of genotyping, one targets genomic regions known to harbor genetic mutations. Until recently, only serial sequencing of the genome of one individual was possible; however, the introduction of the new class of genome sequencing methods dubbed the next-generation sequencing technologies [28] enabled parallel sequencing of the genome. These platforms break the genomic region of interest into short fragments and perform millions of sequence reads in a single run (for the description of one such platform, see Illumina [29]). Due to the high cost of sample preparation for sequencing, and, in order to fully utilize the potential of the sequencing platforms, multiplexing a large number of specimens in a single batch is essential. As a result, group testing presents itself as a natural paradigm to address these challenges, and the first steps in this direction were taken in [30, 31, 25, 26]. Despite the promising results of applying the existing group testing models to genotyping, many practical problems still stand in the way of the wide-scale use of this method.
One such problem arises from the fact that genotyping methods allow for more precise readings at the output than classical GT detectors, but still do not provide full information about the abundance of a target gene in the test. As a result, test matrices constructed for CGT or TGT underutilize the potential of these sequencers, while test matrices constructed for QGT are prone to errors due to “overestimating” the sequencers’ precision. Specifically, since the precision of a sequencer often depends on the number of defectives and the amount of genetic material in the test, the error is signal/design dependent and cannot be modeled easily. In order to overcome this problem, in what follows we propose a new framework called semi-quantitative group testing (SQGT).

In SQGT, the result of a test is a non-binary value that depends on the number of defectives through a given set of thresholds. The thresholds depend on the sequencer and represent its precision. The SQGT paradigm may be viewed as a combination of the adder model (QGT) and a decimator (quantizer). Although QGT has been widely studied in literature, the addition of a system-dependent decimator makes test construction and analysis quite challenging. It is worth emphasizing that the application of the SQGT model is not limited to genotyping, and in general any scheme in which tests are obtained using a test device with limited precision may be modeled as an instance of SQGT. In particular, CGT, TGT (with zero gap), and QGT are all special cases of SQGT.

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 the classic coin weighing problem – in other 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. While in binary test matrices a value 0 or 1 corresponds to the absence or presence of a subject in a test, respectively, in non-binary SQGT the value of an entry of the test matrix reflects the “strength” or “concentration” of a subject in a test. For example in conflict resolution in MAC, different non-binary values in a test correspond to different power levels of the users, while in genotyping they correspond to different amounts of genetic material of different subjects. For example, if the value corresponding to the $j^{th}$ subject in a genotyping test equals 2, while the value corresponding to the $k^{th}$ subject is equal to 1, this indicates
that the amount of DNA of subject \( j \) in this test is twice the amount of the DNA of subject \( k \).

The reason for focusing on integer-valued test matrices, as opposed to real-valued matrices, is that the sample preparation in genotyping is performed by robotic arms that are usually programmed to sample the same amount of DNA. One could program the robotic arm to dispense different amounts of DNA into test wells, but such a process would be extremely complicated and imprecise. A better alternative is to program the robotic arm to dispense the same amount of DNA into a test well multiple times. Since all test wells contain integer multiples of the same volume of DNA, one can model the test parameters using bounded integers.

Note that non-binary integer-valued group testing can be also used in applications where:

- The subjects to be tested come as a whole and cannot be divided into \textit{real-valued} parts. For example, in the coin-weighing problem, if one has \( n \) bags of coins, where each bag contains \( q - 1 \) identical coins, and some of the bags have counterfeit coins, one can use tests with an alphabet of size \( q \) to find the counterfeit bag with fewer experiments than when using binary tests.

- A real-valued alphabet may not be practical due to “limited precision”. With unlimited precision, one could design \textit{one} single experiment to find any number of defectives among any number of subjects.

- Some robustness to errors and noise is needed in the testing schemes; integers, unlike reals, are spaced discretely, which ensures a form of error protection (see for example [32]).

While there exist information theoretic approaches applicable to the study of non-binary test matrices [23, Ch. 6], the results on non-binary test matrix construction relevant to group testing are limited to a handful of papers, including [33] and [34], where constructions are considered for an \textit{adder MAC} channel (i.e. QGT).

For the new model of SQGT with \( Q \)-ary test results and \( q \)-ary test sample sizes, \( Q, q \geq 2 \), we define a new generalization of \textit{disjunct} and \textit{separable} matrices, called “SQ-disjunct” and “SQ-separable”, respectively. Probabilistic constructions as well as explicit constructions are provided for these two
families of test matrices when the number of defectives is much smaller than
the total number of subjects. In addition, the important special case of
SQGT with equidistant thresholds is discussed in detail, and test construc-
tions are provided for this model as well. Furthermore, a generalization of
the Lindström construction for QGT [36] is described, capable of identifying
any number of defectives, even as large as the total number of subjects. The
results corresponding to SQGT described in this dissertation are partially
available in [37, 38, 39, 40, 41].

1.2 Challenges in dynamic testing, and motivation for
Poisson group testing
The group testing literature may be divided into two categories based on
how the number of defectives is modeled. In combinatorial GT, the number
of defectives, or an upper bound on the number of defectives, is fixed and
assumed to be known in advance [7]. On the other hand, in probabilistic GT
(PGT), the number of defectives is a random variable with a given probability
distribution [16]. With almost no exceptions, the PGT literature focuses on a
binomial \((n, p_0)\) distribution for the number of defectives. Such a model arises
when each of the \(n\) subjects is defective with a fixed probability \(0 < p_0 < 1\),
independent of all other subjects. Binomial models are not necessarily sparse,
given that \(p_0\) may be a constant and given that the defective selection process
is random.

In Chapter 6, we propose a novel GT paradigm, termed Poisson PGT,
which models the distribution of the number of defectives via a right-truncated
Poisson distribution with parameter \(\lambda(n) = o(n)\). Our motivation for this
assumption comes from clinical testing, where one is interested in identifying
infected individuals under the assumption that infections gradually die out.
A similar scenario is encountered in screening DNA clones for the presence
of certain DNA substrings, where the clones are test subjects and defectives

\footnote{SQGT with equidistant thresholds may be viewed as a special instance of quantized integer compressive sensing, introduced in [32], where the entries of the sensing matrices as well as the sparse vectors are allowed to be bounded integers. Another topic in the compressive sensing literature related to this SQGT model is quantized compressive sensing, one instance of which was discussed in [35].}
are clones that contain the given substrings. The distribution of clones containing a given DNA pattern is frequently modeled as Poisson [7]. Other applications include testing genetic traits that are negatively selected for (i.e., traits that diminish in time, as they reduce the fitness of a species). The assumption \( \lambda(n) = o(n) \) ensures that the longer the waiting time or the larger the number of test subjects, the smaller the average relative fraction of defectives. In other words, the rate of defectives diminishes with time.

The Poisson PGT model has a number of useful properties that make it an important alternative to classical binomial models. Although a binomial distribution with \( p_0 < 1 \) and a large \( n \), where \( \lambda = np_0 \) is a constant, converges to a Poisson distribution with parameter \( \lambda = np_0 \) [42], our model allows the parameter \( \lambda(n) \) of the (truncated) Poisson distribution to grow with \( n \); more precisely, the model and the results derived in this work are valid even if \( \lim_{n \to \infty} \lambda(n) = \infty \), as long as \( \lim_{n \to \infty} \frac{\lambda(n)}{n} = 0 \). Such a model is useful in settings where test subjects are assumed to arrive sequentially in time, and where tests are performed only once a sufficient number of subjects \( n \) is present. This model is also applicable to streaming and dynamic testing scenarios [43], in which the probability that a subject is defective decreases in time so that newly arriving subjects are less likely to be defective. In such a setting, classical binomial \((n, p_0)\) models are inadequate, as they assume that the probability \( p_0 \) of a subject being defective does not depend on the number of test subjects.

A number of papers have considered a Poisson model to capture the streaming dynamics of the arrivals of subjects to a test center [44], [45]. In contrast, our model does not make any assumptions on the distribution of the general subject population, but instead focuses on modeling the number of defectives using a right-truncated Poisson distribution. In addition, the focus of [44], [45] is on determining the total amount of time (delay) required to test a batch of subjects arriving at random times. However, here we concentrate on the completely unrelated problem of finding necessary and sufficient conditions on the smallest number of tests needed for accurate nonadaptive and semi-adaptive GT. The results corresponding to Poisson PGT described in this dissertation are partially available in [46, 47].

The dissertation is organized as follows. Chapter 2 describes the SQGT model. Chapter 3 introduces SQ-disjunct and SQ-separable matrices and their properties. In Chapter 4, we describe a number of binary and non-
binary combinatorial and probabilistic constructions for test matrices suitable for SQGT in its most general form: SQGT with arbitrary thresholds. For these constructions, we describe efficient decoding algorithms that can be used to identify the defectives with zero error probability in the presence of errors. In Chapter 5, we focus on an important special case of SQGT: SQGT with equidistant thresholds. For this case, we will describe explicit and probabilistic test matrix constructions and efficient decoding algorithms based on algebraic methods and message passing on factor graphs. In addition, we describe a test matrix construction that enables us to identify any number of defectives among the pool of subjects, even if the number of them is as large as $n$. In Chapter 6, we describe the Poisson PGT model and discuss lower bounds and upper bounds on the number of tests required to find the defectives using nonadaptive and semi-adaptive testing methods. The upper bounds are constructive and therefore correspond to practical testing schemes that can be used for the purpose of identifying the defectives. Finally, in Chapter 7, we describe other ideas that we will pursue in the future corresponding to the application of SQGT for learning interpretable rules for decision making.
CHAPTER 2
THE SQGT MODEL

Throughout this dissertation, 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 and sequences. Asymptotic symbols such as $o(\cdot)$ and $O(\cdot)$ are used in the standard manner. More precisely, we say that $f(x) = O(g(x))$ if and only if there exist $M, x_0 \in \mathbb{R}$, with $M > 0$, such that $|f(x)| \leq M|g(x)|$ for all $x \geq x_0$. Also, $f(x) = o(g(x))$ means that for any $\epsilon > 0$, there exists $x_0 \in \mathbb{R}$ such that $|f(x)| \leq \epsilon|g(x)|$ for all $x \geq x_0$.

Let $\mathbb{Z}^+$ denote the set of positive integers. For a positive integer $n \in \mathbb{Z}^+$, we define $[n] := \{0, 1, \ldots, n-1\}$, and $\lbrack n \rbrack := \{1, 2, \ldots, n\}$. For simplicity, we sometimes use $\mathcal{X} = \{x_i\}_{i=1}^s$ to denote a set of $s$ codewords, $\mathcal{X} = \{x_1, x_2, \ldots, x_s\}$.

Let $n$, $m$, and $d$ denote the number of test subjects, the number of tests, and the number of defectives, respectively. Let $S_i$ denote the $i^{th}$ subject, $i \in \lbrack n \rbrack$, and let $D_j$ be the $j^{th}$ defective, $j \in \lbrack d \rbrack$. Furthermore, let $\mathcal{D}$ denote the set of defectives, so that $|\mathcal{D}| = d$. Let $w \in \{0, 1\}^n$ be a binary vector with its $i^{th}$ coordinate equal to 1 if the $i^{th}$ subject is defective, and 0 otherwise.

We assign to each subject a unique $q$-ary vector of length $m$, termed the codeword of the subject. Each coordinate of the codeword corresponds to a test. Due to the one-to-one correspondence between the codewords and test subjects, with some abuse of notation we use $\mathcal{D}$ to denote both the set of defectives and the set of columns assigned to the defectives. If $x_i \in \{0, 1\}^m$ denotes the codeword of the $i^{th}$ subject, then the $k^{th}$ coordinate of $x_i$, denoted by $x_i(k)$, may be viewed as the “amount” (or strength) of $S_i$ used in the $k^{th}$ test.\(^1\) Note that the symbol 0 indicates that $S_i$ is not present in the test. We denote the test matrix, or equivalently, the code, by $C \in \{0, 1\}^{m \times n}$.

The result of each test in SQGT is an integer from the set $\{Q\}$, $Q > 0$. The

\(^1\)Note that $q$ is actually the available alphabet size and not necessarily the effective alphabet size. In many constructions in this report, we use an effective alphabet size smaller than $q$, but if the maximum available entry of the alphabet is $q - 1$, we still call the alphabet size $q$. 

10
Table 2.1: Table of symbols and their definitions

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>Total number of subjects</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of tests</td>
</tr>
<tr>
<td>$d$</td>
<td>Number of defectives</td>
</tr>
<tr>
<td>$Q$</td>
<td>Size of the output alphabet</td>
</tr>
<tr>
<td>$q$</td>
<td>Size of the test matrix alphabet</td>
</tr>
<tr>
<td>$\eta_l$</td>
<td>The $l$th threshold where $l \in {Q}$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
<td>Set of defectives</td>
</tr>
<tr>
<td>$w \in [2]^n$</td>
<td>Indicator vector of defectives</td>
</tr>
<tr>
<td>$y \in [Q]^m$</td>
<td>Vector of test results</td>
</tr>
<tr>
<td>$x_i \in [q]^m$</td>
<td>Codeword assigned to the $i$th subject</td>
</tr>
<tr>
<td>$C \in [q]^{m \times n}$</td>
<td>Code (test matrix)</td>
</tr>
<tr>
<td>$e$</td>
<td>Number of errors in $y$ that $C$ can correct</td>
</tr>
</tbody>
</table>

Results of all $m$ tests are represented using a vector of length $m$ called the vector of test results, $y$. Table 2.1 summarizes these notations.

Each test outcome depends on the number of defectives and their sample amount in the test through a quantization function $f_\eta(\cdot)$ with $Q$ thresholds, $\eta_l (l \in \{1, 2, \ldots, Q\})$, defined as follows.

**Definition 1.** For a set of thresholds $\eta = [\eta_0 = 0, \eta_1, \ldots, \eta_Q]^T$ and a scalar $\alpha \in \mathbb{Z}^+$, we define the quantization function $f_\eta : \mathbb{Z}^+ \rightarrow [Q]$ as

$$f_\eta(\alpha) = r \quad \text{if} \quad \eta_r \leq \alpha < \eta_{r+1},$$

where $r \in [Q]$. In words, the function $f_\eta(\alpha)$ returns the index of the quantization bin that contains its argument.

For a vector of positive integers $\alpha$, $f_\eta(\alpha)$ is a vector with each entry equal to the quantization of the corresponding entry of $\alpha$ according to Def. 1. For two scalars $\alpha, \alpha' \in \mathbb{Z}^+$, and a set of thresholds $\eta$, we write $\alpha >_\eta \alpha'$ to indicate that $f_\eta(\alpha) > f_\eta(\alpha')$. Next, we define the syndrome of a set of codewords using $f_\eta(\cdot)$.

**Definition 2.** Let $\mathcal{X} = \{x_1, x_2, \ldots, x_s\} = \{x_j\}_1^s$ be a set of $s \geq 1$ codewords of length $m$ in a SQGT model with thresholds $\eta = [\eta_0 = 0, \eta_1, \eta_2, \ldots, \eta_Q]^T$. The syndrome of $\mathcal{X}$, denoted by $y_\mathcal{X} \in [Q]^m$, is defined as $y_\mathcal{X} = f_\eta(\Sigma_{j=1}^s x_j)$.

By this definition, in the absence of any errors, the vector of test results is equal to the syndrome of defectives, i.e. $y = y_\mathcal{D}$. However, when errors occur,
The outcome of the $k^{\text{th}}$ test and its relationship with $\sum_{j=1}^{d} x_{ij}(k)$ through the thresholds in a SQGT model with (possibly) non-binary test design.

$$y(k)$$

$$\sum_{j=1}^{d} x_{ij}(k) \begin{array}{ccccccc}
0 & 1 & 2 & \cdots & Q - 1 \\
0, 1, \cdots, \eta_{1} - 1, \eta_{1}, \cdots, \eta_{2} - 1, & \cdots & \eta_{Q - 1}, \cdots, \eta_{Q} - 1.
\end{array}$$

Figure 2.1: The outcome of the $k^{\text{th}}$ test and its relationship with $\sum_{j=1}^{d} x_{ij}(k)$ through the thresholds in a SQGT model with (possibly) non-binary test design.

$$w^T = (0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1)$$

$$C = \begin{pmatrix}
0 & 1 & 0 & 1 & 2 & 0 & 0 & 2 & 1 & 1 \\
1 & 2 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\
2 & 0 & 2 & 2 & 0 & 2 & 1 & 1 & 1 & 1 \\
0 & 2 & 1 & 0 & 2 & 0 & 1 & 2 & 0 & 0 \\
1 & 1 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 1
\end{pmatrix}$$

$$y = \begin{pmatrix}
1 \\
3 \\
0 \\
2
\end{pmatrix}$$

$$S_1 \ S_2 \ S_3 \ S_4 \ S_5 \ S_6 \ S_7 \ S_8 \ S_9 \ S_{10}$$

Figure 2.2: A test matrix $C$, indicator vector of defectives $w$, and the corresponding vector of test results $y$, for an SQGT scheme with $d = 3$, $m = 5$, $n = 10$, $q = 3$, $Q = 4$, and $\eta = [0, 2, 3, 5, 7]^T$.

Some entries of $y$ may differ from $y_D$. In particular, if $e$ tests are erroneous, we assume that $e$ entries of $y_D$ have changed to an arbitrary value in $[Q]^2$.

The relationship between the syndrome of defectives and the strength of the defectives in a test is illustrated in Fig. 2.1. One should note that an underlying assumption in the SQGT model is that $\eta_Q > d(q - 1)$ to ensure that the sum of entries corresponding to defectives is always smaller than $\eta_Q$.

Figure 2.2 provides an example of a SQGT test matrix, an incidence vector of the defectives, and vector of test results, with $d = 3$, $m = 5$, $n = 10$, $q = 3$, $Q = 4$, and $\eta = [0, 2, 3, 5, 7]^T$.

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 CGT. Furthermore, if $Q - 1 = d(q - 1)$ and $\forall r \in [Q]$, $\eta_r = r$, then SQGT reduces to the adder model (QGT), with a possibly non-binary test matrix. Similarly, TGT with zero gap and the model in [21] also represent special instances of SQGT. Another interesting

$^2$Note that this assumption corresponds to the case in which no information is available regarding the pattern of errors (i.e. worst case scenario). However, more informative assumptions regarding the error pattern can be considered to simplify the problem, e.g. errors that change the outcome of a test to the value corresponding to an adjacent bin.
Figure 2.3: Different group testing models for the case $q = 2$. In the figures, $\eta_T$ denotes the threshold in TGT and $\eta_{DR}$ denotes the threshold in the model described in [21].

A special case of the SQGT model is the case where the quantizer has equidistant thresholds. In this case, $\eta_r = r\eta$, where $r \in [Q + 1]$, and the syndrome of $s$ codewords $\mathcal{X}$ simplifies to $y_x = \left\lfloor \frac{x_1 + x_2 + \ldots + x_s}{\eta} \right\rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Figure 2.3 illustrates all these models for $q = 2$. 
In this chapter, we introduce two families of test matrices suitable for SQGT, termed $SQ\text{-}disjunct$ and $SQ\text{-}separable$ matrices. For each of these families, we use a set of parameters as described below. A $[q;Q;\eta;(l : u);e]$-$SQ$-disjunct/separable matrix is a $q$-ary test matrix for a SQGT model with thresholds $\eta = [0,\eta_1,\eta_2,\ldots,\eta_Q]^T$. Such a test matrix is capable of uniquely identifying any number of defectives between $l$ and $u$, $l\leq d \leq u$, from a $Q$-ary vector of test results containing up to $e$ erroneous test results. For simplicity, when the test matrix can only identify exactly $d$ defectives (i.e. $l = u = d$), we use $d$ instead of $(l : u)$.

3.1 SQ-disjunct matrices

The SQ-disjunct matrices are generalizations of binary disjunct matrices introduced in [24] for efficient zero-error identification of defectives in CGT.

We start by defining the binary disjunct and separable matrices for CGT [24].

**Definition 3 (Binary $d$-disjunct matrices for CGT).** A binary $d$-disjunct matrix designed for CGT, capable of correcting up to $e$ errors, is an $m \times n$ matrix such that for any set of $d+1$ columns, $\mathcal{X} = \{x_j\}^{d+1}_1$, and for any column $x_i \in \mathcal{X}$, there exists a set of coordinates $\mathcal{R}_i$ of size at least $2e + 1$, such that $\forall k \in \mathcal{R}_i$, $x_i(k) = 1$ and $x_j(k) = 0$, for $x_j \in \mathcal{X}$ and $j \neq i$.

In order to introduce the new family of SQ-disjunct matrices, we need the following definition.

**Definition 4.** A set of columns $\mathcal{X} = \{x_j\}^s_1$ with syndrome $\mathbf{y}_x$ is said to be included in another set of columns $\mathcal{Z} = \{z_j\}^t_1$ with syndrome $\mathbf{y}_z$, if $\forall i \in [m]$, $\mathbf{y}_x(i) \leq \mathbf{y}_z(i)$. We denote this inclusion property by $\mathcal{X} \bowtie \mathcal{Z}$, or equivalently, $\mathbf{y}_x \bowtie \mathbf{y}_z$. 

14
Remark 1. Using this definition, it can be easily verified that if \( X \subseteq Z \), then \( X \preceq Z \).

Note that for \( q = Q = 2 \) and \( \eta_1 = 1 \), Definition 2 is equivalent to the definition of inclusion for disjunct matrices in CGT [24]. Based on the notion of inclusion, we may define SQ-disjunct matrices for the error-free scenario, \( e = 0 \).

Definition 5. An \( m \times n \) matrix is called a \([q;Q;\eta;(1:d);0]\)-SQ-disjunct matrix if \( \forall s, t \leq d \) and for any sets of \( q \)-ary columns \( X = \{x_j\}_1^s \) and \( Z = \{z_j\}_1^t \), \( X \preceq Z \) implies \( X \subseteq Z \).

The next two theorems describe some properties of SQ-disjunct matrices.

**Theorem 1.** A \([q;Q;\eta;(1:d);0]\)-SQ-disjunct matrix is capable of identifying any number of defectives less than or equal to \( d \) in the absence of test errors. In other words, given an error-free vector of test results \( y \in [Q]^m \), any column with a syndrome included in \( y \) corresponds to a defective, and any column with a syndrome not included in \( y \) corresponds to a non-defective.

**Proof.** Let \( x_i, i \in [n] \), be a column of a \([q;Q;\eta;(1:d);0]\)-SQ-disjunct matrix. Since \( y = y_D \), if \( i \) corresponds to a defective, i.e. \( i \in D \), we have \( y_{(x_i)} \preceq y \). Conversely, by Definition 5, it can be easily verified that if \( i \notin D \) and \( |D| \leq d \), then \( y_{(x_i)} \not\preceq y \).

We also prove the following result used in subsequent derivations.

**Theorem 2.** A matrix is \([q;Q;\eta;(1:d);0]\)-SQ-disjunct if and only if no column is included in a set of \( d \) other columns.

**Proof.** It is easy to verify that if a matrix is \([q;Q;\eta;(1:d);0]\)-SQ-disjunct, then no column is included in the set of \( d \) other columns.

Conversely, let \( X = \{x_j\}_1^s \) and \( Z = \{z_j\}_1^t \) be two sets of columns where \( s, t \leq d \). From the assumption that no column is included in a set of \( d \) other columns, one can conclude that no column is included in a set of \( t \) other columns whenever \( t \leq d \). If \( X \preceq Z \) but \( X \notin Z \), then there exists a column \( x_j \in X, j \in [s] \), such that \( \{x_j\} \notin Z \). But since \( \{x_j\} \preceq X \preceq Z \), then \( \{x_j\} \preceq Z \), which contradicts the assumption that no column is included in \( t \) other columns. \( \square \)
Remark 2. From Theorem 2, one can conclude that a matrix is \([q;Q;\eta;(1:d);0]-SQ\)-disjunct if and only if for any set of \(d+1\) columns, \(\mathcal{X} = \{x_j\}_{1}^{d+1}\), and for any column \(x_i \in \mathcal{X}\), there exists at least one unique coordinate \(k_i\) for which

\[
y(x_i)(k_i) > y_{\mathcal{X}\backslash\{x_i\}}(k_i),
\]

where \(y(x_i)\) is the syndrome of \(\{x_i\}\), and \(y_{\mathcal{X}\backslash\{x_i\}}\) is the syndrome of the other \(d\) columns in \(\mathcal{X}\). Note that for equidistant SQGT, (3.1) implies

\[
\left[ \frac{x_i(k_i)}{\eta} \right] > \left[ \frac{\sum_{j=1,j\neq i}^{d+1} x_j(k_i)}{\eta} \right].
\]

The uniqueness property in Remark 2 can be proved as follows. Fix a set \(\mathcal{X}\) and \(x_i, x_j \in \mathcal{X}\) such that \(i \neq j\) and \(k_i = k_j\). Using Definition 2, it can be easily verified that for any coordinate \(k\),

\[
y_{\mathcal{X}\backslash\{x_i, x_j\}}(k) = y_{\mathcal{X}\backslash\{x_i\}}(k) \geq y_{\{x_j\}}(k). \tag{3.2}
\]

Using (3.1) and (3.2), one has

\[
y(x_i)(k_i) > y_{\mathcal{X}\backslash\{x_i\}}(k_i) \geq y_{\{x_j\}}(k_i). \tag{3.3}
\]

Applying condition (3.1) to \(x_j\) and using (3.2), one similarly obtains

\[
y(x_j)(k_j) > y_{\mathcal{X}\backslash\{x_j\}}(k_j) \geq y_{\{x_i\}}(k_j). \tag{3.4}
\]

Since \(k_i = k_j\), (3.3) and (3.4) contradict each other, which completes the proof.

Using the notion of unique coordinate, we can generalize Definition (5) to SQ-disjunct matrices that are capable of correcting up to \(e > 0\) errors.

Definition 6 (SQ-disjunct matrices). An \(m \times n\) matrix is called a \([q;Q;\eta;(1:d);e]-SQ\)-disjunct matrix if for any set of \(d+1\) columns, \(\mathcal{X} = \{x_j\}_{1}^{d+1}\), and for any column \(x_i \in \mathcal{X}\), there exists a set of coordinates, \(\mathcal{R}_i\), of size at least \(2e+1\) such that \(\forall k_i \in \mathcal{R}_i,\)

\[
y(x_i)(k_i) > y_{\mathcal{X}\backslash\{x_i\}}(k_i),
\]

and \(\mathcal{R}_i\) is disjoint of any \(\mathcal{R}_l\) for which \(x_l \in \mathcal{X}\) and \(l \neq i\); in this equation
\( y_{(x_i)} \) is the syndrome of \( \{x_i\} \), and \( y_{\mathcal{X} \setminus (x_i)} \) is the syndrome of the remaining \( d \) columns in \( \mathcal{X} \).

Such a matrix is capable of uniquely identifying up to \( d \) defectives, in the presence of up to \( e \) errors in the vector of test results. If a column \( x_i \) does not correspond to a defective, its syndrome contains at least \( e + 1 \) coordinates satisfying \( y_{(x_i)}(k) > y(k) \). On the other hand, if \( x_i \) corresponds to a defective, its syndrome contains at most \( e \) coordinates satisfying \( y_{(x_i)}(k) > y(k) \).

**Remark 3.** It can be easily seen from (3.1) and (3.5) that a necessary condition for the existence of a \([q; Q; \eta; (1 : d); e]\)-SQ-disjunct matrix is that \( q - 1 \geq \eta_1 \). As a result, there exist no binary \([2; Q; \eta; (1 : d); e]\)-SQ-disjunct matrices when \( \eta_1 > 1 \).

**Remark 4 (Decoding Algorithm:).** Definition 6 suggests an efficient decoding algorithm for SQ-disjunct matrices with complexity \( O(mn) \), which resembles the decoding algorithm for binary disjunct matrices for CGT. The decoding algorithm for an \( m \times n \) \([q; Q; \eta; (1 : d); e]\)-SQ-disjunct matrix works as follows. For each column \( x_i, i \in [n] \), count the number of coordinates of \( y_{(x_i)} \) for which \( y_{(x_i)}(k) > y(k) \). If the number of such coordinates is at least \( e + 1 \), \( x_i \) does not correspond to a defective. On the other hand, if the number of such coordinates is at most \( e \), the column corresponds to a defective.

### 3.2 SQ-separable matrices

Although SQ-disjunct matrices can be used to find defectives in a SQGT model via a simple decoding procedure, the requirements imposed on such matrices may appear too restrictive for certain applications. As a result, relaxing these structural constraints may lead to a reduction in the number of tests for fixed values of \( n \). Another shortcoming of SQ-disjunct matrices is that their alphabet size must satisfy \( q > \eta_1 \); as a result, in situations where smaller alphabet size is required, such matrices cannot be utilized. SQ-separable matrices are a family of \( q \)-ary matrices that are capable of overcoming the aforementioned issues. The SQ-separable matrices are generalizations of binary separable matrices for CGT [24], defined below.

**Definition 7 (Binary \( d \)-separable matrices for CGT).** A binary \( d \)-separable matrix designed for CGT, capable of correcting up to \( e \) errors, is
an \( m \times n \) matrix such that for any two distinct sets of columns \( \mathcal{X} \) and \( \mathcal{Z} \), 
\[ 1 \leq |\mathcal{X}|, |\mathcal{Z}| \leq d, \]
the Boolean sum of the columns in \( \mathcal{X} \) differs from the Boolean sum of the columns in \( \mathcal{Z} \) in at least \( 2e + 1 \) coordinates.

**Definition 8 (SQ-separable matrices).** An \( m \times n \) matrix is called a \([q; Q; \eta; (l: u); e]\)-SQ-separable matrix if for any two distinct sets of columns \( \mathcal{X} \) and \( \mathcal{Z} \) that satisfy \( l \leq |\mathcal{X}|, |\mathcal{Z}| \leq u \), there exists a set of coordinates \( \mathcal{R} \), with size \( |\mathcal{R}| \geq 2e + 1 \), such that \( \forall k \in \mathcal{R} \)
\[ y_{\mathcal{X}}(k) \neq y_{\mathcal{Z}}(k). \]

Such matrices are capable of identifying defectives when the vector of test results contains at most \( e \) errors, given that the number of defectives is at least \( l \) and at most \( u \). Note that as the next proposition demonstrates, SQ-disjunct matrices are special cases of SQ-separable matrices.

**Proposition 1.** Any \([q; Q; \eta; (1:d); e]\)-SQ-disjunct matrix is a \([q; Q; \eta; (1:d); e]\)-SQ-separable matrix.

**Proof.** Consider a \([q; Q; \eta; (1:d); e]\)-SQ-disjunct matrix and any two distinct sets of columns \( \mathcal{X} \) and \( \mathcal{Z} \) that satisfy \( 1 \leq |\mathcal{X}|, |\mathcal{Z}| \leq d \). Without loss of generality, assume that \( |\mathcal{X}| \leq |\mathcal{Z}| \). Since these two sets are distinct, \( \mathcal{Z} \setminus \mathcal{X} \neq \emptyset \); let \( z \) be a column such that \( z \in \mathcal{Z} \setminus \mathcal{X} \). Since \( |\mathcal{X} \cup \{z\}| \leq d + 1 \), using the definition of SQ-disjunct matrices, one can conclude that there exists a set of coordinates, \( \mathcal{R} \), of size at least \( 2e + 1 \), such that \( \forall k \in \mathcal{R} \),
\[ y_{\mathcal{X}}(k) \neq y_{\mathcal{Z}}(k). \]

On the other hand since \( z \in \mathcal{Z} \), Definition 2 implies that \( \forall k \in \mathcal{R} \), \( y_{\mathcal{Z}}(k) \geq y_{\mathcal{X}}(k) \), which completes the proof. \(\square\)

**Remark 5.** From Definition 8, one can see that a necessary condition for the existence of a \([q; Q; \eta; (l: u); e]\)-SQ-separable matrix is that \( l(q - 1) \geq \eta_1 \). If \( l = 1 \), this condition simplifies to \( q - 1 \geq \eta_1 \), which is the same as the necessary condition for the existence of a \([q; Q; \eta; (1:d); e]\)-SQ-disjunct matrix. This is expected, since any SQ-disjunct matrix is also a SQ-separable matrix, while the converse is not true. On the other hand, if \( q = 2 \), the condition simplifies to \( l \geq \eta_1 \). This implies that if the number of defectives is smaller than \( \eta_1 \), one cannot identify the defectives using a binary matrix.
In this chapter, we describe test matrix constructions and efficient decoding algorithms for the general SQGT model in which the thresholds are arbitrary but are known in advance. We start this chapter by introducing two simple methods for constructing SQ-disjunct and SQ-separable matrices from binary disjunct and binary separable matrices for CGT. These two simple constructions are then used as building blocks (BBs) to construct more sophisticated $q$-ary test matrices for SQGT. These constructions are based on three new families of integer sequences which represent extensions and generalizations of $B_h$ and certain types of super-increasing and lexicographically ordered sequences [48]. These novel sequences are termed “quantized $B_h$ sequences”, “type-s semi-quantitative lexicographically ordered sequences (SQLO_s)” and “type-l semi-quantitative lexicographically ordered sequences (SQLO_l)”. While SQLO_s and SQLO_l sequences are special cases of quantized $B_h$ sequences, they exhibit a special nested structure that allows for computationally efficient decoding algorithms. Finally, in the last section, we discuss constructions of binary test matrices for applications in which $q$-ary test matrices with $q > 2$ are not applicable.

A summary of different properties of the constructions described in this chapter is provided in Tables 4.1 and 4.2. Since several constructions were based on classical binary $d$-disjunct and $d$-separable matrices, we explicitly included these underlying building blocks (BBs) in “Features”. In these cases, the number of tests $m$ as a function of $d$, $e$ and $n$, depends on the specific BBs used. Given that there are many different constructions for classical binary $d$-disjunct and $d$-separable matrices available in the literature, a comprehensive survey of all possible SQGT matrices would be well beyond the scope of this dissertation. We therefore focused on a small set of classical binary disjunct and separable matrices well-documented in the literature, e.g. [7] and [22].
Table 4.1: A comparative summary of SQGT matrices described in Constructions 1-4

<table>
<thead>
<tr>
<th>Test Matrix</th>
<th>Construction 1</th>
<th>Construction 2</th>
<th>Construction 3</th>
<th>Construction 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>$[q; Q; \eta; (1:d); e]$</td>
<td>$[q; Q; \eta; (1:d); e]$</td>
<td>$[q; Q; \eta; (1:d); e]$</td>
<td>$[q; Q; \eta; (1:d); e]$</td>
</tr>
<tr>
<td>Type</td>
<td>SQ-disjunct</td>
<td>SQ-separable</td>
<td>SQ-separable</td>
<td>SQ-separable</td>
</tr>
<tr>
<td>Thresholds</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Construction</td>
<td>Explicit</td>
<td>Explicit</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Num. Tests</td>
<td>$O(ed^2 \log_2 \frac{n}{d})$</td>
<td>$O(ed^2 \log_2 \frac{n}{d})$</td>
<td>$O\left(\frac{ed^2 \log_2 \frac{n}{d} K}{mK}\right)$</td>
<td>$O\left(\frac{ed^2 \log_2 \frac{n}{d} K}{mK}\right)$</td>
</tr>
<tr>
<td>Features</td>
<td>Efficient decoder of complexity $O(mn)$, BB: disjunct (CGT)</td>
<td>BB: $d$-separable (CGT)</td>
<td>Decoder of complexity $O\left(\frac{mn}{K} + 2^K (K + md)\right)$, BB: $d$-disjunct (CGT), Uses $K$ elements of a quantized $B_d$ sequence</td>
<td>Decoder of complexity $O\left(\frac{mn}{K} + dm \log m + \text{deg}_{\text{max}} K\right)$, BB: $d$-disjunct (CGT), Uses $K$ elements of a SQLO$_d(\eta, d)$ sequence</td>
</tr>
</tbody>
</table>
Table 4.2: A comparative summary of SQGT matrices described in Constructions 5-7

<table>
<thead>
<tr>
<th>Test Matrix</th>
<th>Construction 5</th>
<th>Construction 6</th>
<th>Construction 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>$[q;Q;\eta; (1:d); e]$</td>
<td>$[2;Q;\eta; (\eta_\alpha;d); e]$</td>
<td>$[2;Q;\eta; (\eta_\alpha;d); e]$</td>
</tr>
<tr>
<td>Type</td>
<td>SQ-separable</td>
<td>SQ-separable</td>
<td>SQ-separable</td>
</tr>
<tr>
<td>Thresholds</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Construction</td>
<td>Explicit</td>
<td>Probabilistic</td>
<td>Explicit</td>
</tr>
<tr>
<td>Num. Tests</td>
<td>$O\left(e^2d^2\log_2\frac{n}{dK}\right)$</td>
<td>$O_e\left(d^2\log_2\log_2\frac{n}{\eta}\right)$</td>
<td>$O_e\left(\frac{n+1}{\eta_\alpha-1}d^3\log_2\log^3(\log n)\log n\right)$</td>
</tr>
<tr>
<td>Features</td>
<td>Decoder of complexity $O\left(\frac{mn}{K} + dm\log m + deg_{max}K\right)$,</td>
<td>Binary test matrix</td>
<td>Binary test matrix</td>
</tr>
<tr>
<td></td>
<td>BB: $d$-disjunct (CGT), Use $K$ elements of a SQLO$_k(\eta,d)$ sequence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1 Construction of $q$-ary SQGT test matrices using binary disjunct/separable matrices for CGT

Given that SQ-disjunct matrices (Def. 6) represent generalizations of conventional binary disjunct matrices (Def. 3), it is expected that one can construct SQ-disjunct matrices using conventional disjunct matrices. The following proposition describes one such construction.

**Proposition 2 (Construction 1).** Any matrix generated by multiplying a conventional binary $d$-disjunct matrix capable of correcting $e$ errors\(^1\) by $q - 1$, where $q - 1 \geq \eta_1$, is $[q; Q; \eta; (1:d); e]$-SQ-disjunct.

**Proof.** A conventional binary $d$-disjunct matrix, capable of correcting $e$ errors, satisfies the condition that for any set of $d + 1$ columns, $Z = \{z_j\}_{d+1}^1$, and for any column $z_i \in Z$, there exists a set of coordinates $R_i$ of size at least $2e + 1$, such that $\forall k \in R_i$,

$$z_i(k) = 1,$$

$$z_j(k) = 0, \quad \text{for } z_j \in Z \text{ and } j \neq i.$$

Multiplying such a matrix with $q - 1$, where $q - 1 \geq \eta_1$, produces a $q$-ary matrix such that for any set of $d + 1$ columns, $X = \{x_j\}_{d+1}^1$, and for any column $x_i \in X$, there exists a unique set of coordinates, $R_i$, of size at least $2e + 1$, such that $\forall k \in R_i$,

$$y_{(x_i)}(k) > 0,$$

$$x_j(k) = 0, \quad \text{for } x_j \in X \text{ and } j \neq i.$$

As a result, $\forall k \in R_i$,

$$y_{(x_i)}(k) > y_{X \setminus \{x_i\}}(k) = 0.$$

$\square$

Similar to the case of SQ-disjunct matrices, SQ-separable matrices (Def. 8) may also be constructed from classical binary separable matrices (Def. 7), as shown in the next proposition.

---

\(^1\)For constructions of binary $d$-disjunct matrices with error correcting capabilities, see [7], [49], [50] and references therein.
Proposition 3 (Construction 2). Any matrix generated by multiplying a conventional binary $d$-separable matrix capable of correcting up to $e$ errors by $q - 1$, where $q - 1 \geq \eta_1$, represents a $[q; Q; \eta_1; (1:d); e]$-SQ-separable matrix.

Proof. The proof is a direct result of Def. 7 and Def. 8, and is hence omitted. □

Constructions 1 and 2 provide two simple methods to generate SQ-disjunct and SQ-separable matrices from their binary counterparts for CGT. However, these constructions do not fully utilize all the information provided by the $Q$ thresholds and the $Q$-ary vector of test results: the dimensions of the SQ-disjunct and SQ-separable matrices are the same as the dimensions of the binary disjunct and binary separable matrices. In the next section, we will describe more sophisticated constructions that overcome this issue.

4.2 SQ-separable matrices using quantized $B_h$ sequences

In this section, we introduce the notion of quantized $B_h$ sequences and describe how to use their elements in conjunction with binary disjunct matrices to construct SQ-separable matrices. In addition, we describe construction methods for quantized $B_h$ sequences as well as efficient decoding methods for the resulting matrices. The gist of these constructions is horizontal matrix concatenation, defined as follows.

Definition 9 (Horizontal concatenation). Consider $K \geq 2$ matrices $C_j \in \mathbb{R}^{m \times n}$, $1 \leq j \leq K$. The horizontal concatenation of these matrices is a matrix defined by $C = [C_1, C_2 \ldots, C_K]$, such that for $j \in [K]$ and $l \in [n]$, the $((j - 1)n + l)^{th}$ column of $C$ is equal to the $l^{th}$ column of $C_j$.

Before describing the main results of this section, we introduce a simple matrix construction that provides the intuition behind the derivations of the main results.

Theorem 3. Consider a SQGT model with thresholds $\eta = [0, \eta_1, \eta_2, \ldots, \eta_Q]^T$ satisfying $\eta_Q > \eta_1 + \max\{\eta_2, \eta_3 - \eta_1\}$ and $Q \geq 4$. Fix a binary $d$-disjunct matrix $C_b$ of dimensions $m_b \times n_b$, capable of correcting up to $e$ errors. Form a matrix
C of length \( m = m_b \) and size \( n = 2m_b \) by concatenating \( C_1 = \alpha_1 C_b \) and \( C_2 = \alpha_2 C_b \) horizontally, where \( \alpha_1 = \eta_1 \) and \( \alpha_2 = \max\{\eta_2, \eta_3 - \eta_1\} \). The constructed matrix is a \([q; Q; \eta; (1:d); e]\)-SQ-separable matrix with \( q = \max\{\eta_2, \eta_3 - \eta_1\} + 1 \).

**Proof.** Consider two distinct subsets of columns, \( X_1 \) and \( X_2 \), such that \( 1 \leq |X_1|, |X_2| \leq d \). Without loss of generality, assume that \( |X_1| \leq |X_2| \). Since the two sets are distinct, \( X_2 \setminus X_1 \neq \emptyset \). Let \( z' \in X_2 \setminus X_1 \). By construction, \( z' = \alpha z_b \) for some \( \alpha \in \{\alpha_1, \alpha_2\} \) and some binary column \( z_b \) of \( C_b \). Let \( z'' \) be another column of \( C \) with the same support as \( z' \), obtained by multiplying \( z_b \) by \( \{\alpha_1, \alpha_2\} \setminus \{\alpha\} \).

If \( z'' \notin X_1 \), then by the construction of \( C \) and Def. 3, there exists a set of coordinates \( R \) of size at least \( 2\epsilon + 1 \), such that \( \forall k \in R, z'(k) \geq \alpha_1 = \eta_1 \) and \( z(k) = 0 \), \( \forall x \in X_1 \). Since \( \forall k \in R, \sum_{x \in X_2} x(k) \geq z'(k) \geq \eta_1 \), and \( \sum_{x \in X_1} x(k) = 0 \), it follows that

\[
y_{x_2}(k) \geq y_{(\omega', \omega'')} (k) > y_{x_1} (k).
\]

On the other hand, if \( z'' \in X_1 \setminus X_2 \), there exists a set of coordinates \( R \) of size at least \( 2\epsilon + 1 \), such that \( \forall k \in R, z'(k) \in \{\alpha_1, \alpha_2\}, z''(k) \in \{\alpha_1, \alpha_2\}, \) and \( z(k) = 0 \) \( \forall x \in X_1 \setminus \{z''\} \). Since \( \forall k \in R, \sum_{x \in X_2} x(k) \geq z'(k) + z''(k) = \alpha_1 + \alpha_2 = \max\{\eta_1 + \eta_2, \eta_3\} \geq \eta_3 \) and \( \sum_{x \in X_1} x(k) \leq \alpha_2 < \eta_3 \), and since \( \eta_Q > \eta_1 + \max\{\eta_2, \eta_3 - \eta_1\} \), it follows that

\[
y_{x_2}(k) \geq y_{(\omega', \omega'')} (k) > y_{x_1} (k).
\]

If \( z'' \in X_1 \setminus X_2 \), we have to separately analyze two cases: if \( z' = \alpha_2 z_b \), then there exists a set of coordinates \( R \) of size at least \( 2\epsilon + 1 \), such that \( \forall k \in R, z'(k) = \alpha_2, z''(k) = \alpha_1, \) and \( z(k) = 0 \) \( \forall x \in X_1 \setminus \{z''\} \). Since \( \forall k \in R, \sum_{x \in X_2} x(k) \geq z'(k) = \alpha_2 \geq \eta_2 \), and \( \sum_{x \in X_1} x(k) = \alpha_1 = \eta_1 < \eta_2 \), it follows that

\[
y_{x_2}(k) \geq y_{(\omega', \omega'')} (k) > y_{x_1} (k).
\]

However, for the case that \( z'' \in X_1 \setminus X_2 \) and \( z' = \alpha_1 z_b \), there exists a set of coordinates \( R \) of size at least \( 2\epsilon + 1 \), such that \( \forall k \in R, z'(k) = \alpha_1, z''(k) = \alpha_2, \) and \( z(k) = 0 \) \( \forall x \in X_2 \setminus \{z'\} \). Since \( \forall k \in R, \sum_{x \in X_1} x(k) \geq z''(k) = \alpha_2 \geq \eta_2 \), and \( \sum_{x \in X_2} x(k) = \alpha_1 = \eta_1 < \eta_2 \), we conclude that

\[
y_{x_2}(k) < y_{(\omega', \omega'')} (k) \leq y_{x_1} (k).
\]

This completes the proof. \( \square \)
In Construction 1, we showed that multiplying a binary $d$-disjunct matrix of dimension $m_b \times n_b$ by $\eta_1$ results in a SQ-disjunct matrix of the same dimension. On the other hand, Thm. 3 shows that one may increase the number of test subjects twofold, using only $m = m_b$ tests. The increase is achieved by using a carefully chosen multiplier for the second block. More precisely, this choice of $\alpha_2$ satisfies two properties. First, since $\alpha_2 \succ \eta \alpha_1$, none of the two columns of $C$ have the same syndrome, and therefore can be uniquely distinguished. Second, the fact that $\alpha_1 + \alpha_2 \succ \eta \alpha_2 \succ \eta \alpha_1$ ensures that if we can identify a column of $C_b$ that corresponds to at least one defective, denoted by $x_b$, it is possible to determine if $\{\alpha_1 x_b\}$, or $\{\alpha_2 x_b\}$, or $\{\alpha_1 x_b, \alpha_2 x_b\}$ are the columns of $C$ that correspond to the defectives. These two properties, combined with the disjunctness property of $C_b$, ensure that any collection of up to $d$ items has a unique syndrome after quantization, even in the presence of up to $e$ errors. This construction can be generalized to include concatenations of more than two matrices using the new families of quantized $B_h$ sequences described next.

We start by defining the standard $B_h$ sequences [48]. Note that throughout this chapter, with slight abuse of notation, we use $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ to denote both a set and/or a sequence consisting of $K$ positive integers. The exact meaning will be apparent from the context, and it will depend on which property of $A$ is being discussed. Note that for a set of positive integers $A$, one can view the natural ordering of the elements of $A$ as the corresponding sequence.

**Definition 10 ($B_h$ sequence).** A finite sequence of positive integers $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ is a $B_h$ sequence if $\forall A_1, A_2 \subseteq A$ such that $A_1 \neq A_2$, $|A_1| = |A_2| = h$, one has $\sum_{\alpha_i \in A_1} \alpha_i \neq \sum_{\alpha_i \in A_2} \alpha_i$.

Similar to the classical $B_h$ sequences which require distinct subset sums of cardinality $h$, in quantized $B_h$ sequences we require that the quantized sums of subsets of size up to $h$ be distinct. These sequences can be used to generalize Thm. 3 to construct SQ-separable matrices.

**Definition 11 (Quantized $B_h$ sequence).** A finite sequence of positive integers $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ is called a quantized $B_h$ sequence with respect to $\eta$ if

1. $\alpha_K \succ \eta \alpha_{K-1} \succ \eta \ldots \succ \eta \alpha_1 \succ \eta 0$ (i.e., all elements of $A$ lie in different quantization bins).
2. \( \forall A_1, A_2 \in A \) such that \( A_1 \neq A_2 \), \( |A_1| \leq h \) and \( |A_2| \leq h \), one either has 
\[ \sum_{\alpha_i \in A_1} \alpha_i > \eta \sum_{\alpha_i \in A_2} \alpha_i \text{ or } \sum_{\alpha_i \in A_2} \alpha_i > \eta \sum_{\alpha_i \in A_1} \alpha_i \]  
(the sums of elements of distinct subsets lie in different quantization bins).

Intuitively, we require that all the elements of the sequence are located in different quantization bins, none of them is in the same bin as 0, and in addition, all the sums that are formed by adding elements of subsets of cardinality at most \( h \) fall into different bins. Note that when \( K = 2 \), setting \( \alpha_1 = \eta_1 \) and \( \alpha_2 = \max\{\eta_2, \eta_3 - \eta_1\} \) as was done in Thm. 3 ensures that the conditions in the aforementioned definition are met.

**Remark 6.** Note that the cardinality of a finite quantized \( B_h \) sequence may be smaller than the value of \( h \). For example, \( A = \{\eta_1\} \) is a quantized \( B_h \) sequence with respect to \( \eta \), for any \( h \in \mathbb{Z}^+ \). However, one seeks to find the densest such sequence given an upper bound on the values of its largest element.

Quantized \( B_h \) sequences can be used to construct SQ-separable matrices as shown in the next theorem.

**Theorem 4 (Construction 3).** Fix a binary \( d \)-disjunct matrix \( C_b \) of dimensions \( m_b \times n_b \), capable of correcting up to \( e \) errors. Let \( A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\} \) be a quantized \( B_d \) sequence with respect to \( \eta \). Form an \( m \times n \) matrix \( C \) where \( m = m_b \) and \( n = Kn_b \) by concatenating \( K \) matrices \( C_i = \alpha_i C_b, 1 \leq i \leq K \), horizontally. The constructed matrix is a \([q; Q; \eta; (1:d); e]\)-SQ-separable matrix with \( q = \alpha_K + 1 \).

**Proof.** In order to show that the constructed matrix is \([q; Q; \eta; (1:d); e]\)-SQ-separable, we consider two distinct sets of columns \( X_1 \) and \( X_2 \) that satisfy \( 1 \leq |X_1|, |X_2| \leq d \). The idea is to show that the syndrome of these two sets contain at least \( 2e + 1 \) different entries. Without loss of generality, we assume that \( |X_1| \leq |X_2| \). Since the two sets are distinct, one must have \( X_2 \setminus X_1 \neq \emptyset \), and therefore we choose \( z_r \in X_2 \setminus X_1 \). By construction, \( z_r = \alpha_r z_b \) for some binary column \( z_b \) in \( C_b \) and some \( \alpha_r \in A \).

For the fixed binary column \( z_b \), let \( Z \) be the set of columns of \( C \) generated by multiplying \( z_b \) with the elements of \( A \). Let \( Z_1 = X_1 \cap Z \) and \( Z_2 = X_2 \cap Z \) be the set of columns with the same support as \( z_b \) in \( X_1 \) and \( X_2 \), respectively. Also, let \( A_{Z_1} \subseteq A \) and \( A_{Z_2} \subseteq A \) be the set of coefficients used to form the
columns in $Z_1$ and $Z_2$, respectively. Given that $A$ is a quantized $B_d$ sequence, we have to separately consider two different scenarios.

**Case 1:** $\sum_{\alpha_i \in A_{Z_2}} \alpha_i > \eta \sum_{\alpha_i \in A_{Z_1}} \alpha_i$.

By construction of $C$ and Def. 3, there exists a set of coordinates $R_r$ of size at least $2e + 1$, such that $\forall k \in R_r$,

$$\begin{cases}
z_r(k) = \alpha_r, \\
x(k) = 0 & \forall x \in X_1 \setminus Z_1, 
\end{cases}$$

Consequently, $\forall k \in R_r$ we have the following sequence of inequalities:

$$y_{x_2}(k) \geq y_{z_2}(k)$$

$$y_{z_2}(k) > y_{z_1}(k)$$

$$y_{z_1}(k) = y_{x_1}(k)$$

where (4.1) follows since $Z_2 \subseteq X_2$, (4.2) follows since $\sum_{\alpha_i \in A_{Z_2}} \alpha_i > \eta \sum_{\alpha_i \in A_{Z_1}} \alpha_i$, and (4.3) follows since $x(k) = 0, \forall x \in X_1 \setminus Z_1$.

**Case 2:** $\sum_{\alpha_i \in A_{Z_1}} \alpha_i > \eta \sum_{\alpha_i \in A_{Z_2}} \alpha_i$.

In this case, we cannot use the set of coordinates $R_r$, since (4.2) no longer holds. On the other hand, this case happens only if $A_{Z_1} \setminus A_{Z_2} \neq \emptyset$. Consequently, one has $Z_1 \setminus Z_2 \neq \emptyset$; let $z_s \in Z_1 \setminus Z_2$, where $z_s = \alpha_s z_b$ for some $\alpha_s \in A_{Z_1}$. Similar to case 1, by considering $X_2$ instead of $X_1$, there exists a set of coordinates $R_s$ of size at least $2e + 1$, such that $\forall k \in R_s$,

$$\begin{cases}
z_s(k) = \alpha_s, \\
x(k) = 0 & \forall x \in X_2 \setminus Z_2, 
\end{cases}$$

As a result, the following inequalities hold:

$$y_{x_1}(k) \geq y_{z_1}(k)$$

$$y_{z_1}(k) > y_{z_2}(k)$$

$$y_{z_2}(k) = y_{x_2}(k)$$

where (4.4) follows since $Z_1 \subseteq X_1$, (4.5) follows since $\sum_{\alpha_i \in A_{Z_1}} \alpha_i > \eta \sum_{\alpha_i \in A_{Z_2}} \alpha_i$, and (4.6) follows since $x(k) = 0, \forall x \in X_2 \setminus Z_2$. Note that even though $|X_1| \leq |X_2|$, unlike for Case 1, we have $y_{x_1}(k) > y_{x_2}(k)$ for all $k \in R_s$. 

□
4.2.1 Fundamental limits and constructions of quantized $B_h$ sequences

Quantized $B_h$ sequences ensure that a set of integers and their subset sums are placed into different quantization bins. As a result, for a fixed set of $Q$ thresholds $\eta$, the existence of quantized $B_h$ sequences with a predetermined cardinality $K$ depends on the thresholds. As mentioned in Remark 6, the cardinality of a quantized $B_h$ sequence may be smaller than $h$. For example, one can always choose $A = \{\eta_1\}$ as a quantized $B_h$ sequence with $K = 1$. For the case of $K = 2$, the sequence $A = \{\eta_1, \max\{\eta_2, \eta_3 - \eta_1\}\}$ used in Thm. 3 is a quantized $B_h$ sequence with respect to $\eta$ as long as $Q \geq 4$ and $\eta_Q > \eta_1 + \max\{\eta_2, \eta_3 - \eta_1\}$. These two examples imply that for any set of thresholds, there always exists a quantized $B_h$ sequence, which in the worst case scenario has cardinality $K = 1$.

Next, we discuss constructions of quantized $B_h$ sequences with $K > 2$. From a practical perspective, and given that in most applications $q$ cannot be too large, a greedy algorithm for finding a quantized $B_h$ sequence is the simplest constructive approach. In the greedy approach, one starts with $\alpha_1 = \eta_1$; then, given the first $i$ elements of the sequence, to find $\alpha_{i+1}$, one increases the value of $\alpha_i$ until the properties of the quantized $B_h$ sequence are satisfied.

Although this method works for small values of $K$, for large values of $K$ this procedure has a high computational complexity. Alternatively, one can use standard subset-sum distinct sequences\(^2\) [48], and generalizations of standard $B_h$ sequences to construct a family of quantized $B_h$ sequences as described in the next theorem.

**Theorem 5.** Consider a SQGT model with thresholds $\eta = [0, \eta_1, \eta_2, \ldots, \eta_Q]^T$; $\forall s: 1 \leq s \leq Q$, and let $g_s = \max_{1 \leq i \leq s} \eta_i - \eta_{i-1}$ be the largest gap of the first $s$ thresholds. Let $B = \{\beta_1 < \beta_2 < \ldots\}$ be a sequence for which all the subset sums of at most $h$ elements are distinct. For a fixed $s$, $2 \leq s \leq Q$, let $K_s$ be a positive integer small enough to satisfy $\eta_s > g_s \sum_{i=\max(1,K_s-h)}^{K_s} \beta_i$. Then all the sequences of the form $A_s = \{g_s \beta_1, g_s \beta_2, \ldots, g_s \beta_{K_s}\}$ are quantized $B_h$ sequences with respect to $\eta$.

**Proof.** Fix a value of $s: 1 \leq s \leq Q$; consider any two distinct sets $A_1, A_2 \subseteq A_s$,\(^2\)

\(^2\)A subset-sum distinct sequence is a sequence of positive integers such that the sum of the elements of its subsets are distinct.
\[ |A_1| \leq h \text{ and } A_2 \leq h, \] which are obtained by multiplying the elements of \( B_1 \subseteq B \) and \( B_2 \subseteq B \) with \( g_s \), respectively. Suppose \( f_\eta(\sum_{\alpha_i \in A_1} \alpha_i) = f_\eta(\sum_{\alpha_i \in A_2} \alpha_i) \); as a result, there exists \( r, 1 \leq r \leq s \), such that \( \eta_{r-1} \leq \sum_{\alpha_i \in A_1} \alpha_i < \eta_r \) and \( \eta_{r-1} \leq \sum_{\alpha_i \in A_2} \alpha_i < \eta_r \). Consequently,

\[
\left| \sum_{\alpha_i \in A_1} \alpha_i - \sum_{\alpha_i \in A_2} \alpha_i \right| \leq \eta_r - \eta_{r-1} - 1 < g_s. \quad (4.7)
\]

However, since all the sums of up to \( h \) elements of \( B \) are distinct and \( |B_1| \leq h \) and \( |B_2| \leq h \), \( \left| \sum_{\beta_i \in B_1} \beta_i - \sum_{\beta_i \in B_2} \beta_i \right| \geq 1 \). Consequently,

\[
\left| \sum_{\alpha_i \in A_1} \alpha_i - \sum_{\alpha_i \in A_2} \alpha_i \right| = g_s \left| \sum_{\beta_i \in B_1} \beta_i - \sum_{\beta_i \in B_2} \beta_i \right| \geq g_s, \quad (4.8)
\]

which contradicts (4.7).

Given this theorem, one can construct quantized \( B_h \) sequences using the sequences mentioned in the theorem or the more strict subset-sum distinct sequences, for which many constructions are known in the literature [48, 51, 52]. One should note that for a fixed value of \( K \), the construction of quantized \( B_h \) sequences described in this theorem may not generate the densest sequence; however, this construction has the important property that it applies to any set of thresholds and only depends on a condition that can be easily verified given the thresholds.

**Remark 7.** All the subset-sums consisting of at most \( h \) elements of a quantized \( B_h \) sequence must fall into different quantization bins; since there are \( Q \) such bins, the following bounds on the number of elements of a quantized \( B_h \) sequence hold: Let \( A \) be a finite quantized \( B_h \) sequence with respect to \( \eta \) such that \( |A| = K \). If \( K \leq h \), then \( K \leq \log_2 Q \). On the other hand, if \( K > h \), then \( \sum_{i=0}^{h} \binom{K}{i} \leq Q \).

**Remark 8.** Let \( B \) be a subset-sum distinct sequence (i.e. a sequence such that all its subsets sum up to distinct values). Assume that a positive integer \( K \) satisfies the condition in Thm. 5; then, this theorem can be used to construct a quantized \( B_h \) sequence \( A \), \( |A| = K \), using \( B = \{\beta_1, \beta_2, \ldots, \beta_K\} \). There exists a large body of literature describing constructive bounds on \( \beta_K \) [51], [52]. All bounds are of the form \( \beta_K \leq c2^K \), where \( c < 1 \) is a constant that depends on
the construction (e.g. \(c = 0.22002\) in [52]). Given a bound of this form, one has \(\alpha_K < cg2^K\), where \(g\) is the largest gap for the first \(K\) thresholds. The aforementioned bound is exponential in \(K\), where the base of the exponential equals 2. In Lemma 2, we will prove an upper bound on \(\alpha_K\) in which the base of the exponential function is strictly smaller than 2. Although this bound applies to SQLO\(_s\) sequences, given that any SQLO\(_s\) sequence is also a quantized \(B_h\) sequence, it can be considered an upper bound for quantized \(B_h\) sequences as well. This implies that the bound in Lemma 2 is asymptotically tighter than the aforementioned bound.

### 4.2.2 A decoding algorithm for SQGT matrices constructed using quantized \(B_h\) sequences

We describe next a decoding algorithm for matrices constructed using Theorem 4. Let \(D\) denote the set of columns of \(C\) corresponding to the defectives. Also, let \(X_D\) be the set of binary columns each corresponding to the support of at least one column in \(D\); clearly, \(|X_D| \leq |D| \leq d\). As an example, suppose that in a SQGT system \(D = \{[2,0,2,2]^T, [6,0,6,6]^T, [2,0,2,0]^T\}\); in this case one has \(X_D = \{[1,0,1,1]^T, [1,0,1,0]^T\}\).

The decoding procedure is performed in three steps. The idea is to use the disjunctness property of binary disjunct matrices and the property of quantized \(B_h\) sequences to first recover the set \(X_D\) in Step 1 and then use this set to recover \(D\) in Steps 2 and 3. The steps of the decoding algorithm are listed in Algorithm 1.

**Theorem 6.** Algorithm Dec-QBh is capable of identifying up to \(d\) defectives in the presence of at most \(e\) errors in the vector of test results \(y\).

**Proof.** In the first step of the algorithm, and for each column of the binary matrix \(C_b\), we count the number of coordinates for which the test result is smaller than the corresponding entry of the column. In order to show that the set \(X\) recovered in Step 1 is equal to \(X_D\), we first show that \(X \supseteq X_D\). Each column in \(D\) can be written as \(z_i = \alpha x_i, 1 \leq i \leq |D|\), for some \(\alpha \in A\) and some binary column \(x_i\) in \(X_D\). We need to show that if \(x_i \in X_D\), then \(x_i \in X\), or equivalently, the number of coordinates \(j\) for which

\[
x_i(j) \leq y(j)\tag{4.9}
\]
Algorithm 1: Dec-QBh

Input: $y \in [Q]^m$, $C_b \in [2]^{m \times K}$, $\eta$, $A$, $e \geq 0$
Output: $\hat{D}$

Step 1: Initialize $X \leftarrow \emptyset$ and $\hat{D} \leftarrow \emptyset$

For $i = 1, 2, \ldots, \frac{m}{K}$ do

If the number of coordinates $j$ for which the $i$-th column of $C_b$ does not satisfy $x_i(j) \leq y(j)$ is at most equal to $e$, set $X \leftarrow X \cup \{x_i\}$.

End

End

Step 2:

Form $B$ the ordered list of the distinct sums of elements of subsets of $A$ with cardinality at most $d$ and their corresponding subsets.

Step 3:

Form $u_p$ such that $u_p(j)$ is the upper threshold of the quantization bin in which $y(j)$ lies.

For $i = 1, 2, \ldots, |X|$ do

Find $\beta_i$, the largest element of $B$ such that the number of coordinates $j$ for which $\beta_i x_i(j) < u_p(j)$ is not satisfied is at most $e$.

Let $A_{i,l} \subseteq A$ be the set with the sum equal to $\beta_i$.

Set $\hat{D}_i \leftarrow \{\text{columns of } C \text{ of the form } z = \alpha x_i, \ \forall \alpha \in A_{i,l}\}$.

End

Return $\hat{D} = \bigcup_i \hat{D}_i$
is not satisfied is at most \( e \). All the entries of \( y \) which are not erroneous are equal to the corresponding entries of the syndrome of defectives \( y_p \). As a result, (4.9) is trivially satisfied for entries of \( x \) that are equal to zero, since for these entries \( y_p \) is equal to zero and an error can only increase the corresponding coordinate in \( y \). On the other hand, since \( A \) is a quantized \( B_d \) sequence, its smallest element satisfies \( \alpha_1 \geq \eta_1 \). Consequently, a nonzero entry of \( x \) results in a nonzero entry in \( y_p \), which is a nonzero entry in \( y \) unless an error occurs; since the nonzero entries of \( x \) are equal to 1 (the smallest positive integer) and there are at most \( e \) errors, condition (4.9) is satisfied for all except up to \( e \) nonzero entries. Consequently, \( X \supseteq X_D \).

Next, we show that if \( x_i \in X \), then \( x_i \in X_D \), or equivalently \( X \subseteq X_D \). Suppose this is not true and let \( x \in X \setminus X_D \). Since \( C_b \) is a binary disjunct matrix and \( |X_D| \leq d \), then there exists a set of coordinates \( R \) such that \( |R| \geq 2e+1 \) and \( \forall j \in R \) one has \( x(j) = 1 \) while \( x_i(j) = 0 \), \( \forall x_i \in X_D \). Consequently, \( \forall j \in R \), one has \( y_p(j) = 0 \), which implies that \( y(j) = 0 \) unless an error occurred. Since there are at most \( e \) errors, \( x(j) > y(j) \) for at least \( e+1 \) coordinates, which implies that \( x \notin X \). This contradicts the starting assumption. Hence, \( X \subseteq X_D \).

Now given that Step 1 recovered the set \( X = X_D \), we only need to show that Step 3 recovers \( D \) given \( X_D \). For each \( x_i \in X_D \), let \( A_{i,t} \) be the “true” set of coefficients used to generate the columns in \( D \) with the same support as \( x_i \). Also, let \( \beta_l = \sum_{\alpha \in A_{i,t}} \alpha \) be the sum of these coefficients. Since the error-free entries of \( y \) are equal to \( y_p \), then for all \( 1 \leq j \leq m \), one has \( \beta_l x_i(j) < u_D(j) \) unless an error occurred in the \( j \)-th coordinate. Since there are at most \( e \) errors, there are at most \( e \) coordinates for which this condition is not satisfied. As a result, \( \beta_l \geq \beta_t \).

In order to complete the proof, we show that no value of \( \beta' \in B \) such that \( \beta' > \beta_t \) satisfies the condition in Step 3 and hence conclude that \( \beta_l \leq \beta_t \). From the disjunctness property of \( C_b \), there exists a set of coordinates \( R_i \) such that \( |R_i| \geq 2e+1 \) and \( \forall j \in R_i \), \( x_i(j) = 1 \), while all other columns in \( X_D \) have the value zero at that coordinate. As a result, \( \forall j \in R_i \),

\[
\sum_{z \in D} z(j) = \sum_{\alpha \in A_{i,t}} \alpha x_i(j) = \sum_{\alpha \in A_{i,t}} \alpha = \beta_t.
\]

Since there are at most \( e \) errors in \( y \), there exists a set of coordinates \( R'_i \subseteq R_i \)
with $|\mathcal{R}_i'| \geq e + 1$, such that $\forall j \in \mathcal{R}_i'$,

$$ u_D(j) > \sum_{z \in D} z(j) = \beta_t. $$

Consider $\beta' \in \mathcal{B}$ such that $\beta' > \beta_t$. Since $\mathcal{A}$ is a quantized $B_d$ sequence, $\beta' > \eta \beta_t$ implies that $\forall j \in \mathcal{R}_i'$ one has $\beta' \geq u_D(j) > \beta_t$. Given $|\mathcal{R}_i'| \geq e + 1$, the condition in Step 3 is not satisfied for such a choice of $\beta'$ and hence $\beta_t \geq \beta_t$. As a result, Step 3 uniquely recovers $\beta_t = \beta_t$ which corresponds to the set $\mathcal{A}_{i,t}$. Consequently, $\hat{D} = D$ in the presence of up to $e$ errors in the vector of test results, as claimed. \qed

**Remark 9.** The computational complexity of Algorithm 1 is equal to $O(\frac{mn}{K} + 2^K(K + md))$. The computational complexity of Step 1 is $O(\frac{mn}{K})$. The second step requires $\sum_{i=1}^{\min(d,K)} \binom{K}{i}(i - 1) = O(K2^K)$ summations. Finally, the computational complexity of Step 3 is $O(dm2^K)$.

Due to the exponential growth of the computational complexity of the decoding algorithm with $K$, the matrices constructed using quantized $B_h$ sequences are most suitable for small values of $K$. On the other hand, for larger values of $K$, we introduce two other families of sequences that lead to matrices with significantly smaller decoding complexity.

### 4.3 SQ-separable matrices using SQLO$_s$ sequences

In this section, we introduce the notion of semi-quantitative lexicographical order and corresponding sequences, and describe how to construct SQ-separable matrices using these sequences. Then, we provide a decoding algorithm for such matrices and discuss how they compare to the matrices constructed using quantized $B_h$ sequences.

As discussed earlier, the matrices constructed using quantized $B_h$ sequences have a decoding algorithm with computational complexity $O(\frac{mn}{K} + 2^K(K + md))$. Although for small values of $K$ the dominant term is $\frac{mn}{K}$, for large values of $K$ the exponential growth of the complexity with $K$ is problematic. In this section we introduce the notion of SQLO$_s$ sequences and use them to construct SQ-separable matrices with a decoding algorithm that has computational complexity linear in $K$. 

33
**Definition 12 (SQLO$(_s\eta,h)$ sequences).** Given a set of thresholds $\eta$, a sequence of positive integers $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ is termed a SQLO$(_s\eta,h)$ sequence if

1. $\alpha_K >_\eta \alpha_{K-1} >_\eta \ldots >_\eta \alpha_2 >_\eta \alpha_1 >_\eta 0$ (i.e., all elements of $A$ lie in different quantization bins).

2. For any two distinct nonempty nested subsets $A_1 \subset A_2 \subseteq A$ such that $|A_1| \leq h$ and $|A_2| \leq h$, one has $\sum_{i \in A_2} \alpha_i >_\eta \sum_{i \in A_1} \alpha_i$ (i.e., the sums of elements of nested subsets fall into different quantization bins).

3. For any two distinct nonempty subsets $A_1, A_2 \subseteq A$ that are not nested and $|A_1| \leq h$ and $|A_2| \leq h$, one has $\sum_{i \in A_2} \alpha_i >_\eta \sum_{i \in A_1} \alpha_i$ whenever $\exists \alpha \in A_2 \setminus A_1$ such that $\alpha >_\eta \alpha'$, $\forall \alpha' \in A_1 \setminus A_2$ (i.e., two subsets that are not nested are ordered based on their largest distinct element).

The properties above induce a partial order on the subsets of a SQLO$_s$ sequence.

The SQLO$_s$ properties for $K = 2$ and $h \geq 2$ simply become $\alpha_2 + \alpha_1 >_\eta \alpha_2 >_\eta \alpha_1 >_\eta 0$, while for $K = 3$ and $h \geq 3$ it becomes $\alpha_3 + \alpha_2 + \alpha_1 >_\eta \alpha_3 + \alpha_2 >_\eta \alpha_3 + \alpha_1 >_\eta \alpha_2 + \alpha_1 >_\eta \alpha_2 >_\eta \alpha_1 >_\eta 0$. As an example, it can be easily verified that $A = \{3, 6, 12\}$ is a SQLO$_s$ sequence with respect to the thresholds $\eta = [0, 3, 6, \ldots, 24]^T$, since $f_\eta(12+6+3) = 7 > f_\eta(12+6) = 6 > f_\eta(12+3) = 5 > f_\eta(12) = 4 > f_\eta(6+3) = 3 > f_\eta(6) = 2 > f_\eta(3) = 1 > 0$.

SQLO$_s$ sequences obey a more stringent set of constraints compared to the quantized $B_h$ sequences. As a result, one is able to use these constraints to reduce the computational complexity of the decoding. In the next proposition, we show that in fact any SQLO$_s$ sequence is also a quantized $B_h$ sequence, but the converse is not necessarily true.

**Proposition 4.** A sequence of $K$ positive integers $A$ is a SQLO$(_s\eta,h)$ sequence if and only if both of the following properties are satisfied:

1. $A$ is a quantized $B_h$ sequence.

2. $\forall i : 1 \leq i \leq K$ and $\forall A' \subseteq \{\alpha_1, \alpha_2, \ldots, \alpha_{i-1}\}$ such that $|A'| \leq h$, one has $\alpha_i >_\eta \sum_{j \in A'} \alpha_j$.

**Proof.** First, we show that if $A$ is a SQLO$_s(\eta,h)$ sequence, it satisfies properties 1 and 2. It is easy to see that since $A$ is a SQLO$_s(\eta,h)$ sequence, it
satisfies the first property of quantized $B_h$ sequences, i.e. $\alpha_K \gg \eta \alpha_{K-1} \gg \eta \cdots \gg \eta \alpha_1 \gg \eta 0$.

Let $A_1$ and $A_2$ be two arbitrary nonempty distinct subsets of $A$ such that $|A_1| \leq h$ and $|A_2| \leq h$. We need to show that $\sum_{\alpha_i \in A_1} \alpha_i \gg \eta \sum_{\alpha_i \in A_2} \alpha_i$ or $\sum_{\alpha_i \in A_2} \alpha_i \gg \eta \sum_{\alpha_i \in A_1} \alpha_i$. If these two subsets are nested, i.e. if $A_1 \subset A_2$ or $A_2 \subset A_1$, from the second property of a SQLO$_s$ sequence, it follows that $\sum_{\alpha_i \in A_2} \alpha_i \gg \eta \sum_{\alpha_i \in A_1} \alpha_i$ or $\sum_{\alpha_i \in A_1} \alpha_i \gg \eta \sum_{\alpha_i \in A_2} \alpha_i$, respectively. Otherwise, the third property of a SQLO$_s(\eta, h)$ sequence ensures that $A$ is a quantized $B_h$ sequence. On the other hand, from the third property of a SQLO$_s(\eta, h)$ sequence, one can directly conclude that the second property of the proposition holds.

Now we show that if $A$ satisfies the two properties stated in the proposition, then it is a SQLO$_s(\eta, h)$ sequence. Since $A$ is a quantized $B_h$ sequence, the first property of a SQLO$_s(\eta, h)$ sequence is automatically satisfied.

Next, consider two distinct nonempty nested subsets $A_1 \subset A_2 \subset A$ such that $|A_1| \leq h$ and $|A_2| \leq h$. Since $\sum_{\alpha_i \in A_2} \alpha_i$ and $\sum_{\alpha_i \in A_1} \alpha_i$ fall into different quantization bins, due to the second property of a quantized $B_h$ sequence, and since $\sum_{\alpha_i \in A_2} \alpha_i > \sum_{\alpha_i \in A_1} \alpha_i$, one has $\sum_{\alpha_i \in A_2} \alpha_i \gg \eta \sum_{\alpha_i \in A_1} \alpha_i$. Now consider two distinct nonempty subsets $A_1, A_2 \subset A$ that are not nested, such that $|A_1| \leq h$ and $|A_2| \leq h$. Assume that $\exists \alpha \in A_2 \setminus A_1$ such that $\alpha \gg \eta \alpha'$, $\forall \alpha' \in A_1 \setminus A_2$. In this case, it holds that

$$\sum_{\alpha_i \in A_2} \alpha_i \gg \eta \sum_{\alpha_i \in A_1} \alpha_i,$$

where the last inequality follows from the second property of the proposition. This completes the proof of the proposition.

As a result of the first condition in Proposition 4, one can directly use a SQLO$_s(\eta, h)$ sequence instead of a quantized $B_h$ sequence to construct SQ-separable matrices, as formally stated in the next theorem. In addition, the second property in Proposition 4 allows us to reduce the computational complexity of the decoding significantly. This is a consequence of the fact that superincreasing sequences$^3$ are knapsack-solvable in linear time [53]. In other words, given an integer and a finite superincreasing sequence, it is

---

$^3$A superincreasing sequence is a sequence of positive integers such that each element of the sequence is at least as large as the sum of all the elements preceding it.
possible to determine in linear time whether the integer can be expressed as a sum of distinct elements of the sequence, and if so to identify these elements [53].

**Theorem 7 (Construction 4).** Fix a binary \(d\)-disjunct matrix \(C_b\) of dimensions \(m_b \times n_b\), capable of correcting up to \(e\) errors. Let \(A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}\) be a SQLO\(_s\)(\(\eta, d\)) sequence. Form a matrix \(C\) of length \(m = m_b\) and size \(n = Kn_b\) by concatenating \(K\) matrices \(C_i = \alpha_iC_b\), \(1 \leq i \leq K\), horizontally. The constructed matrix is a \([q; Q; \eta; (1:d); e]\)-SQ-separable matrix with \(q = \alpha_K + 1\).

**Proof.** The proof directly follows from Proposition 4 and Thm. 4. Since any SQLO\(_s\)(\(\eta, d\)) sequence is a quantized \(B_d\) sequence, Thm. 4 implies that the matrix \(C\) is a \([q; Q; \eta; (1:d); e]\)-SQ-separable matrix with \(q = \alpha_K + 1\). \(\square\)

### 4.3.1 Fundamental limits and constructions of SQLO\(_s\) sequences

Next, we discuss construction methods and fundamental density limits for SQLO\(_s\) sequences. Given a set of thresholds, a simple greedy algorithm can be used to find a SQLO\(_s\) sequence by checking the properties in Def. 12. For example, assuming that \(\eta = [0, 2, 5, 6, 10, 13, 15, 16, 18, 21]^T\) and that \(h \geq K = 3\), the greedy algorithm produces \(A = \{2, 5, 11\}\). Alternatively, one can use the following theorem to construct SQLO\(_s\) sequences using super-increasing sequences.

**Definition 13.** A sequence of positive integers \(B = \{\beta_1, \beta_2, \ldots\}\) is called \(h\)-superincreasing if \(\forall j > 1, \beta_j > \sum_{i=\max\{1,j-h\}}^{j-1} \beta_i\).

**Theorem 8.** Consider a SQGT system with thresholds \(\eta = [0, \eta_1, \eta_2, \ldots, \eta_Q]^T\); \(\forall s : 1 \leq s \leq Q\), let \(g_s = \max_{1 \leq i \leq s} \eta_i - \eta_{i-1}\) be the largest gap of the first \(s\) thresholds. Let \(B = \{\beta_1 < \beta_2 < \ldots\}\) be a \(h\)-superincreasing sequence. For a fixed \(s\), \(2 \leq s \leq Q\), let \(K_s\) be a positive integer small enough to satisfy \(\eta_s > g_s \sum_{i=\max\{1,K_s-h\}}^{K_s} \beta_i\). Then all the sequences of the form \(A_s = \{g_s \beta_1, g_s \beta_2, \ldots, g_s \beta_{K_s}\}\) are SQLO\(_s\)(\(\eta, h\)) sequences.

**Proof.** Let \(A_s\) be a fixed sequence satisfying the conditions of the theorem. First, we show that \(A_s\) is a quantized \(B_h\) sequence. Fix a value of \(s : 1 \leq s \leq Q\). Consider any two distinct sets \(A_1, A_2 \subseteq A_s\), \(|A_1| \leq h\) and \(|A_2| \leq h\),
which are obtained by multiplying the elements of \( B_1 \subseteq B \) and \( B_2 \subseteq B \) with \( g_s \), respectively. Suppose that \( f_\eta\left( \sum_{\alpha_i \in A_1} \alpha_i \right) = f_\eta\left( \sum_{\alpha_i \in A_2} \alpha_i \right) \); as a result, there exists an integer \( r, 1 \leq r \leq s \), such that \( \eta_{r-1} \leq \sum_{\alpha_i \in A_1} \alpha_i < \eta_r \) and \( \eta_{r-1} \leq \sum_{\alpha_i \in A_2} \alpha_i < \eta_r \). Consequently,

\[
\left| \sum_{\alpha_i \in A_1} \alpha_i - \sum_{\alpha_i \in A_2} \alpha_i \right| \leq \eta_r - \eta_{r-1} - 1 < g_s. \tag{4.10}
\]

Since \( B_1 \neq B_2 \), the set \((B_1 \cup B_2) \setminus (B_1 \cap B_2)\) is nonempty. Let \( \beta_i \) be the largest element of this set, and without loss of generality assume that \( \beta_i \in B_1 \). Since \( B \) is a \( h \)-superincreasing sequence and \( |B_1| \leq h \) and \( |B_2| \leq h \), one has

\[
\beta_i > \sum_{\beta_j \in B_2} \beta_j. \tag{4.11}
\]

This implies that \( \sum_{\beta_j \in B_1} \beta_j > \sum_{\beta_j \in B_2} \beta_j \), or equivalently, that

\[
\left| \sum_{\beta_j \in B_1} \beta_j - \sum_{\beta_j \in B_2} \beta_j \right| \geq 1.
\]

Given this result, one can use \( h \)-superincreasing sequences to construct \( \text{SQLO}_s(\eta, h) \) sequences. For example, the sequence \( B = \{1, 2, 2^2, 2^3, \ldots\} \) is a superincreasing sequence, hence an \( h \)-superincreasing sequence for any value of \( h \), and can be used to construct \( \text{SQLO}_s(\eta, h) \) sequences. Given this sequence, one obtains a \( \text{SQLO}_s(\eta, h) \) sequence such that \( \alpha_K = g_s 2^{K-1} \). Never-
theless, a simple construction based on recursive equations results in a better upper bound on the smallest value for \( \alpha_K \), as described in Lemma 2.

We next state a theorem by Ostrovsky [54, Thm. 1.1.4] which we will use in the proof of Lemma 2. The proof of this result can be found in [54, P. 3].

**Lemma 1.** Let \( P(x) = x^n - a_1x^{n-1} - \cdots - a_n \), where all the coefficients \( a_i \), \( 1 \leq i \leq n \), are non-negative, and at least one is nonzero. If the greatest common divisor of the indices of the positive coefficients equals 1, then the polynomial \( P(x) \) has a unique positive root \( r \); in addition, for any other root of this polynomial denoted by \( r' \), one has \( |r'| < r \).

Given this lemma, we will prove the following result.

**Lemma 2.** Let \( \gamma \) be the largest positive real root of the polynomial \( g(x) = x^{h+1} - 2x^h + 1 \). Also, assume that a positive integer \( K \) satisfies the condition in Thm. 8; then one can construct a SQLO\(_s(\eta, h)\) sequence such that \( \alpha_K = O_g(\gamma^K) \), where \( g \) is the largest gap for the first \( K \) thresholds, and \( \gamma < 2 \).

**Proof.** We construct a sequence \( B \) as follows. First, \( \forall 1 \leq i \leq h \), we set \( \beta_i = 2^i - 1 \). Then, for \( i > h \), we let \( \beta_i = \beta_{i-1} + \beta_{i-2} + \cdots + \beta_{i-h} + 1 \). Clearly, this sequence is a \( h \)-superincreasing sequence. The characteristic equation of this recurrence is of the form \( f(x) = x^h - x^{h-1} - \cdots - x - 1 = 0 \), which satisfies the condition of Lemma 1. In addition, the greatest common divisor of the indices of the positive coefficients is 1, since all these coefficients are equal to 1. Consequently, Lemma 1 implies that this equation has a unique real positive root, \( \gamma \), and that the absolute values of all the other roots are strictly smaller than \( \gamma \). Consequently, \( \beta_K = O(\gamma^K) \). Simplifying this equation by multiplying both sides by \( (x - 1) \), the equation becomes \( x^{h+1} - 2x^h + 1 = 0 \). Consequently, one has \( \alpha_K = O_g(\gamma^K) \), where \( \gamma \) is the largest positive real root of \( g(x) \).

Next, we show that \( \gamma < 2 \). Evaluating \( g(x) = x^{h+1} - 2x^h + 1 \) on the real axis reveals that this function has two local optima at \( x = 0 \) and \( x = \frac{2h}{h+1} \), and is monotonically increasing for \( x > \frac{2h}{h+1} \). On the other hand, \( g(2) > 0 \); in addition, for all \( h \geq 1 \), one has \( 2 > \frac{2h}{h+1} \); consequently, \( \forall x > 2, f(x) > f(2) > 0 \). As a result, the largest positive real solution to \( g(x) = 0 \) is strictly smaller than 2 for any finite value of \( h \), i.e. \( \gamma < 2 \). \( \square \)
4.3.2 A decoding algorithm for SQGT matrices constructed using SQLO\(_s\) sequences

We next describe the Dec-SQLO\(_s\) algorithm, the decoding procedure for matrices based on SQLO\(_s\) sequences. This algorithm comprises of two steps. The first step is identical to the first step of Algorithm 1. However, Steps 2 and 3 in Algorithm 1 are replaced by a single step which has a significantly lower computational complexity than steps 2 and 3. The steps of Dec-SQLO\(_s\) are listed in Algorithm 2. The first step identifies the set \(X_D\). Given this set, Step 2 identifies the set of defectives \(D\). In order to show that the second step can identify up to \(d\) defectives in the presence of up to \(e\) errors, we state the following lemma and proposition which we find useful for our subsequent proofs.

**Lemma 3.** Consider a SQ-separable matrix constructed using Thm. 7 and let \(y\) be the vector of test results with at most \(e\) erroneous entries. Fix any binary column \(x_i \in X_D\), and let \(S_i \) be the set of nonzero coordinates of \(x_i\). Also, let \(S_i' \subseteq S_i\), with \(|S_i'| = 2e + 1\) be the set of coordinates such that for any fixed \(k \in S_i'\), one has \(y(k) \leq y(j)\) \(\forall j \in S \setminus S_i'\). Then, there exists a set \(S_i'' \subseteq S_i'\) such that \(|S_i''| \geq e + 1\), and \(\forall j \in S_i''\), one has \(y(j) = y_D(j) = f_\eta\left(\sum_{\alpha \in A_i,t} \alpha\right)\); in this equation, \(A_i,t \subseteq A\) denotes the set of coefficients corresponding to the defective columns with the same support as \(x_i\).\(^4\)

**Proof.** Let \(R_i\) be the maximal set of coordinates such that \(\forall j \in R_i, x_i(j) = 1\) and \(x(j) = 0 \) for all \(x \in X_D \setminus \{x_i\}\). Since \(x_i\) is a column of \(C_b\) and since \(|X_D| \leq d\), the disjunctness property implies that such a set exists and that \(|R_i| \geq 2e + 1\); clearly, \(R_i \subseteq S_i\). Let \(A_i,t\) be the set of coefficients used to generate the columns in \(D\) with the same support as \(x_i\). For all \(k \in R_i\), one has \(\sum_{z \in D} z(k) = \sum_{\alpha \in A_i,t} \alpha\), and \(\forall j \in S_i \setminus R_i\), one has \(\sum_{z \in D} z(j) > \sum_{\alpha \in A_i,t} \alpha\). Note that the strict inequality follows since \(R_i\) is a maximal set. Since all the sums of up to \(d\) elements of \(A\) fall into different quantization bins, for any \(k \in R_i\) and for any \(j \in S_i \setminus R_i\), one has

\[
\eta\left(\sum_{z \in D} z(k)\right) < \eta\left(\sum_{z \in D} z(j)\right).
\]

\(^4\)As an example, assume that \(x_i \in X_D\) and let \(\{\alpha_j,x_i,\alpha_j,x_i,\alpha_j,x_i\} \in D\) be the only columns in \(C\) with the same support as \(x_i\) in \(D\). In this case, \(A_i,t = \{\alpha_{j1},\alpha_{j2},\alpha_{j3}\}\).
Algorithm 2: Dec-SQLO$_s$

Input: $y \in [Q]^m$, $C_b \in [2]^{m \times n}$, $\eta$, $A$, $e \geq 0$
Output: $\hat{D}$

Step 1: Initialize $\mathcal{X} \leftarrow \emptyset$ and $\hat{D} \leftarrow \emptyset$
For $i = 1, 2, \ldots, \frac{n}{K}$ do
    $x_i \leftarrow$ the $i$-th column of $C_b$
    $N_i \leftarrow$ number of coordinates $j$ for which $x_i(j) > y(j)$
    If $N_i \leq e$ then
        Set $\mathcal{X} \leftarrow \mathcal{X} \cup \{x_i\}$
    End
End

Step 2:
For $i = 1, 2, \ldots, |\mathcal{X}|$ do
    Set $S_i \leftarrow \{\text{the set of nonzero coordinates of } x_i\}$
    Set $S'_i \leftarrow \{\text{subset of } S_i \text{ with } |S'_i| = 2e + 1 \text{ s.t. } \forall k \in S'_i \text{ and } \forall j \in S_i \setminus S'_i, \text{ one has } y(k) \leq y(j)\}$
    Initialize the multiset $B' \leftarrow \emptyset$
    For $j = 1, 2, \ldots, |S'_i|$ do
        $\eta_u \leftarrow$ the upper threshold of the quantization bin of $y(j)$
        $\eta_l \leftarrow$ the lower threshold of the quantization bin of $y(j)$
        $\beta \leftarrow$ the integer $\eta_l \leq \beta < \eta_u$ that can be written as the sum of up to $d$ elements of $A$
        (use Proposition 5)
        Update the multiset $B' \leftarrow B' \cup \{\beta\}$
    End
    Set $\hat{\beta}_i \leftarrow$ the element of $B'$ with at least $e + 1$ repetitions
    Set $\hat{A}_{i,t} \leftarrow \{\text{the unique subset of } A \text{ with the sum equal to } \hat{\beta}_i\}$
    Set $\hat{D}_i \leftarrow \{\text{columns of } C \text{ of the form } z = \alpha x_i, \forall \alpha \in \hat{A}_{i,t}\}$
End

Return $\hat{D} = \bigcup_i \hat{D}_i$
As a result, if there were no errors in $y$, one would have $S'_i \subseteq R_i$. Each erroneous entry of $y$ removes at most one coordinate of $R_i$ from $S'_i$. Since there are at most $e$ errors and $|S'_i| = 2e+1$, there exists a set of coordinates $S''_i \subseteq S'_i \cap R_i$ with cardinality at least $e + 1$ for which the corresponding entries of $y$ are error-free. As a result, $\forall j \in S''_i$ one has $y(j) = y_D(j) = f_\eta(\sum_{\alpha \in A_{i,t}} \alpha)$. □

**Proposition 5.** Given a SQLO$_\eta(d)\times d$ sequence $A$ and a fixed integer $\beta$, one can identify whether $\beta$ can be written as a sum of up to $d$ elements of $A$ with an algorithm of computational complexity $O(K)$, where $K = |A|$. Given that the answer to this question is positive, one can identify the elements of $A$ which sum up to $\beta$ with computational complexity $O(K)$.

*Proof.* This problem is known as the knapsack-solvability problem [53]. From the second property of Proposition 4, $\forall i: 1 \leq i \leq K$ and $\forall A_1 \subseteq \{\alpha_1, \alpha_2, \ldots, \alpha_{i-1}\}$ such that $|A_1| \leq d$, one has $\alpha_i > \eta \sum_{\alpha_j \in A_1} \alpha_j$, which also implies that $\alpha_i > \sum_{\alpha_j \in A_1} \alpha_j$.

To find the answer to the query with linear computational complexity, we perform a standard knapsack recursion [53]. First, we initialize the procedure by setting $\beta' \leftarrow \beta$ and $\hat{A}_\beta \leftarrow \emptyset$. Then, in the $i$-th iteration, we compare the value of $\beta'$ with the $i$-th largest element of $A$, $\alpha_{K-i+1}$. If $\beta' \geq \alpha_{K-i+1}$, then we update $\beta' \leftarrow \beta' - \alpha_{K-i+1}$ and $\hat{A}_\beta \leftarrow \hat{A}_\beta \cup \{\alpha_{K-i+1}\}$; otherwise, we go to the next iteration. The procedure stops with a negative answer to the first query if $|\hat{A}_\beta| > d$ or if $\beta' > 0$ and no element in $A$ is left that is smaller than or equal to $\beta'$. Otherwise, the procedure stops when $\beta' = 0$ with a positive answer to the first query, and $\hat{A}_\beta$ corresponds to the elements of $A$ that sum up to $\beta$. Note that this procedure is based on the superincreasing property of a SQLO$_\eta$ sequence, which implies that the largest element of $A$ that is smaller than $\beta'$ must be present in the sum. □

The previous proposition and lemma provide the core of the second step of Algorithm 2. The idea is that for each $x_i \in X_D$, we use Lemma 3 to find $S'_i$. The majority of elements $y(j), j \in S'_i$, correctly correspond to the bin in which $\beta_t = \sum_{\alpha \in A_{i,t}} \alpha$ is located. Each correctly identified bin contains a finite number of integers, one of which is the true value of $\beta_t$. As a result, by testing each such integer, we can determine whether it can be written as the sum of up to $d$ elements of $A$ or not using the algorithm in Proposition 5. The integer for which the answer to this query is positive is equal to $\beta_t$, which can then be used to identify the elements of $A_{i,t}$.
Theorem 9. The Dec-SQLO\textsubscript{s} algorithm is capable of identifying up to \(d\) defectives in the presence of at most \(e\) errors in the syndrome of defectives.

Proof. Since the first step of this algorithm is identical to the first step of the Dec-QBh algorithm, it follows that \(X = X_D\). Therefore, we only need to show that Step 2 recovers \(D\) given \(X_D\).

Fix a binary vector \(x_i \in X_D\). Fix a coordinate \(j \in S'_i\), and let \(\eta_l\) and \(\eta_u\) be the lower and upper thresholds of the quantization bin corresponding to \(y(j)\). Since all the sums of up to \(d\) elements of \(A\) fall into different bins, there exists exactly one subset sum \(\beta\) in \([\eta_l, \eta_u)\) that corresponds to the sum of up to \(d\) elements of \(A\). As a result, one can test all the \((\eta_u - \eta_l)\) integers in this bin using Proposition 5 to find the unique value of \(\beta\) that can be written as sum of up to \(d\) elements of \(A\). On the other hand, as was shown in Lemma 3, there exists a set \(S''_i \subseteq S'_i\) such that \(|S''_i| \geq e + 1\), and consequently \(\forall j \in S''_i\) one has \(y(j) = f_\eta(\sum_{\alpha \in A_{i,t}} \alpha) = f_\eta(\beta_t)\). As a result, the element \(\beta\) in the multiset \(B'\) with multiplicity at least \(e + 1\) corresponds to \(\beta_t\), or in other words \(\hat{\beta}_t = \beta_t\). This implies that \(\hat{A}_{i,t} = A_{i,t}\), and consequently, \(\hat{D} = D\). \qed

Remark 10. The computational complexity of the Dec-SQLO\textsubscript{s} algorithm is equal to \(O\left(\frac{mn}{K} + dm \log m + \deg_{\max}\right)\), where \(g_{\max} = \max_{i=1,2,...,Q} (\eta_i - \eta_{i-1})\) is the largest gap between the consecutive thresholds. The computational complexity of Step 1 is \(O\left(\frac{mn}{K}\right)\). On the other hand, sorting the elements of \(S_i\) to find \(S'_i\) requires \(O(dm \log m)\) computations. One can identify the elements of \(A\) that sum up to a fixed integer in linear time., i.e. using \(O(K)\) computational steps. As a result, the algorithm for finding \(\beta\) in each iteration has complexity \(O(eg_{\max}\)K). Hence, finding \(\hat{A}_{i,t}\) requires \(O(\deg_{\max}K)\) computational steps.

4.4 SQ-separable matrices using SQLO\textsubscript{t} sequences

The SQLO\textsubscript{s} sequences introduced in the previous section resolve the problem of exponential growth of decoding computational complexity with respect to \(K\). However, due to the superincreasing property of these sequences (the second property in Prop. 4) the multipliers \(\alpha_K\) tend to grow rapidly as a function of \(K\). In order to overcome this issue while preserving efficient decoding, we introduce a new family of integer sequences, termed SQLO\textsubscript{t}
sequences.

**Definition 14 (SQLO$_t$(η, $h$) sequences).** Given a set of thresholds $\eta$, a sequence of positive integers $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ is a SQLO$_t$(η, $h$) sequence if

1. $\alpha_K > \eta \alpha_{K-1} > \eta \ldots > \eta \alpha_2 > \eta \alpha_1 > \eta 0$ (i.e., all elements of $A$ lie in different quantization bins).

2. For any two subsets $A_1 \subseteq A$ and $A_2 \subseteq A$ such that $|A_1| < |A_2| \leq h$, one has $\sum_{\alpha_i \in A_2} \alpha_i > \eta \sum_{\alpha_i \in A_1} \alpha_i$ (i.e., subsets of different cardinality are ordered based on the number of their members).

3. For any two distinct subsets $A_1 = \{\alpha_1', \alpha_2', \ldots, \alpha_s'\}$ and $A_2 = \{\alpha_1'', \alpha_2'', \ldots, \alpha_s''\}$ with elements listed in an increasing order such that $|A_1| = |A_2| = s \leq h$, one has $\sum_{\alpha_i'' \in A_2} \alpha_i'' > \eta \sum_{\alpha_i' \in A_1} \alpha_i'$ if there exists $r : 1 \leq r \leq s$ such that $\forall i : 1 \leq i < r$, $\alpha_i' = \alpha_i''$ and $\alpha_i'' > \eta \alpha_i'$ (i.e., two subsets with the same cardinality are lexicographically ordered).

As an example, consider the set of thresholds $\eta = [0, 2, 5, 6, 10, 11, 15, 18]^T$ and let $h = 2$ and $K = 3$. The sequence $A_1 = \{2, 5, 10\}$ is a SQLO$_s$(η, 2) sequence that has the smallest value for $\alpha_3$, i.e. $\alpha_3 = 10$. On the other hand, the sequence $A_2 = \{4, 5, 6\}$ is a SQLO$_t$(η, 2) sequence that has the smallest positive value for $\alpha_3$, i.e. $\alpha_3 = 6$. This simple example illustrates how SQLO$_t$ properties may lead to denser sequences compared to SQLO$_s$ sequences.

The SQLO$_t$ properties impose a partial order on the subsets of the sequence. For example, if $K = 3$ and $h \geq K$, these properties become $\alpha_3 + \alpha_2 + \alpha_1 > \eta \alpha_3 + \alpha_2 > \eta \alpha_3 + \alpha_1 > \eta \alpha_2 + \alpha_1 > \eta \alpha_3 > \eta \alpha_2 > \eta \alpha_1 > \eta 0$. Similarly to the case of SQLO$_s$ sequences, it is not difficult to see that any SQLO$_t$(η, $h$) sequence is also a quantized $B_h$ sequence; however, the converse is not necessarily true. As a result, the following theorem holds.

**Theorem 10 (Construction 5).** Fix a binary $d$-disjunct matrix $C_b$ of dimensions $m_b \times n_b$, capable of correcting up to $e$ errors. Let $A = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ be a SQLO$_t$(η, $d$) sequence. Form a matrix $C$ of length $m = m_b$ and size $n = Kn_b$ by concatenating $K$ matrices $C_i = \alpha_i C_b$, $1 \leq i \leq K$ horizontally. The constructed matrix is a $[q; Q, \eta; (1:d); e]$-SQ-separable matrix with $q = \alpha_K + 1$.

**Proof.** Since any SQLO$_t$(η, $d$) sequence is a quantized $B_d$ sequence, the proof follows directly from Thm. 4.

\[\square\]
4.4.1 Fundamental limits and construction of SQLO$_l$ sequences

In [55], two types of lexicographically ordered sequences were defined that are closely related to the SQLO$_l$ sequences. For simplicity, we call these sequences “lex($h$)” and “strong-lex($h$)” and we provide their definition for completeness.

**Definition 15.** A sequence of positive integers $B = \{\beta_1, \beta_2, \ldots\}$ is a lex($h$) sequence, if for any two distinct subsets $B_1 = \{\beta_1', \ldots, \beta_h'\}$ and $B_2 = \{\beta_1'', \ldots, \beta_h''\}$ with elements listed in an increasing order, one has $\sum_{i \in B_2} \beta_i'' > \sum_{i \in B_1} \beta_i'$ if there exists an integer $r$, $1 \leq r \leq h$, such that $\forall i$, $1 \leq i < r$, $\beta_i' = \beta_i''$ and $\beta_r'' > \beta_r'$.

**Definition 16.** A sequence of positive integers $B = \{\beta_1, \beta_2, \ldots\}$ is a strong-lex($h$) sequence, if it is a lex($s$) sequence $\forall s \leq h$; in addition, for any two subsets $B_1 \subseteq B$ and $B_2 \subseteq B$ such that $|B_1| < |B_2| \leq h$, one has $\sum_{i \in B_2} \beta_i > \sum_{i \in B_1} \beta_i$.

The strong-lex($h$) sequences can be used to construct SQLO$_l$ sequences as shown in the next proposition.

**Proposition 6.** Consider a SQGT model with thresholds $\eta = [0, \eta_1, \ldots, \eta_Q]^T$; $\forall s : 1 \leq s \leq Q$, let $g_s = \max_{1 \leq s \leq Q} \eta_s - \eta_{s-1}$ be the largest gap of the first $s$ thresholds. Let $B = \{\beta_1 < \beta_2 < \ldots\}$ be a strong-lex($h$) sequence. For a fixed $s$, $2 \leq s \leq Q$, let $K_s$ be a positive integer small enough to satisfy $\eta_s > g_s \sum_{i=1}^{K_s} \beta_i$. All the sequences of the form $A_s = \{g_s \beta_1, g_s \beta_2, \ldots, g_s \beta_{K_s}\}$ are SQLO$_l(\eta, h)$ sequences.

**Proof.** The proof of this proposition follows along the same lines as the proof of Thms. 5 and 8, and is hence omitted. $\square$

As we demonstrated through a simple example earlier, the SQLO$_l$ properties may result in denser sequences compared to SQLO$_s$ properties. However, a SQLO$_l(\eta, h)$ sequence constructed from strong-lex($h$) sequences according to Proposition 6 does not improve the bound $\alpha_K = O_g(\gamma^K)$ derived in Lemma 2. This can be shown as follows. We define an optimal lex($h$) sequence $B = \{\beta_1, \beta_2, \ldots, \beta_K\}$ with the smallest possible value of $\beta_K$. In [55, Thm. 1], it was proven that the largest element of an
optimal lex$(h)$ sequence satisfies $\beta_K = O(\gamma^K)$, where $\gamma$ is the largest root of $x^{h+1} - 2x^h + 1 = 0$. Since any strong-lex$(h)$ sequence needs to also satisfy the lex$(h)$ property, one can conclude that a SQLO$_l(n, h)$ sequence constructed from strong-lex$(h)$ sequences according to Proposition 6 cannot improve the bound $\alpha_K = O(\gamma^K)$.

### 4.4.2 Decoding algorithm for SQGT matrices constructed using SQLO$_l$ sequences

Next, we describe the Dec-SQLO$_l$ algorithm, the decoding procedure for matrices based on SQLO$_l$ sequences. This algorithm resembles the Dec-SQLO$_s$ algorithm, and similar intuition also applies as follows. In the first step, one identifies $\mathcal{X} = \mathcal{X}_D$, the set of binary columns corresponding to the support of the columns in $\mathcal{D}$. To complete the decoding, one needs to identify the set of elements $A_{i,t} \subseteq A$ which are used to form the columns in $\mathcal{D}$ from the binary column $x_i$, $\forall x_i \in \mathcal{X}_D$.

Since in the proof of Lemma 3 we only used the quantized $B_d$ property, this lemma also holds for matrices constructed using Thm. 10. As a result, using the notation defined in Lemma 3, for any binary column $x_i \in \mathcal{X}_D$ there exists at least $e + 1$ elements of $S'_i$, denoted by $S''_i$, such that $\forall j \in S''_i$ one has $y(j) = f_{\eta}(\sum_{\alpha \in A_{i,t}} \alpha)$. This implies that the majority of elements $y(j)$, $j \in S'_i$, correctly correspond to the bin in which $\beta_t = \sum_{\alpha \in A_{i,t}} \alpha$ is located. Each correctly identified bin contains a limited number of integers, one of which is the true value of $\beta_t$. As a result, by testing each such integer, we can determine whether it can be written as the sum of up to $d$ elements of $A$ or not. The integer for which the answer to this query is positive is equal to $\beta_t$, which can then be used to identify the elements of $A_{i,t}$.

As a result, given a SQLO$_l$ sequence $A$ and an integer $\beta$, the main issue is to efficiently determine whether $\beta$ can be written as the sum of up to $d$ elements of $A$; and if so, what those elements are. In Lemma 4, an algorithm with computational complexity of $O(K)$ is described that can perform this task.

---

Note that there exists a typo in the statement of [55, Thm. 1], in which $\gamma$ is defined as the largest root of $x^{h+1} - x^h + 1 = 0$. However, it is evident from the proof of the theorem that $\gamma$ is in fact the largest root of $x^{h+1} - 2x^h + 1 = 0$. 

---

45
Lemma 4. Given a SQLO$_1(\eta, d)$ sequence $\mathcal{A}$ and a fixed integer $\beta$, it is possible to identify whether $\beta$ can be written as a sum of up to $d$ elements of $\mathcal{A}$ with complexity $O(K)$, where $K = |\mathcal{A}|$. Given that the answer to this question is positive, one can identify these elements of $\mathcal{A}$ with complexity $O(K)$.

Proof. Suppose $\beta$ can be written as the sum of $s \leq d$ elements of $\mathcal{A}$, and let $\mathcal{A}_t \subseteq \mathcal{A}$ be the subset such that $\sum_{\alpha \in \mathcal{A}_t} \alpha = \beta$. The value of $s = |\mathcal{A}_t|$ can be easily determined as follows. First, we form the set $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$, where $\gamma_i = \sum_{j=1}^i \alpha_i$, $1 \leq i \leq K$. As a consequence of the second property in Def. 14, for any $1 \leq i \leq K$, $\gamma_i$ is larger than all $j$-subsets of $\mathcal{A}$ for $j < i$. On the other hand, due to the third property in Def. 14, $\gamma_i$ is smaller than all $j$-subsets of $\mathcal{A}$ for $j > i$. Consequently, one can determine $s$ using $s = \min\{i : \beta < \gamma_i\} - 1$.

Given the value of $s$, we can determine the elements of $\mathcal{A}_t$ successively using $K$ iterations. First, we initialize the procedure by setting $s' \leftarrow s$, $\beta' \leftarrow \beta$, and $\mathcal{A}' \leftarrow \emptyset$. In the $i$-th iteration, $1 \leq i \leq K$, we determine whether $\alpha_i \in \mathcal{A}_t$ or not. At the beginning of the $i$-th iteration, $\mathcal{A}'$ equals $\mathcal{A}_t \cap \{\alpha_1, \alpha_2, \ldots, \alpha_{i-1}\}$, and $s'$ is equal to the number of remaining unidentified elements of $\mathcal{A}_t$, i.e. $s' = |\mathcal{A}_t \setminus \mathcal{A}'|$. In addition, $\beta'$ is equal to the sum of $s'$ elements in $\mathcal{A}_t \setminus \mathcal{A}'$. To determine whether $\alpha_i$ is in $\mathcal{A}_t$, we use the following rule: if $\beta' < \alpha_i + \alpha_{i+2} + \cdots + \alpha_{i+s'}$, then $\alpha_i \notin \mathcal{A}_t$; the reason is that the sum of any $s'$ elements of $\{\alpha_1, \alpha_{i+1}, \ldots, \alpha_{K}\}$ that does not include $\alpha_i$ is at least as large as $\alpha_{i+1} + \alpha_{i+2} + \cdots + \alpha_{i+s'}$. Therefore, if $\beta' < \alpha_i + \alpha_{i+2} + \cdots + \alpha_{i+s'}$, then $\alpha_i$ must be in $\mathcal{A}_t$. Given that this condition is satisfied, we update $\mathcal{A}' \leftarrow \mathcal{A}' \cup \{\alpha_i\}$, $\beta' \leftarrow \beta' - \alpha_i$, and $s' \leftarrow s' - 1$. Otherwise, we go to the next iteration. The algorithm stops after $K$ iterations. At the end, if $s' = 0$ and $|\mathcal{A}'| \leq d$, the answer to the first query is positive and $\mathcal{A}' = \mathcal{A}_t$. Otherwise the answer to the query is negative. \qed

The decoding algorithm for matrices constructed using Thm. 10 is described in Algorithm 3. Note that the main difference between this algorithm and Algorithm 2 is that in Step 2, we use Lemma 4 instead of Proposition 5.

4.5 Construction of binary SQ-separable matrices

The constructions considered up to this point used an alphabet size of $q \geq \eta_1 + 1$. On the other hand, it is important to address the issue of constructing
Algorithm 3: Dec-SQL0

\textbf{Input:} \( y \in [Q]^m, C_b \in [2]^{m \times \frac{n}{K}}, \eta, A, e \geq 0 \)
\textbf{Output:} \( \hat{\mathcal{D}} \)

\textbf{Step 1:} Initialize \( \mathcal{X} \leftarrow \emptyset \) and \( \hat{\mathcal{D}} \leftarrow \emptyset \)
\textbf{For} \( i = 1, 2, \ldots, \frac{n}{K} \) \textbf{do}
\hspace{1em} \( x_i \leftarrow \text{the } i\text{-th column of } C_b \)
\hspace{1em} \( N_i \leftarrow \text{number of coordinates } j \text{ for which } x_i(j) > y(j) \)
\hspace{1em} \textbf{If} \( N_i \leq e \) \textbf{then}
\hspace{2em} \text{Set } \mathcal{X} \leftarrow \mathcal{X} \cup \{x_i\}
\hspace{1em} \textbf{End}
\textbf{End}

\textbf{Step 2:}
\textbf{For} \( i = 1, 2, \ldots, |\mathcal{X}| \) \textbf{do}
\hspace{1em} Set \( \mathcal{S}_i \leftarrow \text{the set of nonzero coordinates of } x_i \)
\hspace{1em} Set \( \mathcal{S}_i' \leftarrow \text{subset of } \mathcal{S}_i \text{ with } |\mathcal{S}_i'| = 2e + 1 \text{ s.t. } \forall k \in \mathcal{S}_i' \text{ and } \forall j \in \mathcal{S}_i \setminus \mathcal{S}_i', \text{ one has } y(k) \leq y(j) \}
\hspace{1em} Initialize the multiset \( B' \leftarrow \emptyset \)
\hspace{1em} \textbf{For} \( j = 1, 2, \ldots, |\mathcal{S}_i'| \) \textbf{do}
\hspace{2em} \( \eta_u \leftarrow \text{the upper threshold of the quantization bin for } y(j) \)
\hspace{2em} \( \eta_l \leftarrow \text{the lower threshold of the quantization bin for } y(j) \)
\hspace{2em} \( \beta \leftarrow \text{the unique integer } \eta_l \leq \beta < \eta_u \text{ that can be written as the sum of up to } d \text{ elements of } A \) (use Lemma 4)
\hspace{2em} Update the multiset \( B' \leftarrow B' \cup \{\beta\} \)
\hspace{1em} \textbf{End}
\hspace{1em} Set \( \hat{\beta}_i \leftarrow \text{the element of } B' \text{ with at least } e + 1 \text{ repetitions} \)
\hspace{1em} Set \( \hat{\mathcal{A}}_{i,t} \leftarrow \{\text{the unique subset of } A \text{ with the sum equal to } \hat{\beta}_i\} \)
\hspace{1em} Set \( \hat{\mathcal{D}}_i \leftarrow \{\text{columns of } C \text{ of the form } z = \alpha x_i, \forall \alpha \in \hat{\mathcal{A}}_{i,t}\} \)
\hspace{1em} \textbf{End}
\textbf{End}

\textbf{Return} \( \hat{\mathcal{D}} = \bigcup_i \hat{\mathcal{D}}_i \)
SQGT matrices with alphabet size \( q \leq \eta_1 \), and in particular \( q = 2 \). This problem may be solved by noticing that SQGT can be viewed as a generalization of TGT with a zero gap. While in TGT with zero gap there exist only one threshold, in SQGT one may have more than one threshold if \( Q \)-ary test results are allowed. This implies that any matrix constructed for TGT is also a SQ-separable matrix. In [56], Chen and Fu observed that a variation of binary disjunct matrices, also studied under the name of cover-free families (see [57, 58, 59]), can be used for TGT. In [60] Cheraghchi showed that a weaker notion of disjunct matrices, so called threshold disjunct matrices, are also applicable to the TGT problem and provided constructions with high rates. In the following theorem, we describe a generalization of these matrices that is particularly useful for the SQGT model. This generalization provides binary and non-binary matrices for arbitrary thresholds, \( \eta \).

Theorem 11. Let \( \eta_\alpha \) be the \( \alpha \)th threshold in a SQGT model. Consider a matrix \( C \in [q]^{m \times n} \) such that for any subset of column-indices \( S \subseteq [n] \) with \( \frac{n}{q - 1} \leq |S| \leq d \), and for any index \( l \in S \), any set \( N \in [n] \), where \( |N| \leq |S| \), and \( S \cap N = \emptyset \), there exists a set of row-indices \( R \) with size at least \( 2e + 1 \), such that \( \forall j \in R \) it holds that

\[
\sum_{k \in S} C(j, k) \in \{ \eta_1, \eta_2, \ldots, \eta_\alpha \}, \quad (4.14)
\]

\[
\sum_{k \in N} C(j, k) = 0, \quad (4.15)
\]

\[
C(j, l) \neq 0. \quad (4.16)
\]

Then, \( C \) is a \([q; Q; \eta; (\frac{n}{q - 1}); d; e] \)-SQ-separable matrix.

Proof. Consider two distinct sets of columns (i.e. columns of \( C \)), denoted by \( \mathcal{X} \) and \( \mathcal{Z} \), such that \( \frac{n}{q - 1} \leq |\mathcal{X}|, |\mathcal{Z}| \leq d \). Without loss of generality, assume that \( |\mathcal{X}| \geq |\mathcal{Z}| \). Let \( S \) be the set of column-indices corresponding to \( \mathcal{X} \). Also, let \( N \) be the set of column-indices corresponding to \( \mathcal{Z} \setminus \mathcal{X} \). Consequently, \( \frac{n}{q - 1} \leq |S| \leq d \), \( |N| \leq |S| \), and \( S \cap N = \emptyset \). Let \( l \) be the index of the column \( x_l \in \mathcal{X} \setminus \mathcal{Z} \). Such a column always exists due to the manner in which \( \mathcal{X} \) and \( \mathcal{Z} \) are chosen.

From the definition of \( C \), there exists a set of row-indices with size \( |R| \geq 2e + 1 \) such that \( \forall k \in R \), conditions (4.14)-(4.16) are satisfied. This implies
that $\forall k \in \mathcal{R}$,

$$y_x(k) > y_z(k).$$

As a result, $C$ is a $[q; Q; \eta; ([\eta_n q^{-1}]; d); e]$-SQ-separable matrix. $\square$

The next theorem describes a probabilistic construction for this type of SQ-separable matrices with $q = 2$. This construction can be generalized for $q > 2$ in a similar manner.

**Theorem 12 (Construction 6).** Let $r = \lceil \log_2 \frac{d}{\eta_n} \rceil + 1$, $\mu = \frac{1}{2^r} \left(1 - \frac{1}{\eta_n}\right)$, and $\rho = \frac{1}{2} \sum_{\beta=1}^{\alpha} \left(\frac{\mu}{\eta_n - 1}\right)^{\eta_n \eta_{\beta-1} d / \eta_n - 1}$. Assume that $d = o(n)$. For any $i \in [r]$, form a binary matrix $C_i \in [2]^{(m/\rho) \times n}$ by choosing each entry independently according to a Bernoulli distribution such that the probability of choosing 1 equals $P_i = \frac{1}{2^{r^2/\eta_n}}$. Now, form a matrix $C = [C_1^T, C_2^T, \ldots, C_r^T]^T$, where $T$ denotes the matrix transpose operator. Then $C$ is a $[2; Q; \eta; (\eta_n ; d); 0]$-SQ-separable matrix with probability at least $1 - o(1)$, provided that $m = r \left(\frac{2d}{\rho} + \delta\right) \log \frac{n}{\delta}$, $\forall \delta > 0$. Similarly, $C$ is a $[2; Q; \eta; (\eta_n ; d); e]$-SQ-separable matrix with probability at least $1 - o(1)$, if $m = r \left(\frac{4d}{\rho} + \delta\right) \log \frac{n}{\delta} + \frac{4e}{\rho}$, $\forall \delta > 0$.

**Proof.** The idea behind this construction is that each sub-matrix $C_i$, $i \in [r]$, satisfies conditions (4.14)-(4.16) for different sizes of $S$.

From Theorem 11, we know that for $q = 2$ it is only required to consider $S$ with size $\eta_n \leq |S| \leq d$; therefore, for any such choice of $S$ we can find $i \in [r]$ such that $\eta_n 2^{i-1} \leq |S| < 2^i \eta_n$. Fix a choice of $S$, a choice of $l \in S$, and a choice of $N$ such that $|N| \leq |S|$. Let $A_i$ denote the total number of such choices. Form $C_i$ by choosing each entry independently according to a Bernoulli distribution such that the probability of choosing 1 equals $P_i = \frac{1}{2^{r^2/\eta_n}}$. Let $\pi_i$ denote the probability that a fixed row of $C_i$ denoted by $r$ satisfies conditions (4.14)-(4.16). Note that since the entries of $C_i$ are chosen according to an i.i.d. probability distribution, the choice of $r$ does not affect $\pi_i$. Let $E_{\beta}$, $\beta \in [\alpha]$, be the event that $\sum_{k \in S} r(k) = \eta_{\beta}$, and $\sum_{k \in N} r(k) = 0$, and $r(l) = 1$. Consequently,

$$\pi_i = \Pr\left(\bigcup_{\beta=1}^{\alpha} E_{\beta}\right) = \sum_{\beta=1}^{\alpha} \Pr(E_{\beta}),$$

where the second equality follows from the disjointness of these events. A

49
lower bound on the probability of the event $\mathcal{E}_\beta$ can be found using

$$
\Pr(\mathcal{E}_\beta) = \sum_{T \subseteq \mathcal{S} \setminus \{l\}, \sum_{i} \Pr(r(k) = 1, \ \forall k \in T) \cdot \Pr(r(l) = 1)
\cdot \Pr(r(k) = 0, \ \forall k \in (\mathcal{S} \cup \mathcal{N}) \setminus (T \cup \{l\}))
= \sum_T P_i^{\eta_\beta} \cdot P_i \cdot (1 - P_i)^{|\mathcal{S} + \mathcal{N} - \eta_\beta| - \eta_\beta}
\geq \sum_T P_i^{\eta_\beta} \cdot (1 - (|\mathcal{S} + |\mathcal{N} - \eta_\beta| \cdot P_i).
$$

On the other hand,

$$
P_i(|\mathcal{S} + |\mathcal{N} - \eta_\beta) \leq P_i(2|\mathcal{S} - \eta_\beta) = 2P_i(|\mathcal{S} - \eta_\beta|
= \frac{1}{2^{i+1/\eta_\alpha}} (|\mathcal{S} - \eta_\beta|) \leq \frac{|\mathcal{S}| - \frac{\eta_\beta}{2}}{\eta_\alpha} \leq \frac{1}{2}.
$$

As a result,

$$
\Pr(\mathcal{E}_\beta) \geq \sum_T \frac{1}{2^{i+1/\eta_\alpha}} \frac{1}{2^{i+1/\eta_\alpha}} \frac{1}{2^{i+1/\eta_\alpha}} P_i^{\eta_\beta} \geq \frac{1}{2^{i+1/\eta_\alpha}} \frac{1}{2^{i+1/\eta_\alpha}} \frac{1}{2^{i+1/\eta_\alpha}} P_i^{\eta_\beta}
= \frac{1}{2} \left( \frac{P_i(|\mathcal{S} - \eta_\beta|) \eta_\beta}{|\mathcal{S} - 1|} \right)^{\eta_\beta} \frac{1}{2} \left( \frac{(2^{-3} - 2^{-\frac{1}{2}/\eta_\alpha}) \eta_\beta}{|\mathcal{S} - 1|} \right)^{\eta_\beta}
\geq \frac{1}{2} \left( \frac{\mu}{\eta_\beta - 1} \right)^{\eta_\beta} \frac{\eta_\beta - 1}{|\mathcal{S} - 1|}
\geq \frac{1}{2} \left( \frac{\mu}{\eta_\beta - 1} \right)^{\eta_\beta} \frac{\eta_\beta - 1}{d - 1}
$$

where $\mu = \frac{1}{2^{i+1/\eta_\alpha}} (1 - \frac{1}{\eta_\alpha})$. Consequently, a lower bound on $\pi_\alpha$ reads as

$$
\pi_i = \sum_{\beta=1}^{\alpha} \Pr(\mathcal{E}_\beta) \geq \frac{1}{2} \sum_{\beta=1}^{\alpha} \left( \frac{\mu}{\eta_\beta - 1} \right)^{\eta_\beta} \frac{\eta_\beta - 1}{d - 1} := \rho,
$$

which is independent of $i$.

Using a union bound and (4.17), we arrive at an upper bound on the probability that $\mathcal{C}$ does not satisfy the conditions in Theorem 11, i.e.

$$
P_F \leq \sum_{i=1}^{r} A_i P_{F_i}(\pi_i).
$$

Here, $P_{F_i}(\pi_i)$ is the probability that $\mathcal{C}_i$ does not satisfy the conditions in Definition 8 for a choice of $\mathcal{S}$ that satisfies $\eta_\alpha 2^{i-1} \leq |\mathcal{S}| < 2^i \eta_\alpha$. 

50
Next, let $m'$ denote the number of rows of $C_i$, for all $i \in [r]$. If $e = 0$, then

$$P_F(\pi_i) = (1 - \pi_i)^{m'} \leq (1 - \rho)^{m'} \leq \exp(-m'\rho) \leq p_F(\rho); \quad (4.19)$$

otherwise, for $e > 0$ we can use the Chernoff bound to find

$$P_F(\pi_i) \leq \exp\left(-\frac{m'\rho}{2}\left(1 - \frac{2e}{m'\rho}\right)^2\right) \leq p_F(\rho). \quad (4.20)$$

Since these upper bounds are independent of $i$, (4.18) simplifies to

$$P_F \leq A_\alpha p_F(\rho), \quad (4.21)$$

where $A_\alpha = \sum_{i=1}^{r} A_i$ and $p_F(\rho)$ are defined in (4.19) and (4.20) for $e = 0$ and $e > 0$, respectively.

Since $A_\alpha$ is equal to the total number of choices for $S$, $l$, and $N$, one has

$$A_\alpha = \sum_{s=\eta_\alpha}^{d} \binom{n}{s}^s \sum_{z=0}^{\min(s,n-s)} \binom{n-s}{z},$$

where $s$ denotes the size of $S$ and $z$ denotes the size of $N$. Since $\binom{n-s}{z} \leq \binom{n}{s}$ for any $z \in \{0,1,\ldots,\min(s,n-s)\}$, by assuming that $d \leq \frac{n}{2}$ for simplicity, we may write

$$A_\alpha \leq \sum_{s=\eta_\alpha}^{d} \binom{n}{s}^2 (s+1)s < \sum_{s=\eta_\alpha}^{d} \left(\frac{ne}{s}\right)^{2s} (s+1)s$$

$$< (d-\eta_\alpha)(d+1)\left(\frac{ne}{d}\right)^{2d} < d^3 \left(\frac{ne}{d}\right)^{2d}, \quad (4.22)$$

where $e = \exp(1)$ denotes the base of the natural logarithm and is not to be confused with the number of errors $e$ that the matrix can correct.

Note that the third inequality follows from the fact that the largest term in $\sum_{s=\eta_\alpha}^{d} \left(\frac{ne}{s}\right)^{2s} (s+1)s$ is indexed by $s = d$. This can be easily shown by noting that

$$\frac{\left(\frac{ne}{s}\right)^{2s} (s+1)s}{\left(\frac{ne}{s+1}\right)^{2s+2} (s+1)(s+2)} = \left(1 + \frac{1}{s}\right)^{2s} \frac{(s+1)^2 s^{-2}}{s + 2} \leq \frac{1}{n^2}s(s+1) < 1.$$

Using (4.17), (4.19), (4.21), and (4.22), the probability that $C$ is not an
$rm' \times n \ [2; Q; \eta, \alpha; d, 0]$-SQ-separable matrix is upper bounded by

$$P_F \leq d^3 \left(\frac{ne}{d}\right)^{2d} \exp(-m' \rho) = \exp \left(2d \log n + 3 \log d + 2d - 2d \log d - m' \rho\right).$$

As a result, if $d = o(n)$, for any $\delta > 0$, one has $P_F = o(1)$ if

$$m = rm' = r \left(\frac{2d}{\rho} + \delta\right) \log \frac{n}{d}.$$

Similarly, the probability that $C$ is not an $rm' \times n \ [2; Q; \eta, \alpha; d, e]$-SQ-separable matrix is upper bounded by

$$P_F \leq d^3 \left(\frac{ne}{d}\right)^{2d} \exp \left(-\frac{m' \rho}{2} \left(1 - \frac{2e}{m' \rho}\right)^2\right) = \exp \left(2d \log n + 3 \log d + 2d - 2d \log d - \frac{m' \rho}{2} \left(1 - \frac{2e}{m' \rho}\right)^2\right).$$

Then, if $d = o(n)$, for any $\delta > 0$, one has $P_F = o(1)$ if

$$m = rm' = r \left[\left(\frac{4d}{\rho} + \delta\right) \log \frac{n}{d} + \frac{4e}{\rho}\right].$$

\[\square\]

**Remark 11.** As discussed earlier, any matrix designed for TGT without a gap, such that $\eta_T \in \{\eta_1, \eta_2, \ldots, \eta_Q\}$, can be used for the purpose of SQGT. Hence, the threshold disjunct matrices in [60], constructed probabilistically, provide an alternative to the matrices in Construction 6 for the SQGT model. However, as the next lemma indicates, the rate $^6$ of this family of threshold disjunct matrices, $R_{TD}$, is decreasing function of $\eta_T$ and the highest rate is achieved if $\eta_T = \eta_1$. Consequently, the matrices described in Construction 6 provide an improvement in the rate, quantified as follows. For any $\eta_T \in$

---

$^6$We define the rate as $R = \frac{\log_2 n}{m}$. 

---

52
\{\eta_1, \eta_2, \ldots, \eta_Q\}, it holds that
\[
\frac{R_{SQ}}{R_{TD}} \geq \min_{\eta_T \in \{\eta_1, \eta_2, \ldots, \eta_Q\}} \frac{R_{SQ}}{R_{TD}(\eta_T)} = \frac{\log_2 \frac{d}{\eta_T} + 1}{\log_2 \frac{d}{\eta_1} + 1} \frac{\sum_{\beta=1}^{\eta_{\beta-1}} \left( \frac{\mu_{\beta}}{\eta_{\beta-1}} \right) \eta_{\beta-1} - 1}{d - 1}.
\]

where \(R_{SQ} = \frac{\log_2 n}{m}\) is the rate of the matrix in Construction 6, and \(\mu_{\eta_1} = \frac{1}{2^t} \left( 1 - \frac{1}{\eta_1} \right)\). As an example, if \(d = \eta_\alpha = 4 \eta_1\), then \(R_{SQ} \geq 3\).

Lemma 5. The rate of the family of threshold disjunct matrices constructed probabilistically in [60], denoted by \(R_{TD} = \frac{\log_2 n}{m}\), is a decreasing function of \(\eta_T\) (for a fixed \(d\)) and the highest rate is achieved if \(\eta_T = \eta_1\).

Proof. In order to show that for a fixed \(d\), the rate \(R_{TD}\) is a decreasing function of \(\eta_T = \eta, 2 \leq \eta \leq d\), we express the rate as \(R_{TD} = \frac{C_d}{f(d, \eta)}\), where \(C_d\) is a coefficient that depends on \(d\),
\[
f(d, \eta) = \frac{\left( \log_2 \frac{d}{\eta_1} + 1 \right)}{\left( \frac{\mu_{\eta}}{\eta_1} \right) \frac{\eta_{\eta-1}}{d-1}},
\]
and \(\mu_{\eta} = \frac{1}{2^t} \left( \frac{\eta-1}{\eta} \right)\). Consequently, \(f(d, \eta) = \left( \left( \log_2 \frac{d}{\eta_1} + 1 \right) \frac{\eta_{\eta-1}}{d-1} \right) (8\eta)^\eta\). Now, to prove that \(R_{TD}\) is decreasing in \(\eta\), it suffices to show that \(f(d, \eta)\) is an increasing function of \(\eta, 2 \leq \eta \leq d\). Let \(d\) be fixed, where \(d \geq 3\). In what follows, we prove that \(\forall \eta \in \{2, 3, \ldots, d - 1\}\),
\[
\frac{f(d, \eta + 1)}{f(d, \eta)} \geq 1.
\]
One has
\[
\frac{f(d, \eta + 1)}{f(d, \eta)} = \frac{\log_2 \frac{d}{\eta + 1} + 1}{\log_2 \frac{d}{\eta} + 1} \left( \frac{\eta + 1}{\eta} \right)^\eta + 8(\eta - 1) \geq 27 \frac{\log_2 \frac{d}{\eta + 1} + 1}{\log_2 \frac{d}{\eta} + 1},
\]
where the inequality follows since \(\eta \geq 2\). Let \(K = [\log_2 d] \). Since \(1 \leq \frac{d}{\eta + 1} < \frac{d}{\eta} \leq \frac{d}{2}\), we partition the closed interval \([1, d/2]\) into a union of disjoint intervals
where for \(1 \leq k \leq K\), \(\mathcal{I}_k = [2^{k-1}, 2^k)\), and \(\mathcal{I}_{K+1} = [2^K, d/2]\). If \(d/\eta + 1 \leq \frac{d}{\eta+1} < \frac{d}{\eta} < 2^k\), then

\[
\frac{f(d, \eta + 1)}{f(d, \eta)} \geq 27 \frac{k-1+1}{k-1+1} = 27 > 1.
\]

If for some \(1 \leq k \leq K\), one has \(\frac{d}{\eta+1} \in \mathcal{I}_k\) and \(\frac{d}{\eta} \geq 2^k\), then

\[
\frac{f(d, \eta + 1)}{f(d, \eta)} \geq 27 \frac{\log_2 \frac{d}{\eta+1} + 1}{\log_2 \frac{d}{\eta} + 1} = \frac{27 k}{\left[\log_2 \left(2^k(1 + \Delta)\right)\right] + 1 + k},
\]

where \(\Delta = \frac{d}{\eta 2^k} - 1\). Since \(\Delta \geq 0\), one has \(\log_2 (1 + \Delta) \leq \frac{\Delta}{\ln 2}\). Since \(\frac{d}{\eta+1} < 2^k\), it follows that \(\Delta < 1/\eta\). Consequently, \(\log_2 (1 + \Delta) < \frac{1}{\eta 2^k} < 1\) and therefore \(\log_2 (1 + \Delta) = 0\). As a result,

\[
\frac{f(d, \eta + 1)}{f(d, \eta)} \geq \frac{27 k}{\left[\log_2 (1 + \Delta)\right] + 1 + k} = \frac{27 k}{1 + k} \geq \frac{27}{2} > 1.
\]

This proves the claim that \(R_{T_D}\) is a decreasing function in \(\eta_T\).

Next, we describe an explicit construction of the family of matrices described in Theorem 11. In [60], an explicit construction based on lossless condensers [61] for TGT matrices was described. In what follows, we explain how to use the building blocks of [60, Construction 3] for TGT and leverage the fact that in SQGT we have \(Q\) thresholds at our disposal.

The key ingredients of our method are building block matrices for threshold disjunct matrices (henceforth, BBTD matrices) [60, Construction 3]. BBTDs are obtained from a strong lossless \((k, \bar{e})\)-condenser\(^7\) \(f : \{0, 1\}^\bar{e} \times \{0, 1\}^{\bar{e}'} \rightarrow \{0, 1\}^l\); if the parameters of the BBTD matrix are \(m' \times n'\), then \(n' = 2\bar{n}\) and \(m' = 2^{t+k} \left(\frac{\eta_T}{\bar{e}} 2^{\bar{e}-1}\right) = 2^{i+k} O_{\eta_T} \left(2^{n_T(l-k)}\right)\), where \(\eta_T\) is the threshold in the TGT model, and \(k\) and \(\bar{e}\) denote the entropy and the error in the definition of a lossless condenser, respectively. Also, \(\bar{e} < (1 - p)/16\) for some real parameter

\(^7\) For the definition and a detailed explanation of strong lossless condensers, see [60, Definition 1] and [61].
0 \leq p < 1$. Let $\tilde{\gamma} := \max\{1, 2^{k-i}2^{k}/(10\eta_r)\}$. The following lemma was proved in [60].

Lemma 6. In a BBTD matrix $B$ with parameters described above, and for any subset of column-indices $S \subseteq [n]$ with $2^{k-2} \leq |S| \leq 2^{k-1}$, and for any $N \subseteq [n]$, where $|N| \leq |S|$, and $S \cap N = \emptyset$, there exists a set of row-indices $R$ with size at least $p\tilde{\gamma}2^{k}$, such that $\forall j \in R$

$$\sum_{k \in S} B(j,k) = \eta_r$$  \hspace{1cm} (4.23)

$$\sum_{k \in N} B(j,k) = 0.$$ \hspace{1cm} (4.24)

The BBTD matrices described above are used in [60] to obtain the so-called “regular” matrices, which are then converted into threshold disjunct matrices.

In the next theorem, we use BBTD matrices to construct SQ-separable matrix with rates exceeding their threshold disjunct matrix counterparts with $\eta_r \in \{\eta_1, \eta_2, \ldots, \eta_Q\}$.

Theorem 13 (Construction 7). Assume that $d \geq \eta_\alpha \geq \eta_1 > 1$. Let $\eta'_\alpha = 2^{[\log_2(\eta_\alpha-1)]}$ be the smallest power of 2 that is at least as large as $(\eta_\alpha-1)$, let $r = \lfloor \log_2((d-1)/\eta'_\alpha) \rfloor$, and let $p \in [0, 1)$. Let $B = \{B_i\}_0^r$ be a set of binary BBTD matrices constructed for parameter $\eta_r = \eta_{\alpha-1}$ using a family of strong lossless $(\tilde{k}_1, \tilde{\epsilon})$-condensers $F = \{f_i\}_0^r$, where $\tilde{k}_1 = [\log_2(\eta_{\alpha-1})] + i + 1$ and $\tilde{\epsilon} < (1-p)/16$.

For each $i \in [r+1]$, $f_i : \{0, 1\}^{\tilde{k}_1} \times \{0, 1\}^i \rightarrow \{0, 1\}^{\tilde{l}_i}$, and for the corresponding BBTD matrix, one has $B_i \in [2]^{m_i \times n}$ where $m_i = 2^{i+\tilde{k}_1} O(\eta_{\alpha-1}(2^{(\eta_{\alpha-1})}(\tilde{l}_i-\tilde{k}_i)))$ and $n = 2^{\tilde{n}}$. In step 1, $\forall i \in [r+1]$ construct $B'_i \in [2]^{2^{r-i}m_i \times n}$ by repeating $B_i$, $2^{r-i}$ times according to the rule $B'_i = [B'_i, B'_i, \ldots, B'_i]^{T}$. In step 2, form matrix $C' = [B'_0^{T}, B'_1^{T}, \ldots, B'_r^{T}]^{T}$. In step 3, fix a $d$-disjunct binary matrix $D \in [2]^{m_d \times n}$ capable of correcting $e_1$ errors in the CGT model. Form the binary matrix $C$ such that its $k^{th}$ row is equal to the bit-wise OR of the $i^{th}$ row of $C'$ and the $j^{th}$ row of $D$, where $i = \lfloor \frac{k}{m_d} \rfloor$ and $j = k - (i-1)m_d$.

Then $C$ is a $[2; Q; \eta; (\eta_\alpha : d); e] - SQ$-separable matrix of dimensions $m \times n$, where $m = 2^{i}m_d(d-1)^{\frac{n_i-1}{\eta_{\alpha-1}}}(\sum_{i=0}^{\eta_\alpha} O(\eta_{\alpha-1}(2^{(\eta_{\alpha-1})}(\tilde{l}_i-\tilde{k}_i))))$, $e = \left\lfloor \frac{2^{r+1}p2^{d-1}}{2^{\tilde{n}}} \right\rfloor$, and $\tilde{\gamma}' = \max\{1, \frac{d-1}{5} \min_{\epsilon \in [r+1]} \left\{2^{k_i-l_i} \right\}\}$.

Proof. First, we provide the sketch of the proof in order to build some intuition. The idea behind the proof is to first show that the interval $[\eta_\alpha-1, d-1]$
is a subset of the interval \( [2^{-1} \eta_{\alpha}, 2^r \eta_{\alpha}] \). Then, using the definition of \( \tilde{k}_i, \ i \in [r + 1] \), we show that \( [2^{-1} \eta_{\alpha}, 2^r \eta_{\alpha}] = \bigcup_{i=0}^{r} [2^{k_i-2}, 2^{k_i-1}] \). Then by construction of \( B_i, \ i \in [r + 1] \), we have that \( B_i \) has at least \( p \tilde{\gamma}_i 2^i \) rows that satisfy (4.23) and (4.24) for \( \eta_\gamma = \eta_i - 1 \) and \( 2^{\tilde{k}_i-2} \leq |S| \leq 2^{\tilde{k}_i-1} \), where \( \tilde{\gamma}_i = \max\{1, 2^{k_i-i}2^{\tilde{k}_i}/(10(\eta_i - 1))\} \). Since each \( B_i' \) is formed by concatenating \( B_i \) vertically \( 2^{r-i} \) times, \( i \in [r + 1] \), then \( B_i' \) has at least \( p \tilde{\gamma}_i 2^{r+i-r} \) rows that satisfy (4.23) and (4.24) for \( \eta_r = \eta_i - 1 \) and \( 2^{\tilde{k}_i-2} \leq |S| \leq 2^{\tilde{k}_i-1} \).

Similarly, since \( C' \) is formed by concatenating the \( B_i' \) matrices vertically, \( i \in [r + 1] \), it follows that \( C' \) has at least \( p \tilde{\gamma}_i 2^i \) rows that satisfy (4.23) and (4.24) for \( \eta_r = \eta_i - 1 \) and \( |S| \in [\eta_i - 1, \eta_i - 1 - 1] \subseteq \bigcup_{i=0}^{r} [2^{\tilde{k}_i-2}, 2^{\tilde{k}_i-1}] \). Upon proving these results, one can reduce the rest of the proof to showing that \( C \) formed by performing bit-wise OR on the rows of \( C' \) and \( D \) according to the description in the statement of the theorem gives a \( [2; Q; \eta; (\eta_\alpha : d); e] \)-SQ-separable matrix.

Consider a set of column-indices \( S \) such that \( \eta_\alpha - 1 \leq |S| \leq d - 1 \). Since \( \eta_\alpha = 2^{\log_2(\eta_\alpha - 1)} \), one has

\[
\eta_\alpha'/2 = 2^{\log_2(\eta_\alpha - 1)} = \eta_\alpha - 1. \tag{4.25}
\]

In addition, since \( r = \lceil \log_2 ((d - 1)/\eta_\alpha') \rceil \), one also has

\[
2^r \eta_\alpha' = 2^{\log_2((d - 1)/\eta_\alpha')} \eta_\alpha' \geq 2^{\log_2((d - 1)/\eta_\alpha')} \eta_\alpha' = \frac{d - 1}{\eta_\alpha'} \eta_\alpha' = d - 1. \tag{4.26}
\]

Using inequalities (4.25) and (4.26), one obtains \( \eta_\alpha'/2 \leq \eta_\alpha - 1 \leq |S| \leq d - 1 \leq 2^r \eta_\alpha' \). Since \( \forall i \in [r + 1], \tilde{k}_i \) is chosen as \( \tilde{k}_i = \lceil \log_2(\eta_\alpha - 1) \rceil + i + 1 \), one has

\[
2^{\tilde{k}_i-2} = 2^{\log_2(\eta_\alpha - 1) - 1} = \eta_\alpha'/2 \leq |S| \leq 2^r \eta_\alpha' = 2^{\log_2(\eta_\alpha - 1)} + r = 2^{\tilde{k}_i-1}.
\]

This implies that for any set of column indices \( S \), where \( \eta_\alpha - 1 \leq |S| \leq d - 1 \), there exists an \( i \in [r + 1] \) for which \( 2^{\tilde{k}_i-2} \leq |S| \leq 2^{\tilde{k}_i-1} \). On the other hand, using Lemma 6 we know that \( \forall i \in [r + 1], B_i \) has at least \( p \tilde{\gamma}_i 2^i \) rows that satisfy (4.23) and (4.24) for \( \eta_r = \eta_i - 1 \) and \( 2^{\tilde{k}_i-2} \leq |S| \leq 2^{\tilde{k}_i-1} \), where \( \tilde{\gamma}_i = \max\{1, 2^{k_i-i}2^{\tilde{k}_i}/(10(\eta_i - 1))\} \).

In the first step of the construction, \( \forall i \in [r + 1] \), we formed \( B_i' \in [2]^{2^r-i}_{m \times n} \) by repeating \( B_i \) \( 2^{r-i} \) times according to the rule \( B_i' = [B_i^T, B_i^T, \ldots, B_i^T]^T \). As a result, \( \forall i \in [r + 1], B_i' \) has at least \( p \tilde{\gamma}_i 2^{i+r-i} \) rows that satisfy (4.23)
and (4.24) for \( \eta_r = \eta_i - 1 \) and \( 2^{k_i-2} \leq |S| \leq 2^{k_i-1} \). Since \( \forall i \in [r+1] \), one also has

\[
2^{r-i} \tilde{\alpha}_i = 2^{r-i} \max\{1, 2^{k_i-1} \}\frac{d-1}{5(\eta_i - 1)} \\
\geq \max\{1, 2^{k_i-1} \}\frac{d-1}{5(\eta_i - 1)} \\
\geq \max\{1, \frac{d-1}{5(\eta_i - 1)} \min_{i \in [r+1]} \{2^{k_i-1}\}\},
\]

then \( B'_i \) contains at least \( p2^{r} \tilde{\alpha} \) rows satisfying (4.23) and (4.24) for \( \eta_r = \eta_i - 1 \) and \( 2^{k_i-2} \leq |S| \leq 2^{k_i-1} \), where

\[
\tilde{\alpha} = \max\{1, \frac{d-1}{5(\eta_i - 1)} \min_{i \in [r+1]} \{2^{k_i-1}\}\}.
\]

This result, in addition to the fact that for any set of column indices \( S \) for which \( \eta_0 - 1 \leq |S| \leq d - 1 \), there exists a \( i \in [r+1] \) for which \( 2^{k_i-2} \leq |S| \leq 2^{k_i-1} \), implies that \( C' \) has at least \( e' = p2^{r} \tilde{\alpha} \) rows that satisfy

\[
\sum_{k \in S} C'(j,k) = \eta_i - 1, \quad (4.27)
\]
\[
\sum_{k \in N} C'(j,k) = 0, \quad (4.28)
\]

for any set \( S \) and \( N \), where \( \eta_0 - 1 \leq |S| \leq d - 1, |N| \leq |S| \) and \( S \cap N = \emptyset \).

In order for \( C \) to be a \([2;Q;\eta;e;\alpha,d,e]-SQ\)-separable matrix,\(^8\) we need to show that for any two distinct sets of columns, i.e. columns of \( C \), denoted by \( X_1 \) and \( X_2 \), for which \( \eta_0 \leq |X_2| \leq |X_1| \leq d \), one has \( y_{X_1} \neq y_{X_2} \). Note that this constraint is weaker than the conditions (4.14)-(4.16). Without loss of generality, we made the assumption that \( |X_2| \leq |X_1| \).

Let \( S_1 \) and \( S_2 \) be the set of column-indices corresponding to \( X_1 \) and \( X_2 \), respectively. Since \( S_1 \neq S_2 \) and \( |S_1| \geq |S_2| \), the set \( S_1 \setminus S_2 \) is nonempty. Let \( l \in S_1 \setminus S_2 \). Given that \( |S_2| \leq d \), it follows from the definition of binary \( d \)-disjunct matrices that for the set \( S_2 \cup \{l\} \) there exists a set of row indices of

\(^8\)Although this construction resembles the construction of threshold disjunct matrices in [60], one should notice that the matrix \( C' \) generated in Step 2 of Construction 7 is not a regular matrix (i.e. it is neither a \((d-1,e';\eta_0 - 1)\)-regular matrix, nor a \((d-1,e';\eta_0 - 1)\)-regular matrix). Consequently, [60, Lemma 6] cannot be used directly to show that \( C \) is a SQ-separable matrix.
\( \mathbf{D} \), denoted by \( \mathcal{R}_D \), with size at least \( 2e_1 + 1 \), such that

\[
\sum_{k \in S_2} \mathbf{D}(j, k) = 0, \quad \forall j \in \mathcal{R}_D, \quad (4.29)
\]

\[
\mathbf{D}(j, l) = 1, \quad \forall j \in \mathcal{R}_D. \quad (4.30)
\]

Let \( \mathcal{S} = \mathcal{S}_1 \setminus \{l\} \). Also, if \( \mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset \) and \( |\mathcal{S}_1| = |\mathcal{S}_2| \), let \( \mathcal{N} = \mathcal{S}_2 \setminus \{k_0\} \) where \( k_0 \) is an arbitrary column-index of \( \mathcal{S}_2 \). Otherwise, let \( \mathcal{N} = \mathcal{S}_2 \setminus \mathcal{S}_1 \). Clearly, \( |\mathcal{N}| \leq |\mathcal{S}|. \)

Next, let \( \mathcal{R}_{C'} \) be the set of row-indices of \( C' \) for which (4.27) and (4.28) are satisfied for the sets \( \mathcal{S} \) and \( \mathcal{N} \). Consider some \( i \in \mathcal{R}_{C'} \) and some \( j \in \mathcal{R}_D \). The \( (j + (i-1)m_d) \text{th} \) row of \( C \) is formed by finding the bit-wise OR of the \( i \text{th} \) row of \( C' \) and the \( j \text{th} \) row of \( D \). Consequently,

\[
\sum_{k \in \mathcal{S}_1} C(j + (i-1)m_d, k) = \sum_{k \in \mathcal{S}} C(j + (i-1)m_d, k) + C(j + (i-1)m_d, l)
\]

\[
= \eta_1 - 1 + 1 = \eta_1, \quad (4.31)
\]

\[
\sum_{k \in \mathcal{S}_2} C(j + (i-1)m_d, k) < \eta_1, \quad (4.32)
\]

where \( C(j + (i-1)m_d, l) = 1 \) follows from (4.30), and (4.32) is a consequence of the following argument. First, note that using (4.24) and (4.29), one has \( \sum_{k \in \mathcal{N}} C(j + (i-1)m_d, k) = 0 \). As a result, if \( \mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset \) and \( |\mathcal{S}_1| = |\mathcal{S}_2| \), then

\[
\sum_{k \in \mathcal{S}_2} C(j + (i-1)m_d, k) = \sum_{k \in \mathcal{N}} C(j + (i-1)m_d, k) + C(j + (i-1)m_d, k_0)
\]

\[
\leq 1 < \eta_1.
\]

Otherwise, one has

\[
\sum_{k \in \mathcal{S}_2} C(j + (i-1)m_d, k) = \sum_{k \in \mathcal{N}} C(j + (i-1)m_d, k) + \sum_{k \in \mathcal{S}_2 \cap \mathcal{S}_1} C(j + (i-1)m_d, k)
\]

\[
= \sum_{k \in \mathcal{S}_2 \cap \mathcal{S}_1} C(j + (i-1)m_d, k)
\]

\[
\leq \sum_{k \in \mathcal{S}_1 \setminus \{l\}} C(j + (i-1)m_d, k)
\]

\[
= \eta_1 - 1 < \eta_1.
\]

Since \( |\mathcal{R}_{C'}| \geq e' \) and \( |\mathcal{R}_D| \geq 2e_1 + 1 \), \( C \) has a set of row indices \( \mathcal{R} \), \( |\mathcal{R}| \geq \).
\[ e'(2e_1+1), \text{ for which (4.31) and (4.32) are satisfied. This implies that } \forall j \in \mathcal{R}, \quad y_{x_1}(j) > y_{x_2}(j), \quad \text{and therefore } C \text{ is a } [2; Q; \eta; (\eta_\alpha : d); e]-\text{SQ-separable matrix, where } e = \left[\frac{(2e_1+1)p2^{e'}-1}{2}\right]. \quad \text{Note that } C \text{ is an } m \times n \text{ matrix, where } n = 2^n, \text{ and}

\[ m = m_d \cdot \left(\sum_{i=0}^{r} 2^{r-i}m_i\right) \approx m_d \left(\sum_{i=0}^{r} 2^{2^{i+\log_2(\eta_1-1)+1}O_{\eta_1}\left(2^{(\eta_1-1)(\tilde{t}_i-k_i)}\right)}\right) \]

\[ = 2^t m_d (d-1) \eta_1 - 1 \frac{\eta_1 - 1}{\eta_\alpha - 1} \left(\sum_{i=0}^{r} O_{\eta_1}\left(2^{(\eta_1-1)(\tilde{t}_i-k_i)}\right)\right). \]

\[ \square \]

\textbf{Remark 12.} A comparison between the rate of the matrix described in Construction 7, denoted by } R_{SQ}, \text{ and the rate of the threshold disjunct matrix described in [60] for } \eta_T = \eta_1, \text{ denoted by } R_{TD}, \text{ reveals that}

\[ \frac{R_{SQ}}{R_{TD}} = \frac{\eta_\alpha - 1}{\eta_1 - 1}. \]

In order to compute this ratio, one needs to carefully calculate } R_{TD}, \text{ keeping track of the constant values that may be hidden in the asymptotic expressions. It turns out that if the same } d\text{-disjunct binary matrix } D \text{ is used in both constructions, } n_{SQ} = n_{TD}, \text{ and } m_{TD} = \frac{\eta_\alpha - 1}{\eta_1 - 1} m_{SQ}.

Different properties of this construction are summarized in Table 4.2. Note that the parameters of the BBTD matrices used by Construction 7 depend on the underlying lossless condenser. Different forms of condensers were discussed in [60], and we refer an interested reader to this paper for more information. For an asymptotic bound on the number of measurements } m \text{ obtained via Construction 7, we used the parameters and condensers outlined in Construction M8 of [60, Table 1].}
CHAPTER 5

TEST MATRIX CONSTRUCTIONS AND DECODING ALGORITHMS FOR SQGT WITH EQUIDISTANT THRESHOLDS

In Chapter 4, we described test matrix constructions and decoding algorithms for the general case of SQGT with arbitrary thresholds. In this chapter, we focus on the special case of SQGT with equidistant thresholds, i.e. $\eta_k = r\eta$ and $r \in [Q + 1]$, and use the relationship between the thresholds to construct test matrices that improve upon the results of previous chapter.

We will describe probabilistic and combinatorial constructions of SQ-disjunct and SQ-separable matrices, in addition to efficient decoding algorithms. We first consider the case in which the number of defectives is much smaller than the total number of items; then, we relax this condition and consider the case in which the number of defectives can be any number between 0 and $n$. Finally, we provide an approximate message passing algorithm for identification of defectives, for constructions for which no efficient decoding algorithm is known.

Throughout this chapter, we use the notation $[q; Q; \eta; (1:d); e]$ instead of $[q; Q; \eta; (1:d); e]$ for simplicity and to emphasize that the test matrices are constructed for SQGT with equidistant thresholds.

A summary of different properties of the constructions described in this chapter is provided in Table 5.1.
Table 5.1: A comparative summary of SQGT matrices described in Constructions 8-10

<table>
<thead>
<tr>
<th>Test Matrix</th>
<th>Construction 8</th>
<th>Construction 9</th>
<th>Construction 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>$[q; Q; \eta; (1:d); e]$</td>
<td>$[q; Q; \eta; d; 0]$</td>
<td>$[q; Q; \eta; (1:n); 0]$</td>
</tr>
<tr>
<td>Type</td>
<td>SQ-disjunct</td>
<td>SQ-separable</td>
<td>SQ-separable</td>
</tr>
<tr>
<td>Thresholds</td>
<td>Equidistant</td>
<td>Equidistant</td>
<td>Equidistant</td>
</tr>
<tr>
<td>Construction</td>
<td>Probabilistic</td>
<td>Explicit</td>
<td>Explicit</td>
</tr>
<tr>
<td>Num. Tests</td>
<td>$O\left(\frac{2n}{\pi^2} (d^2 \log_2 \frac{n}{d} + 2ed)\right)$</td>
<td>$O(d \log_d n)$</td>
<td>$\sim \frac{2n}{\log_2 n}$</td>
</tr>
<tr>
<td>Features</td>
<td>Efficient decoder of complexity $O(mn)$</td>
<td>Number theoretic (Bose-Chowla)</td>
<td>No restriction on $d$, Efficient decoder</td>
</tr>
</tbody>
</table>
5.1 Construction of $q$-ary SQ-disjunct matrices

The following lemma describes how to efficiently choose the entries of a $q$-ary SQ-disjunct matrix for SQGT with equidistant thresholds.

**Lemma 7.** Given a $[q;Q;\eta;(1:d);e]$-SQ-disjunct matrix $C \in [q]^{m \times n}$ exists, one can construct a $[q;Q;\eta;(1:d);e]$-SQ-disjunct matrix $C' \in [q]^{m \times n}$ that effectively uses only an $(I+1)$-ary alphabet, $\{0, \eta, 2\eta, \ldots, I\eta\}$, where $I = \lfloor \frac{q-1}{\eta} \rfloor$.

**Proof.** Form $C'$ by the following substitution: $\forall i \in [m]$ and $\forall j \in [n]$, let $C'(i,j) = \lfloor \frac{C(i,j)}{\eta} \rfloor \eta \in \{0, \eta, 2\eta, \ldots, I\eta\}$. Consider a set of $d+1$ column-indices $S$ and fix a column-index $l \in S$. If $C(i,l), i \in [m]$, is a unique coordinate of the $l$th column of $C$ for which (3.5) is satisfied for the given set $S$, the same condition will still be satisfied in $C'$ for $l$ and $S$. The reason is that after the substitution, the $i$th coordinate of the syndrome of the $l$th column remains unchanged, while the $i$th coordinate of the syndrome of the other $d$ columns indexed by $S\setminus\{l\}$ will have a smaller value. Since this is true for any $S \subseteq [n]$ with $|S| = d+1$ and for any $l \in S$, $C'$ is a $[q;Q;\eta;(1:d);e]$-SQ-disjunct matrix. On the other hand, if for $i \in [m]$, none of the columns of $C$ indexed by $S$ has a unique coordinate in the $i$th row, then this substitution may generate a unique coordinate in a column and therefore improve the error correcting capability of the matrix.

Lemma 7 implies that given an available alphabet $[q]$, in order to design a $[q;Q;\eta;(1:d);e]$-SQ-disjunct matrix with minimum number of rows $m$ for a fixed value of $n$, one only needs to use a $(I+1)$-ary alphabet, $\{0, \eta, 2\eta, \ldots, I\eta\}$, where $I = \lfloor \frac{q-1}{\eta} \rfloor$. We use this lemma to describe a probabilistic construction for SQ-disjunct matrices with equidistant thresholds.

**Theorem 14 (Construction 8).** Form a matrix $C \in \{0, \eta, 2\eta, \ldots, I\eta\}^{m \times n}$ by choosing each entry independently according to the probability distribution

$$P_X(x) = \begin{cases} P_0, & \text{if } x = 0 \\ P_1, & \text{if } x \in \{\eta, 2\eta, \ldots, I\eta\} \end{cases},$$

where $I = \lfloor \frac{q-1}{\eta} \rfloor$, $P_0 = \frac{d}{d+1}$, and $P_1 = \frac{1}{I(d+1)}$. Then $C$ is an $m_I \times n_I$ $[q;Q;\eta;(1:d);e]$-SQ-disjunct matrix with probability at least $1 - o(1)$; asymptotically, $m_I$
equals

\[ m_I \sim \frac{m_1}{\left(1 + \frac{1}{d-1} \sum_{k=0}^{d-1} \binom{d}{k} (Id)^k\right)^{1/(d-1)}} \]

where \( m_1 \) is the number of rows of a \([q; Q; \eta; (1:d); e]\)-SQ-disjunct matrix with \( n_1 = n_I \) columns, obtained by multiplying the best probabilistically constructed binary \( d \)-disjunct matrix, capable of correcting up to \( e \) errors, by \( \eta \).

**Proof.** Fix a choice of \( d+1 \) column indices, \( S \subseteq [n] \), and among them choose one index, \( l \in S \). There are \( \binom{n}{d+1} (d+1) \) ways to choose \( S \) and \( l \). Let \( \pi_I \) be the probability of “success” of a row, i.e., the probability that for a row of \( C \) denoted by \( r \), one has \( r(l) > \frac{1}{d} \sum_{i \in S \setminus \{l\}} r(i) \). Due to the fact that the alphabet consists of integer multiples of \( \eta \), the aforementioned condition is equivalent to

\[ r(l) > \sum_{i \in S \setminus \{l\}} r(i). \tag{5.1} \]

Let \( \mathcal{E}_\beta \) be the event that (5.1) is satisfied and that \( r(l) = \beta \eta \). From this definition, and the law of total probability, it follows that

\[ \pi_I = \Pr\left( \bigcup_{\beta=1}^{l} \mathcal{E}_\beta \right) = \sum_{\beta=1}^{l} \Pr(\mathcal{E}_\beta). \tag{5.2} \]

On the other hand, one has

\[ \Pr(\mathcal{E}_\beta) = P_1 \left( P_0^d + P_1^d \sum_{k=0}^{d-1} \binom{d}{k} \left( \frac{P_0}{P_1} \right)^k \left( \sum_{i=d-k}^{\beta-1} \binom{i-1}{d-k-1} \right) \right), \]

where \( \binom{i-1}{d-k-1} \) counts the number of compositions of \( i \) with \( d-k \) parts, or equivalently the number of positive integer solutions to \( \Sigma_{j=1}^{d-k} x_j = i \). Since

\[ \sum_{i=d-k}^{\beta-1} \binom{i-1}{d-k-1} = \binom{\beta-1}{d-k}, \]

\( \text{By “best”, we mean a matrix designed probabilistically in a way to have the minimum \( m \) for a fixed \( n \).} \)
equation (5.2) simplifies to

\[ \pi_I = \sum_{\beta=1}^{l} P_1 \left( P_0^d + P_1^d \sum_{k=0}^{d-1} \binom{d}{k} \left( \frac{P_0}{P_1} \right)^k \right) \]

\[ = IP_0 P_1^d + P_1^{d+1} \sum_{k=0}^{d-1} \binom{d}{k} \left( \frac{P_0}{P_1} \right)^k \sum_{\beta=2}^{I} \binom{I}{d-k-1} \]

\[ = IP_0 P_1^d + P_1^{d+1} \sum_{k=0}^{d-1} \binom{d}{k} \left( \frac{P_0}{P_1} \right)^k \left( \frac{P_0 I}{1-P_0} \right)^k \left( \frac{I}{d-k+1} \right). \]

(5.3)

Consequently, using the union bound, we can derive an upper bound on the probability that \( \mathbf{C} \) is not a \([q; Q; \eta; (1:d); 0]-\text{SQ-disjunct matrix}, \)

\[ P_F = \binom{n}{d+1} (d+1)(1-\pi_I)^m \leq \binom{n}{d+1} (d+1) \exp\left(-m\pi_I\right) \]

\[ \leq \exp\left((d+1) \log n - d \log(d+1) + d + 1 - m\pi_I\right). \]

As a result, for any \( \delta > 0 \), one has \( P_F = o(1) \) if

\[ m = \left( \frac{d+1}{\pi_I} + \delta \right) \log \frac{n}{d}. \]

This result can be generalized for \([q; Q; \eta; (1:d); e]-\text{SQ-disjunct matrices, where} \ e \ \text{is allowed to grow with} \ n. \) For a fixed \( \mathcal{S} \) and \( l \), \( \forall j \in [m] \), let \( N_j \) be a Bernoulli random variable with value 1 if the \( j^{th} \) row of \( \mathbf{C} \) satisfies (5.1), and 0 otherwise. By definition, the random variables \( N_j \) are independent identically distributed (i.i.d.) and \( \Pr(N_j = 1) = \pi_I \), for \( j \in [m] \). Based on the Chernoff bound, for \( 0 < \delta < 1 \), one obtains

\[ \Pr\left( \sum_{j=1}^{m} N_j \leq (1-\delta)m\pi_I \right) \leq \exp\left(-\frac{\delta^2 m\pi_I}{2}\right). \]

By setting \( \delta = 1 - \frac{2e}{m\pi_I} \), it follows that

\[ \Pr\left( \sum_{j=1}^{m} N_j \leq 2e \right) \leq \exp\left(-\frac{m\pi_I}{2} \left(1 - \frac{2e}{m\pi_I}\right)^2\right), \]

which provides an upper bound on the probability that for a fixed \( \mathcal{S} \) and \( l \), at most \( 2e \) rows of \( \mathbf{C} \) satisfy (5.1). As a result, the probability that \( \mathbf{C} \) is not
a \([q; Q; \eta; (1:d); e]\)-SQ-disjunct matrix is upper bounded by

\[
P_F \leq \left( \frac{n}{d+1} \right)(d+1) \exp \left( -\frac{m\pi_I}{2} \left( 1 - \frac{2e}{m\pi_I} \right)^2 \right)
\]

\[
\leq \exp \left( (d+1) \log n + d + 1 - d \log(d+1) - \frac{m\pi_I}{2} - \frac{2e^2}{m\pi_I} + 2e \right).
\]

It can be easily seen that for any \(\delta > 0\), \(P_F = o(1)\) if

\[
m = \left( \frac{2(d+1)}{\pi_I} + \delta \right) \log \frac{n}{d} + \frac{4e}{\pi_I}.
\]

We can compare the number of tests \(m_I\) for a matrix constructed using this method with the number of tests \(m_1\) in a matrix constructed by multiplying a conventional binary \(d\)-disjunct matrix with \(\eta\) (Construction 1), provided that they have the same number of columns \(n\). It can be easily verified – see for example [7] – that for a fixed \(n\), the distribution \(P_X(x)\) that minimizes the number of tests of a conventional binary \(d\)-disjunct matrix is the one that assigns \(P_0 = \frac{d}{d+1}\) to \(x = 0\) and \(P_1 = \frac{1}{d+1}\) to \(x = 1\). Consequently, \(\pi_1 = \frac{d}{(d+1)^{d+1}}\) maximizes the probability of “success” of a row.\(^2\) Since Construction 1 does not change the dimensions of the underlying binary \(d\)-disjunct matrix, asymptotically it holds that

\[
\frac{m_I}{m_1} \sim \frac{\pi_1}{\pi_I}, \quad (5.4)
\]

On the other hand,

\[
\pi_I = \pi_1 + \gamma_I,
\]

where \(\gamma_I = \frac{1}{\pi_I \pi_I^d} \sum_{k=0}^{d-1} \binom{d}{k} (d-k+1) (Id)^k\). Consequently,

\[
\lim_{n \to \infty} \frac{m_1}{m_I} = 1 + \frac{1}{Id^d} \sum_{k=0}^{d-1} \binom{d}{k} (d-k+1) (Id)^k.
\]

\(\Box\)

Figure 5.1 shows the asymptotic reduction in the number of tests, \(\frac{m_1}{m_I}\), as a function of \(I\) for different values of \(d\). Note that in this theorem, we assumed

\(^2\)Note that even though \(\pi_1\) is the optimal probability of success of a row when \(q-1 < 2\eta\), the same statement does not necessarily hold for \(\pi_I\) found in this construction.
that $I$ and $d$ do not grow with $n$. However, we can also consider the case in which $d \to \infty$ (for a fixed value of $I$) to obtain

\[
\lim_{d \to \infty} \lim_{n \to \infty} \frac{m_1}{m_I} = \lim_{d \to \infty} \left(1 + \frac{1}{I^{d+1}d^d} \sum_{k=0}^{d-1} \binom{d}{k} (d - k + 1)(Id)^k\right)
\]

\[
= \lim_{d \to \infty} \left(1 + \frac{1}{I^{d+1}d^d} \sum_{k=d-1}^{d-1} \binom{d}{k} (d - k + 1)(Id)^k\right)
\]

\[
= \lim_{d \to \infty} \left(1 + \frac{1}{I^{d+1}d^d} \sum_{k=0}^{I-2} \binom{I}{k} \frac{1}{I-k} \frac{d}{d^{I-k-1}}\right)
\]

\[
= 1 + \sum_{k=0}^{I-2} \binom{I}{k} \frac{1}{I^{I-k}} \lim_{d \to \infty} \left(\frac{d}{d^{I-k-1}}\right)
\]

\[
= 1 + \sum_{k=0}^{I-2} \binom{I}{k} \frac{1}{I^{I-k}} \frac{1}{(I-k-1)!},
\]

where we changed the order of the limit and the summation operations, since the sum was over a finite number of terms.

**Remark 13.** It is worth mentioning that instead of setting $P_0 = \frac{d}{d+1}$, one can
Figure 5.2: Reduction in the number of tests of a SQ-disjunct matrix constructed based on Construction 8 for the optimum choice of $P_0$. The parameter $u$, as before, denotes a known upper bound on the number of defectives.

Consider $P_0$ to be a parameter that may be optimized so as to minimize the number of tests in the matrix. Making this change does not affect the validity of (5.3) and (5.4), but it may increase the ratio $\frac{m_1}{m_I}$. Although finding a simple closed-form expression for the maximum $\pi_I$ over $P_0$ does not seem to be straightforward, we evaluated (5.3) numerically to find the maximum probability of “success” of a row. The resulting ratio $\frac{m_1}{m_I}$ is shown in Fig. 5.2 as a function of $I$, for different values of $d$.

5.2 Construction of $q$-ary SQ-separable matrices

In the previous chapter, we described different methods of constructing $q$-ary SQ-separable matrices for an arbitrary set of thresholds, using classical binary separable matrices for CGT and new sequences. On the other hand, in the case of SQGT with equidistant thresholds, SQ-separable matrices are
closely related to separable matrices for the additive model (QGT) defined below.

**Definition 17 (Binary $d$-separable matrices for QGT).** An $m \times n$ binary $d$-separable matrix designed for QGT, capable of correcting up to $e$ errors, is a matrix such that for any two distinct sets of columns $\mathcal{X}$ and $\mathcal{Z}$, $1 \leq |\mathcal{X}|, |\mathcal{Z}| \leq d$, the arithmetic sum of the columns in $\mathcal{X}$ differs from the arithmetic sum of the columns in $\mathcal{Z}$ in at least $2e + 1$ coordinates.

Similar to Construction 2, one can use $C_b$, a binary $d$-separable matrix for QGT capable of correcting up to $e$ errors, in order to form $C = (q-1)C_b$, where $q-1 \in \{\eta, 2\eta, \ldots\}$. Then $C$ represents a $[q; Q; \eta; (1:d); e]$-SQ-separable matrix.

An interesting matrix design for the additive model is the construction by Lindström, described in [36, Theorem 8]. In his approach, Lindström used a theorem by Bose and Chowla in additive number theory [62] to construct binary matrices for an adder channel. Multiplying this matrix with $\eta$ results in an $m \times n [q; Q; \eta; d; 0]$-SQ-separable matrix, where $m = [d \log_q L]$ and $L$ is a power of a prime such that $n \leq L$. A similar idea can be used to further improve the rate of SQ-separable matrices for equidistant SQGT. The idea is based on a result, proved in [62], that shows that if $L$ is power of a prime, there exist $L$ nonzero integers smaller than $L^d$ such that the sums of any $d$ such integers, i.e., their $d$-sums, are all distinct modulo $L^d - 1$.

**Theorem 15 (Construction 9).** Let $L$ be a power of a prime such that $n \leq L$; also, let $q' = \lfloor \frac{2-1}{\eta} \rfloor + 1$. Using the construction in [62], find $L$ non-zero integers with distinct $d$-sums. Let the $q'$-ary representation of these integers serve as columns of a matrix $C_{q'}$. Form the $m \times L$ matrix $C = \eta C_{q'}$, where $m = [d \log_{q'} L]$. A matrix obtained by choosing any $n$ columns of $C$ is an $m \times n [q; Q; \eta; d; 0]$-SQ-separable matrix.

**Proof.** We only need to show that $C_{q'}$ is capable of identifying $d$ defectives in an adder model. Assume that there exist two sets of $d$ columns $\mathcal{X} = \{x_i\}_{i=1}^d$ and $\mathcal{Z} = \{z_j\}_{j=1}^d$ such that $|\mathcal{X} \cap \mathcal{Z}| < d$, and $\sum_{i=1}^d x_i = \sum_{j=1}^d z_j$. Consequently, $\forall k \in [m]$, $\sum_{i=1}^d x_i(k) = \sum_{j=1}^d z_j(k)$. Then,

$$\sum_{k=1}^m \left( \sum_{i=1}^d x_i(k) \right) q^{k-1} = \sum_{k=1}^m \left( \sum_{j=1}^d z_j(k) \right) q^{k-1},$$

68
which implies that there exist two sets of \( d \) integers with the same sum. This contradicts the assumptions behind the construction of \( C_{q'} \) and completes the proof.

5.3 Construction of SQ-separable matrices for arbitrary number of defectives

The constructions described up to this point are able to identify up to \( d \) defectives in a pool of \( n \) subjects whenever \( d \) is significantly smaller than \( n \), say \( d = o(n) \) or \( d \) constant. It is also of interest to address the same questions when \( d \) is not constrained in size, so that one allows \( 0 \leq d \leq n \). This “dense” testing regime may be of use whenever no bound on the number of defectives is known a priori or when the number of defectives is inherently large.

In [36], Lindström described a binary construction for the adder model capable of identifying up to \( n \) defectives. In the next theorem we describe a generalization of this construction that employs a \( q \)-ary alphabet; using this generalization, we construct a SQ-separable matrix capable of identifying up to \( n \) defectives in an equidistant SQGT model. Extensions of [36] to a \( q \)-ary alphabet were also addressed in [33]. Multiplying these matrices with \( \eta \) results in a SQ-separable matrix with the same rate as our construction. But unlike our direct and very simple approach, the methods of [33] and [34] may only be used in a recursive and rather complicated manner.

Before describing our construction, we state a lemma from [36] that will be useful in proving the next theorem.

**Lemma 8.** Let \( \mathcal{F} \) be a collection of sets such that if \( B \in \mathcal{F} \), then \( \mathcal{F} \) contains all the subsets of \( B \) as well. In other words, \( \forall B \in \mathcal{F} \), if \( A \subseteq B \), then \( A \in \mathcal{F} \). Let \( g : \mathcal{F} \mapsto \{0,1\} \) be a function defined on \( \mathcal{F} \) such that for some fixed set \( S \in \mathcal{F} \), one has \( g(A \cap S) = g(A) \) whenever \( A \in \mathcal{F} \). If \( C \in \mathcal{F} \) and \( C \notin S \), then

\[
\sum_{\substack{A \subseteq C \\
|A| \text{ is odd}}} g(A) = \sum_{\substack{A \subseteq C \\
|A| \text{ is even}}} g(A).
\]

**Proof.** See [36].

**Theorem 16 (Construction 10).** Let \( \kappa \in \mathbb{Z}^+ \) and \( m = 2^\kappa - 1 \). Consider the set \( [\kappa] \) and label each of its non-empty subsets by \( S_i, i \in [m] \), such that for
any two subsets $S_{i_1}, S_{i_2} \subseteq [\kappa]$, the inequality $|S_{i_1}| < |S_{i_2}|$ implies $i_1 < i_2$. Let $q' = \lfloor \log_2 \frac{Q - 1}{\eta} \rfloor + 1$ and $q'' = \lfloor \log_2 \frac{Q - 1}{\eta} \rfloor$; for each $S_i$, form a matrix $C_i \in [q']^{m \times (q'' + |S_i|)}$ as follows. For $j \in [m]$ and $k \in [q'' + 1]$, set

$$C_i(j, k) = \begin{cases} 2^{q''-k+1}, & \text{if } |S_i \cap S_j| \text{ is odd,} \\ 0, & \text{if } |S_i \cap S_j| \text{ is even.} \end{cases} \quad (5.5)$$

Let $T_{i,q''+1} = S_i$. For $k \in \{q'' + 2, q'' + 3, \ldots, q'' + |S_i|\}$, fix any $T_{i,k} \subset T_{i,k-1}$ of size $|T_{i,k}| = |S_i| - k + q'' + 1$. Set

$$C_i(j, k) = \begin{cases} 1, & \text{if } C_i(j, k - 1) > 0 \text{ and } |S_j \cap T_{i,k}| \text{ is odd,} \\ 0, & \text{otherwise,} \end{cases} \quad (5.6)$$

where $j \in [m]$. Form a matrix $C' = \eta C$ where $C = [C_1, C_2, \ldots, C_m]$. The matrix $C'$ is an $m \times n [q; Q; \eta; (1:n); 0]$-SQ-separable matrix, where $m = 2^{\kappa} - 1$ and $n = \kappa 2^{\kappa-1} + q''(2^\kappa - 1)$.

**Proof.** As before, we define $w \in [2]^n$ to be a binary vector such that its $l^{th}$ coordinate is equal to 1 if the $l^{th}$ subject is defective, and 0 otherwise. From the construction, the matrix $C$ is formed from $m$ sub-matrices $C_i$, each corresponding to a subset of $[\kappa]$, $S_i$. This implies that each $S_i$ corresponds to a set of variables, i.e. coordinates of $w$. In addition, we label rows of $C$ using subsets $S_i$, $i \in [m]$, such that the $i^{th}$ row is labeled by $S_i$. Since each row of $C$ corresponds to an equation in $y = Cw$, each $S_i$ corresponds to exactly one equation.

The decoding includes $m$ steps, and in each step one solves for the variables corresponding to $S_i$, given all the variables corresponding to $S_{i+1}, S_{i+2}, \ldots, S_m$. To find the variables corresponding to $S_i$, we form two equations: the first equation is obtained by adding all the equations corresponding to the odd subsets of $S_i$ while the second equation is obtained by adding all the equations corresponding to the even subsets of $S_i$. These two equations can be represented by $s_{\text{odd}_i}^T w = y_{\text{odd}},$ and $s_{\text{even}_i}^T w = y_{\text{even}_i},$ respectively. Finally, we form the equation

$$(s_{\text{odd}_i} - s_{\text{even}_i})^T w = y_{\text{odd}_i} - y_{\text{even}_i}. \quad (5.7)$$

For simplicity, let $w_{i_k}$ be the $k^{th}$ variable corresponding to $S_i$, where $k \in [q'' + |S_i|]$. The key in the proof of the theorem is to show that (5.7) is of the
where $a$ is a scalar that depends on $y$ and the known variables corresponding to $S_{i+1}, S_{i+2}, \ldots, S_m$. This implies that all the coefficients of the variables corresponding to $S_1, S_2, \ldots, S_{i-1}$ are zero; also, given that $w_{i_k} \in [2]$ for all $k \in \llbracket q''+|S_i| \rrbracket$, the unknown variables can be determined by finding the unique binary representation of $a$. Note that the coefficient of the variable $w_{i_l}$, $l \leq i$, in the aforementioned expression equals

$$
\sum_{j: S_j \subseteq S_i, |S_j| \text{ is odd}} C_l(j, k) - \sum_{j: S_j \subseteq S_i, |S_j| \text{ is even}} C_l(j, k).
$$

We now show that $\forall l < i$, the coefficients of the variables in $S_l$ of (5.7) are all zero. Although Lemma 8 cannot be directly applied to our problem since the matrix $C$ is not binary, we make use of this lemma in our proof as follows.

Let $F = \{ S \}^m_1$; this set satisfies the condition of Lemma 8. Let $l < i$, due to the specific ordering of the elements of $F$, we have $S_i \notin S_l$, and can consequently set $C = S_i$ and $S = S_l$. Consider the $k^{th}$ column of $C_l$, where $k \in \llbracket q''+1, q''+2, \ldots, q''+|S_i| \rrbracket$. For this column, let $g_{l,k}(S_j) = C_l(j, k)$. Careful inspection shows that $g_{l,k}(S_j \cap S_l) = g_{l,k}(S_j)$, $\forall j \in \llbracket m \rrbracket$, and $g_{l,k}(\cdot) \in \{0, 1\}$. Using Lemma 8, we conclude that

$$
\sum_{j: S_j \subseteq S_i, |S_j| \text{ is odd}} g_{l,k}(S_j) = \sum_{j: S_j \subseteq S_i, |S_j| \text{ is even}} g_{l,k}(S_j).
$$

Next, consider the $k^{th}$ column of $C_l$, where $k \in \llbracket q'' \rrbracket$. For this column, let
\[ g_{l,k}(S_j) = C_l(j, k). \] Since \[ g_{l,k}(S_j) = 2q''^{k+1}g_{l,q''+1}(S_j), \] using (5.8) one obtains

\[
\sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} g_{l,k}(S_j) = 2q''^{k+1} \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} g_{l,q''+1}(S_j)
\]

\[
= 2q''^{k+1} \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is even}} g_{l,q''+1}(S_j)
\]

\[
= \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} g_{l,k}(S_j).
\]

As a result, \( \forall l < i \) and \( k \in \lbrack q'' + \lvert \mathcal{S}_i \rvert \rbrack \) one has

\[
\sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} C_l(j, k) - \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is even}} C_l(j, k) = 0. \quad (5.9)
\]

To complete the proof, consider the \( k^{\text{th}} \) column of \( C_i \), where \( k \in \lbrack q'' + 1 \rbrack \). Since (5.7) is formed using the rows labeled by odd and even subsets of \( \mathcal{S}_i \), the coefficient of \( w_{ik} \) is equal to

\[
\sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} C_i(j, k) - \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is even}} C_i(j, k) = 2q''^{k+1} \cdot 2^{\lvert \mathcal{S}_i \rvert - 1} - 0 = 2q''^{\lvert \mathcal{S}_i \rvert - k}, \quad (5.10)
\]

where \( 2^{\lvert \mathcal{S}_i \rvert - 1} \) is the number of odd subsets of \( \mathcal{S}_i \). Next, consider the \( k^{\text{th}} \) column of \( C_i \), where \( k \in \{ q'' + 2, q'' + 3, \ldots, q'' + \lvert \mathcal{S}_i \rvert \} \). From the definition of \( \mathcal{T}_{i,k} \) and its relationship to \( \mathcal{T}_{i,k-1} \), it can be shown that the coefficient of the variable \( w_{ik} \) equals

\[
\sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} C_i(j, k) - 0 = \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} C_i(j, k)
\]

\[
= \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} 1 [\{ \lvert S_j \cap \mathcal{T}_{i,q''+2} \rvert \text{ odd} \} \cap \cdots \cap \{ \lvert S_j \cap \mathcal{T}_{i,k-1} \rvert \text{ odd} \} \cap \{ \lvert S_j \cap \mathcal{T}_{i,k} \rvert \text{ odd} \}]
\]

\[
= \frac{1}{2} \sum_{j: \lvert S_j \rvert \leq \mathcal{S}_i \text{ is odd}} 1 [\{ \lvert S_j \cap \mathcal{T}_{i,q''+2} \rvert \text{ odd} \} \cap \cdots \cap \{ \lvert S_j \cap \mathcal{T}_{i,k-1} \rvert \text{ odd} \}] = \cdots = 2q''^{\lvert \mathcal{S}_i \rvert - k}.
\]

\[
(5.11)
\]
Using (5.9), (5.10), and (5.11), one can write (5.7) in the form

\[ q'' + |S_i| \sum_{k=1}^{q''+|S_i|} w_{ik} = a, \]

where \( a \) depends on \( y \) and the known variables corresponding to \( S_{i+1}, S_{i+2}, \ldots, S_m \).

This completes the proof of the claimed result. \( \square \)

As an example, let \( \kappa = 3, \eta = 2, \) and \( q = 5; \) consequently, \( m = 7, q' = 9, \) and \( q'' = 2. \) We label the non-empty subsets of \([3]\) as follows: \( S_1 = \{1\}, S_2 = \{2\}, S_3 = \{3\}, S_4 = \{1,2\}, S_5 = \{1,3\}, S_6 = \{2,3\}, S_7 = \{1,2,3\}. \) In construction \( C_7, \) corresponding to \( S_7, \) fix \( T_{7,4} = \{1,2\} \) and \( T_{7,5} = \{1\}. \)

Based on (5.5) and (5.6), one has

\[
C_7 = \begin{pmatrix}
4 & 2 & 1 & 1 & 1 \\
4 & 2 & 1 & 1 & 0 \\
4 & 2 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
4 & 2 & 1 & 0 & 0 \\
\end{pmatrix}.
\]

Using (5.5) and (5.6), we obtain

\[
C = \begin{pmatrix}
S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 \\
S_1 & 4 & 2 & 1 & 0 & 0 & 0 \\
S_2 & 0 & 0 & 0 & 4 & 2 & 1 \\
S_3 & 0 & 0 & 0 & 0 & 4 & 2 \\
C = S_4 & 4 & 2 & 1 & 0 & 0 & 0 \\
S_5 & 4 & 2 & 1 & 0 & 0 & 0 \\
S_6 & 0 & 0 & 0 & 0 & 0 & 0 \\
S_7 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}.
\]

In order to prove that \( C' = 2C \) is a SQ-separable matrix, we only need to show that \( C \) is a separable matrix for an adder model.

\(^3\)Note that there exist other choices for \( T_{7,4} \) and \( T_{7,5} \) that provide for valid matrix constructions.
Let \( w \in [2]^n \) be a binary vector such that its \( l \)th coordinate is equal to 1 if the \( l \)th subject is defective and 0 otherwise. In the adder model, the vector of test results equals \( y = Cw \), which is a system of linear equations with \( n \) variables and \( m \) equations. Note that each set \( S_i \) corresponds to \( q'' + |S_i| \) variables.

We solve the system of equations in a recursive manner, by first solving for variables corresponding to \( S_m \), subtracting their effect on the syndrome and then solving for variables corresponding to \( S_{m-1} \), and so on.

Returning to our example, we can solve for the variables corresponding to \( S_7 \) as follows. Add all the equations corresponding to odd subsets of \( S_7 \). The result is an equation of the form

\[
(s_{\text{odd}}^T) w = y(1) + y(2) + y(3) + y(7),
\]

where

\[
s_{\text{odd}} = (8 4 2 8 4 2 8 4 2 8 4 2 1 8 4 2 1 8 4 2 1 16 8 4 2 1) \nonumber .
\]

Also, add all the equations corresponding to even subsets of \( S_7 \). The result is an equation of the form

\[
(s_{\text{even}}^T) w = y(4) + y(5) + y(6),
\]

where

\[
s_{\text{even}} = (8 4 2 8 4 2 8 4 2 8 4 2 1 8 4 2 1 8 4 2 1 0 0 0 0 0 0) \nonumber .
\]

Since the first 21 entries of \( s_{\text{odd}} \) and \( s_{\text{even}} \) are identical, one has

\[
(s_{\text{odd}} - s_{\text{even}})^T w = 16w(22) + 8w(23) + 4w(24) + 2w(25) + w(26)
\]

\[
= y(1) + y(2) + y(3) + y(7) - y(4) - y(5) - y(6). \quad (5.14)
\]

The equation in (5.14) provides a binary representation of the integer \( y(1) + y(2) + y(3) + y(7) - y(4) - y(5) - y(6) \). Therefore, the variables \( w(22) \), \( w(23) \), \( w(24) \), \( w(25) \), and \( w(26) \) are uniquely determined by the equation. Now, given these variables, one can add all the equations corresponding to odd and even subsets of \( S_6 \) to similarly identify \( w(18) \), \( w(19) \), \( w(20) \), and
This process can be applied iteratively until all the variables are uniquely determined.

**Remark 14.** Construction 10 provides matrices capable of identifying any number of defectives among \( n = \kappa 2^{\kappa - 1} + q''(2^\kappa - 1) \) subjects, using \( m = 2^\kappa - 1 \) experiments. It can be easily shown that the same approach applies for an arbitrary number of subjects. For a fixed value of \( q'' \), one can find the smallest number \( \kappa \) such that \( n \leq \kappa 2^{\kappa - 1} + q''(2^\kappa - 1) \). Removing the \((\kappa 2^{\kappa - 1} + q''(2^\kappa - 1) - n)\) right most columns of \( C' \) in Construction 10 results in an \( m \times n \) SQ-separable matrix, where \( m = 2^\kappa - 1 \).

Different properties of this construction are summarized in Table 5.1. Note that the underlying assumption in all our constructions is that \( q \) is fixed and does not grow with \( n \). For example, in Construction 10, we have \( m = 2^\kappa - 1 \) and \( n = \kappa 2^{\kappa - 1} + q''(2^\kappa - 1) \), where \( q'' = \left\lfloor \log_2 \frac{q-1}{\eta} \right\rfloor \).

Consequently, \( n = 1/2(m+1) \log_2(m+1) + q''m \), and for \( q'' = o(\log_2 m) \), one has

\[
\lim_{\kappa \to \infty} \frac{m}{2n/\log_2 n} = \lim_{m \to \infty} \frac{\log_2 \left(1/2(m+1) \log_2(m+1) + q''m\right)}{\log_2(m+1) + 2q'' + \frac{\log_2(m+1)}{m}} = \lim_{m \to \infty} (1 + o(1)) = 1.
\]

On the other hand, if \( q = \eta 2^{\kappa \alpha} \), for some fixed \( \alpha > 0 \), similar calculations reveal that

\[
m \sim \left(\frac{2}{1 + 2\alpha}\right) \frac{n}{\log_2 n}.
\]

In addition, if \( q \) grows faster than exponential with \( \kappa \) (or equivalently, \( q'' \) grows faster than logarithmic with \( m \)), then \( m \sim \frac{1}{q''} n \).

### 5.4 Belief propagation decoders for SQGT

The SQ-disjunct matrices, as well as many of the SQ-separable matrices described in this and the previous chapter have efficient decoding algorithms based on algebraic methods. On the other hand, for constructions for which no efficient decoding algorithm is known, we can consider a different approach. More specifically, since most proposed SQGT matrices are sparse, methods based on belief propagation (BP) [63] emerge as a viable decoding
option for these matrices. In this section, we focus on BP decoders suitable for SQGT matrices based on probabilistic constructions (such as Constructions 6 and 8). The theoretical guarantees for these matrices hold in the asymptotic domain, and when the number of subjects is small, these guarantees may not apply. Nevertheless, in what follows, we show that BP decoders perform reasonably well even for a small number of subjects and large coding rates, and their performance may be further improved by tailoring the SQGT constructions to the decoder.

BP is an iterative message passing algorithm for inference on graphical models, and it is centered around calculating the marginal distributions of the variables corresponding to the vertices of the underlying graph. BP decoding for binary disjunct matrices was originally proposed by one of the authors in [64]. Later on, BP decoding was also considered in [65] for CGT decoding. Motivated by these two methods, we propose a BP decoder for SQGT, which performs an approximate bitwise maximum a posteriori (MAP) decoding of SQGT matrices in the presence of errors. Note that BP decoding can be used for different error models and assumptions; however, in the rest of this chapter, we focus on the following model.

Consider a SQGT model with thresholds $\eta$ as defined in Chapter 2. Assume that each subject is defective with probability $d/n$ independent of other subjects. Note that one consequence of this assumption is that the number of defectives $|D|$ is a random variable. Consider a set of $n$ subjects and let $W \in [2]^n$ be a random vector representing the incidence vector of defectives. Also, let $w_t \in [2]^n$ denote the true incidence vector of defectives, i.e. the realization of $W$ that we want to reconstruct. Also, let $C \in [q]^{m \times n}$ and $z \in [Q]^m$ be the test matrix and the observed vector of (possibly) erroneous test results, respectively.

The messages passed in a BP decoder depend on the message error model. We focus on one simple substitution error model for the test results. Let $Y \in [Q]^m$ and $Z \in [Q]^m$ be the random vectors corresponding to the error-free test results and the erroneous test results, respectively. We model the effect of false positives and false negatives using two probabilities, $\gamma_p$ and $\gamma_n$, respectively. In other words, for the $t$th test, if $Y(t) \in \{1, 2, \ldots, Q-2\}$ then $Z(t) = Y(t)$ with probability $1 - \gamma_p - \gamma_n$, $Z(t) = Y(t) + 1$ with probability $\gamma_p$, and $Z(t) = Y(t) - 1$ with probability $\gamma_n$. If $Y(t) = 0$ then $Z(t) = Y(t)$ with probability $1 - \gamma_p$ and $Z(t) = Y(t) + 1$ with probability $\gamma_p$. Finally, if
where the last equality follows by marginalizing out all the \( w \) from \( W = w \) in (5.16) yields

\[
\hat{w}_{\text{MAP}}(i) = \arg \max_{w(i) \in \{0,1\}} P_{W(i)|Z}(w(i)|z),
\]

(5.15)

where \( P_{W(i)|Z}(\cdot) \) denotes the conditional probability distribution of \( W(i) \) given \( Z \). Henceforth, we use \( P(\cdot) \) as a generic symbol for probability distribution and for simplicity, do not explicitly display the random variables in the subscript of \( P(\cdot) \).

Using the definition of conditional probability, \( P(w(i)|z) = \frac{P(z,w(i))}{P(z)} \). Since the maximization in (5.15) is performed over different values of \( w(i) \), the value of \( P(z) \) does not affect \( \hat{w}_{\text{MAP}}(i) \). For a function \( f(w) : [2]^n \rightarrow \mathbb{R} \), let the sum of \( f(w) \) over all configurations of the variables other than \( w(i) \) be denoted by \( \sum_{\sim w(i)} f(w) \). In this case, one has

\[
\hat{w}_{\text{MAP}}(i) = \arg \max_{w(i) \in \{0,1\}} P(w(i)|z)
= \arg \max_{w(i) \in \{0,1\}} P(z,w(i))
= \arg \max_{w(i) \in \{0,1\}} \sum_{\sim w(i)} P(z,w),
\]

(5.16)

where the last equality follows by marginalizing out all the \( w(j) \)’s, \( j \neq i \), from \( P(z,w) \).

Since the result of the tests are independent of each other conditioned on \( W = w \), it holds that \( P(z|w) = \prod_{t=1}^n P(z(t)|w) \). Substituting this equality in (5.16) yields

\[
\hat{w}_{\text{MAP}}(i) = \arg \max_{w(i) \in \{0,1\}} \sum_{\sim w(i)} \prod_{t=1}^n P(z(t)|w) P(w)
= \arg \max_{w(i) \in \{0,1\}} \sum_{\sim w(i)} \prod_{t=1}^m P(z(t)|w) \prod_{j=1}^n P(w(j))
\]

77
where the last equality follows since we assumed that the event that a subject is defective is independent of the event of other subjects being defective. Finally, given that each subject is defective with probability $d/n$, one obtains

$$
\hat{w}_{\text{MAP}}(i) = \arg \max_{w(i) \in \{0, 1\}} \sum_{w(i)} \left[ \prod_{t=1}^{m} P(z(t)|w) \cdot \prod_{j=1}^{n} \left( \frac{d}{n} I(w(j) = 1) + \left( 1 - \frac{d}{n} \right) I(w(j) = 0) \right) \right],
$$

(5.17)

where $I(\cdot)$ denotes the indicator function, equal to 1 if the statement in the brackets holds, and equal to 0 otherwise.

Using (5.17), we can form a factor graph that corresponds to the bitwise MAP estimator with $n$ variable nodes and $m$ factor nodes; a factor node corresponding to test $t$ is only connected to variable nodes corresponding to subjects present in the $t^{th}$ test. Similarly, a variable node corresponding to the $i^{th}$ subject is only connected to the factor nodes corresponding to the tests in which the $i^{th}$ subject is used. As a result, the complexity of the BP decoder depends on the sparsity of the test matrix, $C$.

Let $\mathcal{N}(t)$ denote the neighbors of the factor node corresponding to test $t$ in the factor graph. Also, let $\mathcal{N}(i)$ denote the neighbors of the variable node corresponding to the $i^{th}$ subject. Let $\chi_{i \rightarrow t}^{(l)}(w(i))$ denote the message from the $i^{th}$ variable node to the $t^{th}$ factor node in the $l^{th}$ iteration, $1 \leq l \leq L$. Similarly, let $\hat{\chi}_{t \rightarrow i}^{(l)}(w(i))$ denote the message at the $l^{th}$ iteration from the $t^{th}$ factor node to the $i^{th}$ variable node. The BP message update rules for finding the marginal distributions of each subject according to the MAP estimator of (5.17) take the form:

$$
\chi_{i \rightarrow t}^{(l+1)}(w(i)) \propto \left( \frac{d}{n} I(w(i) = 1) + \left( 1 - \frac{d}{n} \right) I(w(i) = 0) \right) \prod_{\tau \in \mathcal{N}(i) \setminus \{t\}} \hat{\chi}_{\tau \rightarrow i}^{(l)}(w(i)),
$$

(5.18)

and

$$
\hat{\chi}_{t \rightarrow i}^{(l+1)}(w(i)) \propto \sum_{w(i)} \left[ P(z(t)|w) \prod_{j \in \mathcal{N}(t) \setminus \{i\}} \chi_{j \rightarrow t}^{(l)}(w(j)) \right],
$$

(5.19)

where $\propto$ denotes “equal up to a multiplicative constant”. For an in-depth explanation regarding message updates for marginals of a distribution, we
refer the interested reader to [63] and the references therein.

In order to get an explicit form for the message updates, we need to calculate the term \( P(z(t)|w) \) in (5.19) for different values of \( z(t) \). For this purpose, let \( w_i := \sum_{l=1}^n w(l) C(t,l) \). Then, one has

\[
P_{Z(t)|w}(0|w) = \begin{cases} 
\gamma_n \left( \eta_1 \leq w_i < \eta_2 \right) + (1 - \gamma_p) \left( \omega_i < \eta_1 \right), & \text{if } w(i) = 0, \\
\gamma_n \left( \eta_1 - C(t,i) \leq \omega_i < \eta_2 - C(t,i) \right) + (1 - \gamma_p) \left( \omega_i < \eta_1 - C(t,i) \right), & \text{if } w(i) = 1,
\end{cases}
\]

\[
P_{Z(t)|w}(Q-1|w) = 
\begin{cases} 
(1 - \gamma_n) \left( \eta_{Q-1} \leq \omega_i < \eta_Q \right) + \gamma_p \left( \eta_{Q-2} \leq \omega_i < \eta_{Q-1} \right), & \text{if } w(i) = 0, \\
(1 - \gamma_n) \left( \eta_{Q-1} - C(t,i) \leq \omega_i < \eta_Q - C(t,i) \right) + \gamma_p \left( \eta_{Q-2} - C(t,i) \leq \omega_i < \eta_{Q-1} - C(t,i) \right), & \text{if } w(i) = 1,
\end{cases}
\]

and for \( z(t) = r \) and \( r \in \{1, 2, \ldots, Q-2\} \), one has

\[
P_{Z(t)|w}(r|w) = 
\begin{cases} 
(1 - \gamma_n - \gamma_p) \left( \eta_r \leq w_i < \eta_{r+1} \right), & \text{if } w(i) = 0, \\
(1 - \gamma_n - \gamma_p) \left( \eta_r - C(t,i) \leq \omega_i < \eta_{r+1} - C(t,i) \right) + \gamma_p \left( \eta_{r-1} \leq \omega_i < \eta_r \right) + \gamma_n \left( \eta_{r+1} \leq \omega_i < \eta_{r+2} \right), & \text{if } w(i) = 1.
\end{cases}
\]

Using standard BP message independence assumptions, the marginal distribution of the \( i \)th subject after the \( L \)th iteration may be written as:

\[
P_{W(i)|z}^{(L)}(w(i)|z) \propto \left( \frac{d}{n} \mathbb{I}(w(i) = 1) + \left( 1 - \frac{d}{n} \right) \mathbb{I}(w(i) = 0) \right) \prod_{\tau \in N(i)} \tilde{\chi}_{r \to i}^{(L)}(w(i)).
\]

Upon computing the marginals, the set of defectives may be determined based on the following two methods. In the first method,

\[
\hat{\mathcal{D}} = \left\{ i : P_{W(i)|z}^{(L)}(1|z) > P_{W(i)|z}^{(L)}(0|z) \right\}, \tag{5.20}
\]
while in the second method

\[ \hat{D} = \left\{ i : S_i \text{ has one of the } d \text{ largest } P_{W(i,z)}^{(L)}(1|z) \right\}. \]

(5.21)

Note that the complexity of this BP decoder can be further reduced by adapting approaches such as the ones described in the context of \(q\)-ary BP decoding in [66, 67, 68, 69], which will be discussed elsewhere.

For demonstrative purposes, we applied the BP algorithm to an equidistant SQGT model with \(\eta = 2\). We used Construction 8 to generate matrices with \(n = 100\) and \(d = 15\), which represent reasonable parameter choices for the application at hand. In Fig. 5.3 we plotted the probability of error, \(P_e\), as a function of \(q\) for different values of \(\gamma_p\) and \(\gamma_n\), when \(m = 50\). We generated 400 different sets of defectives (trials) for each choice of \(q\) and fixed the number of iterations in the BP algorithm to \(L = 20\). The set of defectives was obtained using (5.21). Figure 5.4 shows the performance of the BP algorithm in a similar setting when (5.20) was used to obtain the set of defectives. To keep the waterfall curves sufficiently uncluttered, we only reported on noisy SQGT performance. Note that the probability of false negatives, \(P_{FN}\), is defined as the probability that a defective is not detected, while the probability of false positives, \(P_{FP}\), is defined as the probability that a non-defective subject is detected as defective. Note that in method (5.21), \(P_e = P_{FN} = P_{FP}\).

As may be seen from the simulation results, there is a clear advantage to using matrices with \(q \geq 3\) from the perspective of BP decoding in the presence of errors. Unfortunately, this effect is accompanied by an increase in the complexity of non-binary BP decoding, which may be mitigated by applications of the aforementioned methods of [66, 67, 68, 69]. One may also notice that the decoding error probability of the BP decoder for the matrices with the considered parameters remains bounded above a value close to 0.1. We believe that this phenomenon is not a result of the unsuitability of BP decoding in SQGT, but rather a consequence of the fact that testing matrices constructed in the paper were not optimized with respect to the requirements of loopy BP. Furthermore, the high probability of error may also be attributed to the fact that the random matrices were generated for parameters that are not in the range of values that guarantee high probability for the SQ disjunctness property.\(^4\) Particularly, in Construction 8, the

\(^4\)Testing the SQ disjunctness property for large matrices is computationally demanding
asymptotic guarantees were results of an upper bound on the probability that \( C \) is not a \([q; Q; \eta; (1:d); 0]-\text{SQ-disjunct} \) matrix. This bound took the form

\[
\Pr(C \text{ is not } [q; Q; \eta; (1:d); 0]-\text{SQ-disjunct}) \leq P_F = \binom{n}{d+1}(d+1)(1-\pi_I)^m,
\]

where \( \pi_I \) was the probability of “success” of a row, as defined in the proof of Construction 8. However, as an example, when \( n = 100, m = 50, \eta = 2, q = 11, \) and \( d = 15 \), this upper bound is larger than 1, i.e. \( P_F > 1 \), and we can therefore not guarantee that the matrix considered for these parameters is \([q; Q; \eta; (1:d); 0]-\text{SQ-disjunct} \) with high probability. A probability of error of approximately 0.15 for \( q \geq 11 \) shows that even though the considered matrices may not satisfy the distinctness property, one is still able to correctly identify the set of defectives with empirical probability approximately 0.85, which is sufficiently high for the described genotyping applications.

and we did not attempt to determine the exact parameters of the SQGT matrix through simulation.

Figure 5.3: Probability of error as a function of the test matrix alphabet size \( q \), for different choices of noise parameters. In the model, we fixed \( \eta = 2, n = 100, d = 15, \) and \( m = 50 \).
Figure 5.4: Probability of false negatives and false positives as a function of the test matrix alphabet size $q$, for different choices of noise parameters. The solid lines represent the probability of false negatives, while the dashed lines represent the probability of false positives. We fixed $\eta = 2$, $n = 100$, $d = 15$, and $m = 50$. 
In order to demonstrate the effect of $m$ on the performance of the algorithm, we applied the BP algorithm on an equidistant SQGT model with $\eta = 2$. Using Construction 8, we generated matrices with $n = 100$, $d = 15$, and $q = 11$. Figure 5.5 shows the probability of error as a function of $m$ for noisy and noise-free scenarios when (5.21) was used to obtain the set of defectives. For each $m$, the BP algorithm was applied on 400 random matrices and terminated with no more than $L = 20$ iterations. Similarly, Fig. 5.6 shows the probabilities of false negatives and false positives when (5.20) was used to find the set of defectives.
Figure 5.6: Probability of false negatives and false positives as a function of $q$ for different noise parameters. The solid lines represent the probability of false negatives while the dashed lines represent the probability of false positives. In this model, we fixed $\eta = 2$, $n = 100$, $d = 15$, and $q = 11$. 
CHAPTER 6

POISSON GROUP TESTING FOR THE CGT MODEL

In the previous chapter, we discussed test matrix constructions and efficient decoding algorithms for a special case of the SQGT model: SQGT with equidistant thresholds. In this chapter, we focus on the CGT model [16], another important special case of SQGT in which $q = Q = 2$ and $\eta_1 = 1$. For this model, we introduce a novel probabilistic group testing framework, termed Poisson group testing, in which the number of defectives follows a right-truncated Poisson distribution. The Poisson model has a number of new applications, including dynamic testing with diminishing relative rates of defectives. We consider both nonadaptive and semi-adaptive identification methods. For nonadaptive methods, we derive a lower bound on the number of tests required to identify the defectives with a probability of error that asymptotically converges to zero; in addition, we propose test matrix constructions for which the number of tests closely matches the lower bound. For semi-adaptive methods, we describe a lower bound on the expected number of tests required to identify the defectives with zero error probability. In addition, we propose a stage-wise reconstruction algorithm for which the expected number of tests is only a constant factor away from the lower bound. The methods rely only on an estimate of the average number of defectives, rather than on the individual probabilities of subjects being defective.

6.1 Introduction

As discussed earlier, the group testing literature may be divided into two categories based on how the number of defectives is modeled. In combinatorial GT, the number of defectives, or an upper bound on the number of defectives, is fixed and assumed to be known in advance [7]. On the other hand, in probabilistic GT (PGT), the number of defectives is a random variable
with a given probability distribution [16]. With almost no exceptions, the PGT literature focuses on a binomial \((n, p_0)\) distribution for the number of defectives. Such a model arises when each of the \(n\) subjects is defective with a fixed probability \(0 < p_0 < 1\), independent of all other subjects. Binomial models are not necessarily sparse, given that \(p_0\) may be a constant and given that the defective selection process is random.

Here, we propose a novel GT paradigm, termed Poisson PGT, which models the distribution of the number of defectives via a right-truncated Poisson distribution with parameter \(\lambda(n) = o(n)\). Our motivation for this assumption comes from clinical testing, where one is interested in identifying infected individuals under the assumption that infections gradually die out. A similar scenario is encountered in screening DNA clones for the presence of certain DNA substrings, where the clones are test subjects and defectives are clones that contain the given substrings. The distribution of clones containing a given DNA pattern is frequently modeled as Poisson [7]. Other applications include testing genetic traits that are negatively selected for (i.e., traits that diminish in time, as they reduce the fitness of a species). The assumption \(\lambda(n) = o(n)\) ensures that the longer the waiting time or the larger the number of test subjects, the smaller the average relative fraction of defectives. In other words, the rate of defectives diminishes with time.

The Poisson PGT model has a number of useful properties that make it an important alternative to classical binomial models. Although a binomial distribution with \(p_0 \ll 1\) and a large \(n\), where \(\lambda = np_0\) is a constant, converges to a Poisson distribution with parameter \(\lambda = np_0\) [42], our model allows the parameter \(\lambda(n)\) of the (truncated) Poisson distribution to grow with \(n\); more precisely, the model and the results derived in this chapter are valid even if \(\lim_{n \to \infty} \lambda(n) = \infty\), as long as \(\lim_{n \to \infty} \frac{\lambda(n)}{n} = 0\). Such a model is useful in settings were test subjects are assumed to arrive sequentially in time, and where tests are performed only once a sufficient number of subjects \(n\) is present. This model is also applicable to streaming and dynamic testing scenarios [43], in which the probability that a subject is defective decreases in time so that newly arriving subjects are less likely to be defective. In such a setting, classical binomial\((n, p_0)\) models are inadequate, as they assume that the probability \(p_0\) of a subject being defective does not depend on the number of test subjects.

A number of papers have considered a Poisson model to capture the stream-
ing dynamics of the \textit{arrivals of subjects} to a test center \cite{44}, \cite{45}. In contrast, our model does not make any assumptions on the distribution of the general subject population, but instead focuses on modeling the number of defectives using a right-truncated Poisson distribution. In addition, the focus of \cite{44}, \cite{45} is on determining the total amount of time (delay) required to test a batch of subjects arriving at random times. However, here we concentrate on the completely unrelated problem of finding necessary and sufficient conditions on the smallest number of tests needed for accurate nonadaptive and semi-adaptive GT.

In addition, a number of papers have considered the problem of binomial group testing with different subjects having different probabilities of being defective. This line of work was introduced in \cite{19} under the name \textit{generalized binomial group testing} (GBGT). Recently, this problem has received renewed interest under the name of \textit{heterogeneous binomial group testing} \cite{70}. In \cite{19}, a two-stage algorithm for GBGT was proposed, resulting in a complicated minimization problem for the expected number of tests required; unfortunately, no closed-form expression, nor any simply calculable expression, was provided for the expected number of tests. In \cite{18}, a similar problem was considered in which the goal was to isolate a \textit{single} defective in the GBGT model. For this problem, the authors proposed an optimal adaptive procedure using a binary testing tree, which was obtained for a set of weights that depend on the probabilities of the subjects being defective. In addition, an upper bound on the expected number of tests was provided in the form of a complicated sum. Other papers that consider the GBGT model include \cite{70}, \cite{71}, and \cite{72}. As we explain in the next section, although related to our Poisson model through Le Cam’s theorem, GBGT operates under very different prior knowledge assumptions and cannot be considered within the same analytical framework.

The main contributions of this chapter are three-fold. First, we introduce a novel probabilistic GT model with applications in streaming and dynamic testing scenarios. This model generalizes probabilistic group testing models beyond the binomial GT paradigm with constant $p_0$ and other models previously considered in the literature. Second, we bridge the gap between combinatorial GT and probabilistic GT methodology by showing how the algorithms and analytical tools developed for combinatorial GT can be generalized and adapted for probabilistic GT. To the best of our knowledge, this
is the first attempt to analyze combinatorial GT and probabilistic GT within the same framework. Finally, we derive closely matching lower and upper bounds on the number of tests required for finding the defectives in Poisson testing using both nonadaptive and semi-adaptive algorithms.

The chapter is organized as follows. Section 6.2 introduces the Poisson GT model. In Section 6.3, we first use an adaptation of Fano’s inequality to find a lower bound on the number of nonadaptive tests required to identify defectives under the Poisson PGT model, with a probability of error converging to zero as the number of subjects grows. We then proceed to describe a simple nonadaptive method based on binary disjunct matrices [24]. The test matrix is constructed probabilistically, with the entries of the matrix being independent and Bernoulli distributed. Given that the number of tests obtained via this method does not tightly match the lower bound, we describe an alternative nonadaptive method with a number of tests differing from the lower bound by only an arbitrary slowly-growing function in $n$. The test matrix in this method does not rely on the disjunctness property and the entries of the matrix are not i.i.d. distributed. Next, we demonstrate that this non-i.i.d. construction outperforms standard information-theoretic sufficient conditions for i.i.d. constructions. Following the practice of binomial probabilistic group testing, in Section 6.4 we use Huffman coding to find a lower bound on the expected number of tests required by adaptive and semi-adaptive methods to identify the defectives with zero error probability. Then, we show that a simple semi-adaptive algorithm identifies all the defectives with an expected number of tests only a constant factor away from the lower bound.

### 6.2 Problem setup

Throughout this chapter we adopt the following notation. Simple uppercase letters are used to denote random matrices, random vectors, and random variables; similarly, simple lowercase letters are used for scalars. The symbols $\log(\cdot)$ and $\log_2(\cdot)$ are used to denote the natural logarithm and the base 2 logarithm, respectively. For a finite integer $K \geq 1$, we also make use of the
$K$-fold logarithm function, defined as
\[ \log^{(K)} n = \log \log \cdots \log n. \]  
(6.1)

Note that for $K > 1$, this function grows slower than $\log(n)$.

Let $S$ denote the set of test subjects with cardinality $n$, among which a subset of $D$ subjects is defective. In the Poisson PGT model, we assume that the number of defectives follows a right-truncated Poisson distribution with parameters $\lambda(n)$ and $n$, i.e.,
\[ P_D(d) = \begin{cases} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)}, & 0 \leq d \leq n \\ 0, & \text{otherwise.} \end{cases} \]  
(6.2)

Here, $D = |D|$ denotes the number of defectives and $c(n) = e^{\lambda(n)} / \sum_{d=0}^{n} \frac{\lambda(n)^d}{d!}$ is a normalization coefficient. Note that $c(n)$ is a decreasing function of $n$, such that $\lim_{n \to \infty} c(n) = 1$. In addition we assume that all subsets of $S$ with equal cardinality have the same probability of being defective. This assumption is used to model the setup in which given $D = d$, the decoder has no information as to which set of cardinality $d$ is most likely to be the set of defectives.

Let $\bar{\lambda}(n)$ be the expected number of defectives in the model. It can be easily verified that
\[ \bar{\lambda}(n) = \mathbb{E}[D] = \lambda(n) \left(1 - \frac{\lambda(n)^n}{n!} e^{-\lambda(n)}\right) = \lambda(n)(1 - o(1)). \]  
(6.3)

A right-truncated Poisson distribution is closely related to a finite support version of the non-uniform Bernoulli model on the set of test subjects, in which the $i^{th}$ subject is defective with probability $p_i$, $0 \leq p_i \leq 1$, independent of all other test subjects. From Le Cam’s theorem [73], it may be deduced that the number of defectives $D$ under this model satisfies
\[ \sum_{k=0}^{\infty} \left| \Pr(D = d) - e^{-\lambda(n)} \frac{\lambda(n)^d}{d!} \right| \leq 2 \sum_{i=1}^{n} p_i^2, \]  
(6.4)

where $\lambda(n) = \sum_i^n p_i$. As an example, one can choose $p_i = c/i$, for some constant $c > 0$, to arrive at a model where individual subjects have decreasing probabilities in $i$ of being defective, so that $\lambda(n) = O(\log n)$. The approxima-
tion error with respect to the Poisson distribution scales as \(2c \zeta(2) = c\pi^2/3\), where \(\zeta(\cdot)\) denotes the Riemann zeta function. By choosing \(c\) sufficiently small, the approximation error can be reduced to any desired positive level. Although adaptive and other classes of non-uniform Bernoulli models were reported in the literature [19], [18], [74], they rely on the exact knowledge of each probability \(p_i\), \(i = 1, \ldots, n\). However, even in applications in which a subject is defective independently from all other subjects, estimating each of the \(p_i\) values may be prohibitively difficult. In contrast, Poisson PGT only makes use of a single aggregate value of the probabilities, \(\lambda(n)\), which is less informative but usually much easier to estimate.

In the GT framework, each test is performed on a subset of the subjects, and the result of a test equals 1 if at least one defective is present in the test, and 0 otherwise. The total number of tests is denoted by \(m\). For non-adaptive PGT, despite the fact that the defectives are chosen randomly, the number of tests is deterministic. The question of interest is to find the smallest number of tests that guarantee a probability of detection error that converges to zero asymptotically with the number of subjects \(n\). In contrast, adaptive and semi-adaptive algorithms, in which tests are performed sequentially or grouped into different stages with the choice at one stage used to inform the choice at the following stages, call for a random number of tests. The goal is then to compute the expected number of tests that allows for zero error probability. In this chapter, we focus on nonadaptive and semi-adaptive testing schemes, and our main results are summarized and discussed in Section 6.5.

6.3 Nonadaptive methods for Poisson PGT

Nonadaptive group testing refers to group testing methods in which all tests are designed simultaneously. In other words, in nonadaptive GT the choice of a test is not allowed to depend on the outcomes of previous tests [7]. The main advantage of nonadaptive methods is that all the tests can be performed in parallel, which is of great practical importance for large-scale problems. A clear disadvantage compared to adaptive methods is the sometime significant increase in the number of tests.

As mentioned earlier, the CGT model is a special case of SQGT in which \(q = Q = 2\) and \(\eta_1 = 1\). For completeness, we describe this special case of
SQGT in more detail in what follows. In nonadaptive CGT, the assignment of subjects to different tests is usually specified via a binary test matrix \( C \in \{0, 1\}^{m \times n} \), where \( m \) denotes the number of tests and \( n \) denotes the number of subjects. If \( C(i, j) = 1 \), for \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \), the \( j \)th subject is present in the \( i \)th test; on the other hand, if \( C(i, j) = 0 \), then the \( j \)th subject is excluded from the \( i \)th test. The test results are captured by the binary vector of test results, \( y \in \{0, 1\}^m \).

It can be easily observed that the vector of test results is equal to the Boolean OR of columns of \( C \) corresponding to the defectives. Figure 6.1 illustrates the notion of a test matrix, the set of defectives, and the vector of test results in the CGT model. Note that \( S_i \) denotes the \( i \)th subject in \( S \).

For a fixed test matrix on \( n \) subjects, \( C \), and a decoding algorithm \( f : (C, y) \mapsto \hat{D} \), let \( \mathcal{E}(n) \) denote the event that the decoding algorithm cannot identify the set of defectives, i.e. the event that \( f(C, y) \neq D \). The ultimate goal of most combinatorial nonadaptive GT methods is to ensure that \( P(\mathcal{E}(n)) = 0 \). Due to the probabilistic nature of the Poisson PGT model, any subset of subjects may be defective with a non-zero probability. As a result, since in nonadaptive GT each test is designed independently from previous tests, for any fixed test matrix \( C \) with \( m < n \), one can always find a choice of \( D \) for which \( f(C, y) \neq D \).

To verify the correctness of this claim, consider a fixed test matrix \( C \) and a set of subjects \( S \) such that each column of \( C \) is assigned to one subject in \( S \); for any set \( D' \subseteq S \), let \( y_{D'} \) denote the Boolean OR of the columns of \( C \) corresponding to \( D' \). Since in Poisson PGT each subset of \( S \) may correspond to the set of defectives with a nonzero probability, in order to ensure \( P(\mathcal{E}(n)) = 0 \), the test matrix must be able to distinguish between any two distinct subsets of \( S \); in other words, for any two distinct sets \( D_1, D_2 \subseteq S \),
we must have \( y_1 \neq y_2 \). Since in total, there exist \( \sum_{d=0}^{n} \binom{n}{d} = 2^n \) choices for the set of defectives, at least \( m = n \) tests are required (i.e. one has to test each subject individually).

The discussion above implies that for Poisson PGT, there exists no non-adaptive test matrix with fewer than \( n \) rows and an accompanying decoding algorithm for which \( P(\mathcal{E}(n)) = 0 \); as a result, we instead focus on the requirement that the test matrix satisfy the asymptotic condition

\[
\lim_{n \to \infty} P(\mathcal{E}(n)) = 0.
\] (6.5)

In what follows, we propose two nonadaptive test matrix constructions and decoding algorithms to guarantee that the aforementioned condition is met.\(^1\) In order to evaluate how effectively each method uses its tests, we first find a lower bound on the minimum number of tests required by a nonadaptive algorithm to ensure (6.5), and then use this bound as a benchmark. The constructive methods provide upper bounds on the minimum number of tests.

In Section 6.3.1, we use Fano’s inequality [75] to find a lower bound on the number of tests of the form \((1 - \epsilon)\lambda(n) \log_2 n \left(1 - o(1)\right)\), where \( \epsilon \) is an arbitrarily small fixed scalar such that \( 0 < \epsilon < 1 \). In Section 6.3.2, we propose a test matrix construction using binary disjunct matrices. The entries of the test matrix are chosen according to an i.i.d. distribution, and the method requires \( m = C_2 \alpha(n) \lambda(n)^2 \log_2 n \left(1 + o(1)\right) \) measurements, where \( \alpha(n) \) is an arbitrary chosen slowly-growing function of \( n \).

Given the gap between the number of tests in this method and the lower bound, we propose another method that requires \( m = C_1 \alpha(n) \lambda(n) \log_2 n \left(1 + o(1)\right) \) tests. The second method is also based on a probabilistic construction, but the entries of the test matrix no longer follow an i.i.d. distribution. One major difference between these two methods is that the first method uses the disjunctness property [24, 22], while for the second method we relaxed this constraint. Both of these constructions can be extended to identify the set of defectives in the presence of errors in the vector of test results, and also employ a decoding algorithm with computational complexity \( O(mn) \).

Finally, we use a standard information theoretic approach – combined with a maximum likelihood decoder – to determine an upper bound on the

\(^{1}\) Also, we omit the parameter \( n \) whenever possible and tacitly assume the dependence of the error probability on this parameter.
minimum number of tests required by any nonadaptive method based on an i.i.d. test matrix. The number of tests for this approach is of the form 
\[ m = C_3 n \lambda(n)^2 \log_2 n (1 + o(1)) \], and hence still a factor of \( \lambda(n) \) larger than the second method. A summary of these results is provided in Section 6.5.

6.3.1 Lower bound on the minimum number of tests

Let \( \hat{D} \) be the set of defectives recovered by some decoding algorithm using a fixed test matrix \( C \in \{0, 1\}^{m \times n} \), and given the random vector of test results, \( Y \in \{0, 1\}^m \). Conditioned on \( D = d, 0 \leq d \leq n \), let \( \mathcal{E}_d \) be the error event that \( \hat{D} \neq D \); consequently, \( P(\mathcal{E}) = \mathbb{E}_D[P(\mathcal{E}_d)] \). Using Fano’s inequality [75], one has

\[
H(D \mid Y, C, D = d) \leq 1 + P(\mathcal{E}_d) \log_2 \left( \frac{n}{d} \right), \quad (6.6)
\]

where \( H(\cdot) \) denotes the Shannon entropy function [75]. Since conditioned on \( D = d \), the set of defectives \( D \) is chosen uniformly at random, independent on \( C \), one has

\[
H(D \mid C, D = d) = \log_2 \left( \frac{n}{d} \right). \quad (6.7)
\]

Using the definition of mutual information [75], we may write

\[
H(D \mid C, D) = H(D \mid Y, C, D) + I(D \mid Y \mid C, D)
\]

\[
= H(D \mid Y, C, D) + H(Y \mid C, D) - H(Y \mid C, D, D)
\]

\[
\leq H(D \mid Y, C, D) + H(Y \mid D) - H(Y \mid C_D, D)
\]

\[
= H(D \mid Y, C, D) + I(Y \mid C_D \mid D), \quad (6.8)
\]

where the inequality follows since conditioning reduces entropy; also, the test results \( Y \) only depend on the columns of the test matrix assigned to the set \( D \) and hence \( H(Y \mid C, D, D) = H(Y \mid C_D, D) \), where \( C_D \) is the set of columns of \( C \) corresponding to \( D \). Substituting (6.6) and (6.7) in (6.8) yields

\[
\log_2 \left( \frac{n}{d} \right) \leq 1 + P(\mathcal{E}_d) \log_2 \left( \frac{n}{d} \right) + I(Y \mid C_D \mid D)
\]

\[
\Rightarrow P(\mathcal{E}_d) \geq 1 - \frac{I(Y \mid C_D \mid D) + 1}{\log_2 \left( \frac{n}{d} \right)}. \quad (6.9)
\]
On the other hand, from the following chain of inequalities

\[ I(Y; C_D | D) \leq H(Y | D) \leq H(Y) \leq \sum_{i=1}^{m} H(Y(i)) \leq m, \]

it follows that

\[ P(E_d) \geq 1 - \frac{I(Y; C_D | D) + 1}{\log_2 \binom{n}{d}} \geq 1 - \frac{m + 1}{\log_2 \binom{n}{d}}. \]  

(6.9)

Since \( P(\mathcal{E}) = \mathbb{E}_D[P(\mathcal{E}_d)] \), (6.9) may be used to find a lower bound on \( m \) that ensures \( P(\mathcal{E}) = o(1) \), as formally stated in the next theorem.

**Theorem 17.** Let \( 0 < \epsilon < 1 \) be an arbitrarily small fixed scalar, and suppose that \( \lambda(n) = o(n) \). Any nonadaptive group testing method designed for Poisson PGT that satisfies \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \) requires at least \( m = (1-\epsilon)\lambda(n) \log_2 n \) \((1-o(1))\) tests.

**Proof.** Let \( 0 < \epsilon < 1 \). Then, since \( \lambda(n) = o(n) \), for large enough values of \( n \), \( \lambda(n)(1 + \epsilon) < n \). On the other hand, since \( P(\mathcal{E}_d) \geq 0, 0 \leq d \leq n \), then

\[ P(\mathcal{E}) = \mathbb{E}_D[P(\mathcal{E}_d)] = \sum_{d=0}^{n} P(D = d)P(\mathcal{E}_d) \]

\[ \geq \sum_{d=\lambda(n)(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d)P(\mathcal{E}_d). \]

As a result, a necessary condition for \( P(\mathcal{E}) = o(1) \) is that for \( n \) large enough,

\[ o(1) \geq \sum_{d=\lambda(n)(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d)P(\mathcal{E}_d) \]

\[ \geq \sum_{d=\lambda(n)(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d) \left( 1 - \frac{m + 1}{\log_2 \binom{n}{d}} \right) \]

\[ \geq \left[ \min_{d: \lambda(n)(1-\epsilon) \leq d \leq \lambda(n)(1+\epsilon)} \left( 1 - \frac{m + 1}{\log_2 \binom{n}{d}} \right) \right] \sum_{d=\lambda(n)(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d), \]  

(6.10)

where the second inequality is a consequence of Equation (6.9).

Using the Chernoff bound for a standard Poisson distribution, it may be shown that

\[ \sum_{d=\lambda(n)(1+\epsilon)}^{\infty} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \leq \exp (-\lambda(n)(1+\epsilon) \log (1+\epsilon) + \epsilon \lambda(n)), \]
\[
\sum_{d=0}^{\lambda(n)(1+\epsilon)} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \leq \exp\left(-\lambda(n)(1 - \epsilon) \log(1-\epsilon) - \epsilon \lambda(n)\right).
\]

Let \( f_1(\epsilon) = (1 + \epsilon) \log(1 + \epsilon) - \epsilon \) and \( f_2(\epsilon) = (1 - \epsilon) \log(1 - \epsilon) + \epsilon \). It can be easily verified that for \( \epsilon > 0 \), \( f_1(\epsilon) > 0 \) and \( f_2(\epsilon) > 0 \). As a result, from the above inequalities we obtain

\[
\sum_{d=\lambda(n)(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d) \geq c(n) \left(1 - e^{-\lambda(n)f_1(\epsilon)} - e^{-\lambda(n)f_2(\epsilon)}\right). \tag{6.11}
\]

On the other hand, for \( n > 2(1 - \epsilon)\lambda(n) \),

\[
\min_{d: \lambda(1-\epsilon) \leq d \leq \lambda(1+\epsilon)} \left(1 - \frac{m + 1}{\log_2\left(\frac{n}{d}\right)}\right) = \left(1 - \frac{m + 1}{\log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)}\right). \tag{6.12}
\]

Substituting (6.11) and (6.12) in (6.10), a necessary condition for \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \) is of the form

\[
o(1) > \left[ \min_{d: \lambda(1-\epsilon) \leq d \leq \lambda(1+\epsilon)} \left(1 - \frac{m + 1}{\log_2\left(\frac{n}{d}\right)}\right) \right]^{\lambda(n)} \sum_{d=\lambda(1-\epsilon)}^{\lambda(n)(1+\epsilon)} P(D = d)
\]

\[
> c(n) \left(1 - \frac{m + 1}{\log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)}\right) \left(1 - e^{-\lambda f_1(\epsilon)} - e^{-\lambda f_2(\epsilon)}\right)
\]

\[
> c(n) \left(1 - \frac{m + 1}{\log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)} - e^{-\lambda f_1(\epsilon)} - e^{-\lambda f_2(\epsilon)}\right).
\]

As a result, one has

\[
\left(1 - \frac{m + 1}{\log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)}\right) < o(1) \Rightarrow m \geq \log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)(1 - o(1)).
\]

This inequality can be further simplified as

\[
m \geq \log_2\left(\frac{n}{(1-\epsilon)\lambda}\right)(1 - o(1))
\]

\[
\geq (1 - \epsilon)\lambda(n) \log_2 \frac{n}{(1-\epsilon)\lambda} (1 - o(1))
\]

\[
= (1 - \epsilon)\lambda(n) \log_2 n (1 - o(1)),
\]

where the last equality follows since \( \lambda(n) = o(n) \). \( \square \)

95
6.3.2 Constructive upper bounds on the minimum number of tests

Next, we describe two nonadaptive methods for Poisson PGT and find the number of tests that ensures \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \). For this purpose, we consider two separate asymptotic regimes for \( \lambda(n) \): one in which \( \lambda(n) = o(n) \) and \( \lim_{n \to \infty} \lambda(n) = \infty \), and another in which \( \lambda(n) = o(n) \) and \( 0 < \lim_{n \to \infty} \lambda(n) < \infty \). Note that the case of constant \( \lambda \) is covered by the latter scenario. We start by proving the following simple large deviations results, which we find useful in our subsequent derivations.

**Lemma 9.** Let \( D \) be a random variable following the right-truncated Poisson distribution, with \( \lambda(n) = o(n) \) and \( \lim_{n \to \infty} \lambda(n) = \infty \). Then, for any fixed \( \epsilon > 0 \), one has \( \lim_{n \to \infty} P(D > \Delta) = 0 \), where \( \Delta = [\lambda(n)(1+\epsilon)] - 1 \).

**Proof.** Using Markov’s inequality, one has

\[
P(D > \Delta) \leq \frac{\mathbf{E}[D]}{[\lambda(n)(1+\epsilon)]} = \frac{\lambda(n)}{[\lambda(n)(1+\epsilon)]}(1-o(1))
\]

\[
= \frac{1}{\lambda^n}(1 + o(1)) = o(1),
\]

where the last claim follows since \( \lim_{n \to \infty} \lambda(n) = \infty \). \( \square \)

Although this lemma is applicable when \( \lim_{n \to \infty} \lambda(n) = \infty \), for the case when \( 0 < \lim_{n \to \infty} \lambda(n) < \infty \) (including the case when \( \lambda \) is a constant), the above arguments do not hold. For this case, we prove a lemma in which a slowly-growing function of \( n \), i.e. \( \beta(n) = \log(K) n \) defined in (6.1), is used to provide the needed guarantees.

**Lemma 10.** Let \( D \) be a random variable following the right-truncated Poisson distribution, with \( \lambda(n) = o(n) \) and \( 0 < \lim_{n \to \infty} \lambda(n) < \infty \). Also, let \( \beta(n) = \log(K) n \), for some finite \( K > 1 \). Then, \( \lim_{n \to \infty} P(D > \Delta) = 0 \) for \( \Delta = [\beta(n)\lambda(n)] - 1 \).

**Proof.** Before proving the lemma, Using Markov’s inequality, one has

\[
P(D > \Delta) \leq \frac{\mathbf{E}[D]}{[\beta(n)\lambda(n)]} = \frac{\lambda(n)}{[\beta(n)\lambda(n)]}(1-o(1))
\]

\[
= \frac{1}{\beta(n)}(1 + o(1)) = o(1),
\]

96
where the last equality follows since $\lim_{n \to \infty} \beta(n) = \infty$.

In our first construction, we use disjunct matrices to devise practical Poisson PGT schemes. Although we have already defined these matrices earlier, we repeat this definition as a reminder.

**Definition 18 (Binary $\Delta$-disjunct matrices [24, 22]).** A binary $\Delta$-disjunct matrix for CGT is an $m \times n$ matrix, such that for any set of $\Delta + 1$ columns, $\mathcal{X} = \{x_j\}_{1}^{\Delta+1}$, and for any column $x_i \in \mathcal{X}$, there exists at least one coordinate $k$ such that $x_i(k) = 1$ and $x_j(k) = 0$, for some $x_j \in \mathcal{X}$, where $j \neq i$.

It is well known that binary $\Delta$-disjunct matrices are capable of identifying up to $\Delta$ defectives in the CGT model. In addition, these matrices are endowed with an efficient decoder with computational complexity $O(mn)$. The decoding procedure is based on the fact that a column corresponds to a defective if and only if its support is a subset of the support of the vector of test results, $y$. Hence, given $y$ and $C$, the set of defectives may be identified with zero probability of error through

$$\hat{D} = \{i : \text{supp}(x_i) \subseteq \text{supp}(y)\},$$

(6.13)

where $x_i$ is the $i^{th}$ column of $C$ and supp(·) stands for the support of a vector (i.e. the set of its nonzero entries).

We consider a simple probabilistic construction for the test matrix: the entries of the test matrix follow an i.i.d. Bernoulli($p$) distribution, such that each entry of $C$ is equal to 1 with probability $p$, and 0 with probability $1 - p$. Let $\Delta = \Delta(n, \lambda(n))$ be a properly chosen function of $n$ and $\lambda(n)$. The idea is to identify $m$, $p$ and $\Delta$ so that $C$ is a $\Delta$-disjunct matrix with high probability, while at the same time, the probability that the number of defectives exceeds $\Delta$ is small, as formally stated in the following theorem.

**Theorem 18.** Assume that $D$ follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $\lim_{n \to \infty} \lambda(n) = \infty$. Construct a test matrix by choosing each entry according to a Bernoulli($p$) distribution, where $p = \left(\left\lfloor \lambda(n)^{(1+\epsilon)} \right\rfloor \right)^{-1}$, and where $\epsilon > 0$ is arbitrarily small. Then $m = e(\left\lfloor \lambda(n)^{(1+\epsilon)} \right\rfloor)^2 \log n = e \lambda(n)^{(1+\epsilon)} \log n (1+o(1))$ tests suffice to ensure $\lim_{n \to \infty} P(E) = 0$ using a decoding algorithm with computational complexity $O(mn)$.
Proof. For any value of $\Delta > 0$, we may write $P(\mathcal{E})$ as

\[
P(\mathcal{E}) = P(\mathcal{E}|D \leq \Delta)P(D \leq \Delta) + P(\mathcal{E}|D > \Delta)P(D > \Delta)
\leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta).
\]

From Lemma 9, we know that $\Delta = \lceil \lambda(n)^{(1+\epsilon)} \rceil - 1$ ensures $\lim_{n \to \infty} P(D > \Delta) = 0$. In order to bound $P(\mathcal{E}|D \leq \Delta)$, we use the following argument. The test matrix is constructed in a probabilistic, i.i.d. manner using the Bernoulli($p$) distribution. Given a fixed test matrix $C$ and a vector of test results $y$, we use the decoder in (6.13) to find $\hat{D}$. Let $\mathcal{E}'$ be the event that $C$ is not $\Delta$-disjunct. Since a $\Delta$-disjunct test matrix can identify up to $\Delta$ defectives with zero error probability, then conditioned on $D \leq \Delta$, one has $\mathcal{E} \in \mathcal{E}'$. As a result, $P(\mathcal{E}|D \leq \Delta) \leq P(\mathcal{E}'|D \leq \Delta) = P(\mathcal{E}')$, where the last equality follows since the events $\mathcal{E}'$ and $\{D \leq \Delta\}$ are independent.

It has been shown in [7, Thm. 8.1.3] that by choosing $p = \frac{1}{\Delta+1}$ and $\pi_N = p(1-p)^{\Delta}$, one can bound $P(\mathcal{E}')$ as

\[
P(\mathcal{E}') \leq (\Delta + 1)\left(\frac{n}{\Delta + 1}\right)^m(1-\pi_N)^m
\leq \exp(-m\pi_N + (\Delta+1) + (\Delta+1) \log n - \Delta \log(\Delta+1)).
\]

Hence, $\frac{(\Delta+1) \log n}{\pi_N}$ tests suffice to ensure $\lim_{n \to \infty} P(\mathcal{E}') = 0$. Substituting $\pi_N = \frac{\Delta^\Delta}{(\Delta+1)^{\Delta+1}}$ yields

\[
\frac{(\Delta + 1)}{\pi_N} \log n = (\Delta + 1)^2\left(1 + \frac{1}{\Delta}\right) \log n
\leq e(\Delta + 1)^2 \log n = e(\lceil \lambda(n)^{(1+\epsilon)} \rceil)^2 \log n.
\]

In addition, since $P(\mathcal{E}) \leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta) \leq P(\mathcal{E}') + P(D > \Delta)$, $m = e(\lceil \lambda(n)^{(1+\epsilon)} \rceil)^2 \log n = e \lambda(n)^{2(1+\epsilon)} \log n (1+o(1))$ tests suffice to ensure $\lim_{n \to \infty} P(\mathcal{E}) = 0$. \hfill $\square$

The previous theorem relies on the assumption that $\lim_{n \to \infty} \lambda(n) = \infty$. A similar approach can be used for the case $0 < \lim_{n \to \infty} \lambda(n) < \infty$, as described in the theorem to follow.

**Theorem 19.** Assume that $D$ follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $0 < \lim_{n \to \infty} \lambda(n) < \infty$. Let $\beta(n) = \log^{(K)} n$, where
for some finite $K > 1$. Construct a test matrix by choosing each entry according to a Bernoulli$(p)$ distribution, where $p = ([\beta(n)\lambda(n)])^{-1}$. Then $m = e\left(\lceil \beta(n)\lambda(n) \rceil \right)^2 \log n = e (\beta(n)\lambda(n))^2 \log n (1 + o(1))$ tests suffice to ensure $\lim_{n \to \infty} P(E) = 0$ using a decoding algorithm with computational complexity of $O(mn)$.

**Proof.** Similarly as in the proof of Theorem 18, we may write $P(E) \leq P(E|D \leq \Delta) + P(D > \Delta)$, for any $\Delta > 0$. From Lemma 10, we know that setting $\Delta = \lceil \beta(n)\lambda(n) \rceil - 1$ ensures $\lim_{n \to \infty} P(D > \Delta) = 0$. By choosing $p = \frac{1}{\Delta+1}$ and invoking the same arguments as those in Theorem 18, we conclude that $m = e (\Delta + 1)^2 \log n$ tests suffice for $\lim_{n \to \infty} P(E|D \leq \Delta) = 0$. Substituting the previously computed value of $\Delta$ into the expression for the number of tests results in $m = e \left(\lceil \beta(n)\lambda(n) \rceil \right)^2 \log n$. \hfill $\square$

Theorems 18 and 19 do not account for the presence of errors in the vector of test results. In order to address this issue, we invoke the following definition of an error-tolerant binary disjunct matrix.

**Definition 19 (Error tolerant binary $\Delta$-disjunct matrices [22]).** A binary $\Delta$-disjunct matrix designed for CGT, capable of correcting up to $v$ errors, is an $m \times n$ matrix such that for any set of $\Delta+1$ columns, $\mathcal{X} = \{x_j\}_{j=1}^{\Delta+1}$, and for any column $x_i \in \mathcal{X}$, there exists a set of coordinates $R_i$ of size at least $2v+1$, such that $\forall k \in R_i$, $x_i(k) = 1$ and $x_j(k) = 0$, for some $x_j \in \mathcal{X}$ with $j \neq i$.

In order to identify the set of defectives using these matrices with a zero error probability, we use the following decoder. For each column $x_i$, $i \in \{1, 2, \ldots, n\}$, let $N_i$ denote the number of coordinates $j \in \{1, 2, \ldots, m\}$ for which $x_i(j) = 1$ and $y(j) = 0$ hold simultaneously. Then

$$\hat{D} = \{i : N_i \leq v\}.$$  \hfill (6.14)

Note that the computational complexity of this decoding method is $O(mn)$.

The next theorems use error-tolerant disjunct matrices in order to bound the number of tests for a Poisson PGT model that guarantees (6.5) in the presence of up to $v$ errors in the vector of test results $y$.

**Theorem 20.** Assume that the number of defectives follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $\lim_{n \to \infty} \lambda(n) = \infty$. Construct
a test matrix by choosing each entry according to a Bernoulli\((p)\) distribution, where \(p = ([\lambda(n)^{(1+\epsilon)}])^{-1}\) and \(\epsilon > 0\) is arbitrarily small. Then \(m = (2e\lambda(n)2(1+\epsilon)\log n + 4e\log n)^{(1+\epsilon)}\) tests suffice to ensure \(\lim_{n \to \infty} P(\mathcal{E}) = 0\) in the presence of not more than \(v(n)\) errors, using a decoding algorithm with computational complexity \(O(mn)\).

Proof. Similar to the proof of Theorem 18, we may write \(P(\mathcal{E}) \leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta)\), for any value of \(\Delta > 0\). Lemma 9 can be used directly to show that \(\lim_{n \to \infty} P(D > \Delta) = 0\), if \(\Delta = [\lambda(n)^{(1+\epsilon)}] - 1\). In order to bound \(P(\mathcal{E}) \leq P(\mathcal{E}|D \leq \Delta)\), the approach of [7, Thm. 8.1.3] used in Theorem 18 can be generalized to show that \(P(\mathcal{E}|D \leq \Delta) \leq P(\mathcal{E}'|D \leq \Delta) = P(\mathcal{E}')\), where \(\mathcal{E}'\) is the event that \(\mathcal{C}\) is not an \(v\) error correcting \(\Delta\)-disjunct test matrix. To bound \(P(\mathcal{E}')\), we first fix a set of column-indices \(\mathcal{I} : |\mathcal{I}| = \Delta + 1\) and let \(k \in \mathcal{I}\) be fixed. There are \((\Delta + 1)\binom{n}{\Delta + 1}\) ways to choose \(k\) and \(\mathcal{I}\). For a fixed choice of \(\mathcal{I}\) and \(k\), \(\forall j \in \{1, 2, \ldots, m\}\), let \(N_j\) be a Bernoulli random variable such that it has a value 1 if the \(j^{th}\) row of \(\mathcal{C}\) has a value 1 in the \(k^{th}\) column while having 0 in each column indexed by \(\mathcal{I}\setminus\{k\}\), and \(N_j\) has a value 0 otherwise. By definition, the random variables \(N_j\) are i.i.d., and for \(j \in m\) one has

\[
\Pr(N_j = 1) = p(1 - p)^\Delta \approx \pi_N.
\]

Using the Chernoff bound for binomial random variables for \(0 < \delta < 1\), one obtains

\[
\Pr\left(\sum_{j=1}^{m} N_j \leq (1 - \delta)m\pi_N\right) \leq \exp\left(-\frac{\delta^2 m\pi_N}{2}\right).
\]

By setting \(\delta = 1 - \frac{2v}{m\pi_N}\), it follows that

\[
\Pr\left(\sum_{j=1}^{m} N_j \leq 2v\right) \leq \exp\left(-\frac{m\pi_N}{2}\left(1 - \frac{2v}{m\pi_N}\right)^2\right),
\]

which provides an upper bound on the probability that for a fixed \(\mathcal{I}\) and \(k\), at most \(2v\) rows of \(\mathcal{C}\) satisfy the disjunctness property. As a result,

\[
P(\mathcal{E}') \leq \left(\begin{array}{c} n \\ \Delta + 1 \end{array}\right)(\Delta + 1) \exp\left(-\frac{m\pi_N}{2}\left(1 - \frac{2v}{m\pi_N}\right)^2\right)
\]

\[
\leq \exp((\Delta + 1)\log n + \Delta + 1 - \Delta \log(\Delta + 1) - \frac{m\pi_N}{2} - \frac{2v^2}{m\pi_N} + 2v).
\]

Hence, \(2\frac{\Delta+1}{\pi_N}\log n + 4\frac{v}{\pi_N}\) tests suffice to ensure \(\lim_{n \to \infty} P(\mathcal{E}') = 0\). Substituting
\[ \pi_N = \frac{\Delta^\lambda}{(\Delta+1)^{\lambda N}}, \text{ yields} \]
\[ 2 \left( \frac{\Delta+1}{\pi_N} \right) \log n + \frac{4v}{\pi_N} = 2(\Delta+1) \left( 1 + \frac{1}{\Delta} \right) \left( (\Delta+1) \log n + 2v \right) \]
\[ \leq 2 e(\Delta+1) \left( (\Delta+1) \log n + 2v \right) \]
\[ = 2 e \left( [\lambda(n)]^{(1+\epsilon)} \right) \left( [\lambda(n)]^{(1+\epsilon)} \right) \log n + 2v. \]

Consequently, \( m = (2e \lambda(n)^2(1+\epsilon) \log n + 4e v(n)\lambda(n)^{1+\epsilon}) (1+o(1)) \), tests suffice to ensure \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \).

**Theorem 21.** Assume that the number of defectives follows the right-truncated Poisson distribution, with \( \lambda(n) = o(n) \) and \( 0 < \lim_{n \to \infty} \lambda(n) < \infty \). Let \( \beta(n) = \log(K)n \), for some finite \( K > 1 \). Construct a test matrix by choosing each entry according to a Bernoulli \( p \) distribution, where \( p = ([\beta(n)\lambda(n)])^{-1} \). Then \( m = (2e (\beta(n)\lambda(n))^2 \log n + 4e v(n)\beta(n)\lambda(n)) (1+o(1)) \) tests suffice to ensure \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \) in the presence of not more than \( v(n) \) errors, using a decoding algorithm with computational complexity \( O(mn) \).

**Proof.** Since \( P(\mathcal{E}) \leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta) \), for any \( \Delta > 0 \), Lemma 10 ensures that \( \lim_{n \to \infty} P(D > \Delta) = 0 \) if \( \Delta = [\beta(n)\lambda(n)] - 1 \). Repeating the arguments of Theorem 20, we conclude that \( m = 2e (\Delta+1)^2 \log n + 4e v(\Delta+1) \) tests suffice for \( \lim_{n \to \infty} P(\mathcal{E}|D \leq \Delta) = 0 \). Substituting the previously computed value of \( \Delta \) into the expression for the number of tests results in \( m = (2e(\beta(n)\lambda(n))^2 \log n + 4e v(n)\beta(n)\lambda(n)) (1+o(1)) \).

In [76], Cheng and Du described the construction of a probabilistic test matrix for the nonadaptive combinatorial GT model, and proved that their test matrix can identify up to \( \Delta \) defectives from \( n \) subjects with high probability. Although the underlying matrices are not binary disjunct, the decoder in (6.13) can be used to identify the defectives with high probability. The construction consists of two steps: in the first step, a nonbinary test matrix with i.i.d. entries is created; in the second step, a transformation is used to convert this nonbinary matrix into a binary matrix [76, Thm. 1]. One should note that as a consequence of this transformation, the entries of the binary test matrix are no longer i.i.d. We use this construction technique to identify the set of defectives in Poisson PGT and achieve this with a suitable choice of \( \Delta \). The following lemma is a restatement of the results in [76, Thm. 10], suitable for our application.
Lemma 11. The nonadaptive group testing method in [76] can identify up to $\Delta$ defectives among $n$ subjects, using no more than $\frac{3\Delta}{\log_2 3} \left( \log_2 n + \log_2 \frac{1}{1-p} \right)$ tests, with probability at least $p$.

Proof. See [76, Thm. 10] and its proof.

Next, we show how this pooling design can be used to identify the set of defects in Poisson PGT, while ensuring a probability of error that diminishes asymptotically.

Theorem 22. Assume that $D$ follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $\lim_{n \to \infty} \lambda(n) = \infty$. Then, one can identify the set of defectives such that $\lim_{n \to \infty} P(\mathcal{E}) = 0$, using $m = \frac{3\Delta}{\log_2 3} \lambda(n)(1+\epsilon) \log_2 n \left( 1 + o(1) \right)$ tests.

Proof. We first write $P(\mathcal{E})$ as

$$ P(\mathcal{E}) = P(\mathcal{E}|D \leq \Delta) P(D \leq \Delta) + P(\mathcal{E}|D > \Delta) P(D > \Delta) $$

$$ \leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta). $$

Given $\Delta = [\lambda(n)^{(1+\epsilon)}] - 1$, for a fixed $\epsilon > 0$, we use Lemma 9 to conclude that $\lim_{n \to \infty} P(D > \Delta) = 0$. By setting $p = 1 - \frac{1}{\log n}$ and using Lemma 11, we can show that one can identify up to $\Delta$ defectives with no more than $m = \frac{3\Delta}{\log_2 3} \lambda(n)(1+o(1))$ tests, so that the probability of error is bounded as

$$ P(\mathcal{E}|D \leq \Delta) \leq 1 - p = \frac{1}{\log n}. $$

Consequently, one has

$$ \lim_{n \to \infty} P(\mathcal{E}) \leq \lim_{n \to \infty} P(\mathcal{E}|D \leq \Delta) + \lim_{n \to \infty} P(D > \Delta) = 0. $$

Theorem 23. Assume that $D$ follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $0 < \lim_{n \to \infty} \lambda(n) < \infty$. Let $\beta(n) = \log(K) n$, for some value of $K > 1$. Then, one can identify defectives with $\lim_{n \to \infty} P(\mathcal{E}) = 0$, using $m = \frac{3\Delta}{\log_2 3} \beta(n) \lambda(n) \log_2 n \left( 1 + o(1) \right)$ tests.

Proof. Similar to what was done in Theorem 22, we may write $P(\mathcal{E}) \leq P(\mathcal{E}|D \leq \Delta) + P(D > \Delta)$, for any value of $\Delta > 0$. Given $\Delta = [\beta(n) \lambda(n)] - 1$,
we use Lemma 10 to conclude that \( \lim_{n \to \infty} P(D > \Delta) = 0 \). By setting \( p = \frac{1}{\Delta + 1} \)
and invoking the same arguments provided in the proof of Theorem 22, we conclude that \( m = \frac{3\Delta}{\log_2 3} \log_2 n \left(1 + o(1)\right) \) tests are sufficient to ensure that \( \lim_{n \to \infty} P(\mathcal{E}|D \leq \Delta) = 0 \). Substituting the previously computed value of \( \Delta \)
into the expression for \( m \) results in \( m = \frac{3}{\log_2 3} \beta(n) \lambda(n) \log_2 n \left(1 + o(1)\right) \).

Information theoretic approaches have been used in the study of combinatorial CGT problem by several authors [23], [77, 78, 79, 65]. In what follows, we apply these approaches to the Poisson PGT model in order to derive an upper bound on the minimum number of nonadaptive tests that satisfy (6.5). We assume that the test matrix is constructed probabilistically: the entries of the test matrix follow an i.i.d. Bernoulli \((p)\) distribution, such that each entry of \( C \) is equal to 1 with probability \( p \), and 0 with probability \( 1 - p \). For this construction method, we consider a maximum likelihood (ML) decoding procedure, which given the vector of test results and the test matrix reduces to:

\[
\hat{D} = \arg \max_{D'} P(y, C|D') \arg \max_{D'} P(y|C, D').
\]

(6.15)

Here, \( P(y|C, D') \) denotes the conditional distribution of observing \( y \) given the test matrix \( C \) and set of defectives \( D' \). Note that the second equality holds since the test matrix is constructed independent of the set of defectives.

The goal is to find the number of tests required to satisfy (6.5). We define the error event \( \mathcal{E}' \) as the event that there exists a set of subjects \( D' \neq D \) such that \( P(y|C, D') \geq P(y|C, D) \). It can be easily verified that \( P(\mathcal{E}) \leq P(\mathcal{E}') \). As a result, a number of tests that guarantees \( \lim_{n \to \infty} P(\mathcal{E}') = 0 \) also guarantees (6.5). Given \( D = d, 1 \leq d \leq n \), let \( \mathcal{E}_i', 1 \leq i \leq d \), denote the event that there exists a set of subjects with cardinality \( d \), that differ from \( D \) in exactly \( i \) items and is at least as likely as \( D \) to the decoder. Given these definitions, one has

\[
P(\mathcal{E}') = \mathbb{E}_D \left[ P(\mathcal{E}'|D) \right] = \sum_{d=1}^{n} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} P(\cup_{i=1}^{d} \mathcal{E}_i') \leq \sum_{d=1}^{n} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} P(\mathcal{E}_i'),
\]

(6.16)

where the last inequality follows from the union bound.
At first glance, it may seem that a bound on $P(\mathcal{E}^i)$ may be obtained using an upper bound on $P(\mathcal{E}^i)$ for a fixed value of $d$ (such as the bound presented in [79]) and subsequent averaging; however, there are two subtle yet important issues that prohibit us from using this approach. First, in (6.16) the value of $d$, and hence $i$, may be as large as $n$. Since we are interested in the asymptotic regime where $n \to \infty$, a bound on $P(\mathcal{E}^i)$ should account for the growth of $d$ and $i$ with respect to $n$. Second, all known bounds on $P(\mathcal{E}^i)$ (see [79] and references therein) rely on a test matrix $C$ with i.i.d. Bernoulli$(1/d)$ entries. However, in Poisson PGT, the true value of $d$ is unknown (more precisely, $D$ is a random variable) and cannot be used as a design parameter in a natural way. In order to overcome the aforementioned problems, we derive special functions that bound $P(\mathcal{E}^i)$ for different ranges of $d$ and in addition derive new bounds that do not rely on the value of $d$ as a design parameter.

We start by observing that in [79], it was shown that for $d = o(n)$, and for all $\rho, 0 \leq \rho \leq 1$, one has

$$P(\mathcal{E}^i) \leq 2^{-m \left( E_o(\rho, i, d, n) - \frac{\rho \log (\frac{n^{i-d}}{m})}{m} \right)}$$

(6.17a)

where the error exponent $E_o$ satisfies

$$E_o(\rho, i, d, n) = -\log \sum_{Y \in \{0,1\}} \sum_{T_2} \left( \sum_{T_1} P(t_1) P(y, t_2 | t_1, D) \right)^{\frac{1}{1+\rho}}.$$

(6.17b)

Here, we diverge slightly from the previously used notation and let $Y$ denote a random variable corresponding to the result of a single test and let $y$ be a realization of $Y$. Let $(D_1, D_2)$ be a partition of $D$ into disjoint sets with cardinalities $|D_1| = i$ and $|D_2| = d - i$, respectively. The vectors $T_1$ and $T_2$ are binary-valued row-vectors of length $i$ and $d - i$, indicating which subjects in $D_1$ and $D_2$ are present in a given test, respectively. Also, $t_1$ and $t_2$ are realizations of $T_1$ and $T_2$, respectively.

In order to prove the main results of this section, we need the following lemma.

**Lemma 12.** Let $h(n) : \mathbb{N} \to \mathbb{R}^+$ be an increasing function of $n$ such that $\lim_{n \to \infty} h(n) = \infty$. Assume that each entry of the binary test matrix is an i.i.d. Bernoulli$(p)$ random variable, such that $[h(n)] p = o(n)$. Then $\forall i, d$
such that $1 \leq i \leq d \leq \lceil h(n) \rceil$, and $\forall \rho$ such that $0 < \rho < 1$, one has the following bound on the error exponent:

$$E_o(\rho, i, d, n) \geq \rho (1 - p)^{d ip} \left( 1 - \frac{p}{2} \log^2 (ip) + o(1) \right).$$

**Proof.** Given $h(n) : \mathbb{N} \mapsto \mathbb{R}^+$, construct the test matrix $C$ such that each entry follows an i.i.d. Bernoulli($p$) distribution, where $p = \frac{h(n)}{1+\epsilon'}$ for any fixed $\epsilon'$ such that $0 < \epsilon' < 1$. In order to prove this lemma, we use the results in [79, Lemma VII.1] and [79, Lemma VII.2] which state that $\forall i, d, n : 1 \leq i \leq d \leq n$,

$$E_o(\rho, i, d, n) \geq \rho I(T_1; Y|T_2) - \frac{\rho^2}{2} \max_{0 < \rho < 1} \left| \frac{\partial^2}{\partial \rho^2} E_o(\rho, i, d, n) \right|. \tag{6.18}$$

If $\mathbb{E}_{T_1} [u_\rho \log u_\rho]$ is a non-increasing function of $\rho$, then

$$\left| \frac{\partial^2}{\partial \rho^2} E_o(\rho, i, d, n) \right| \leq \sum_{T_2} \sum_{Y} g_\rho \mathbb{E}_{T_1} [u_\rho \log^2 (u_\rho)], \tag{6.19}$$

where we used the following notation

$$u_\rho = \frac{P(y|t_1, t_2)^{1/(1+\rho)}}{\sum_{T_1} P(t_1) P(y|t_1, t_2)^{1/(1+\rho)}},$$

$$g_\rho = \left( \sum_{T_1} P(t_1) P(y, t_2|t_1)^{1/(1+\rho)} \right)^{(1+\rho)}.$$

In order to simplify the previous equations, we consider different realizations of $u_\rho$ for different values of $y$, $t_2$ and $t_1$. In particular, we consider four cases based on the realizations of the pair $(y, t_2)$. For each case, we find $\mathbb{E}_{T_1} [u_\rho \log u_\rho]$ and show that this expectation is independent of $\rho$. In addition, when $\mathbb{E}_{T_1} [u_\rho \log u_\rho] \neq 0$, we find an expression for $g_\rho$.

**Case 1:** Let $y = 0$ and $t_2 = 0$. Then, we have

$$P(y = 0|t_1, t_2 = 0) = \begin{cases} 0 & \text{if } t_1 \neq 0 \\ 1 & \text{if } t_1 = 0, \end{cases}$$

105
which implies that

\[ u_\rho = \begin{cases} 
0 & \text{if } t_1 \neq 0 \\
\frac{1}{(1-p)^i} & \text{if } t_1 = 0.
\end{cases} \]

As a result,

\[
\mathbb{E}_{T_1}[u_\rho \log^2(u_\rho)] = \frac{(1-p)^i}{(1-p)^i} \log \left( \frac{1}{(1-p)^i} \right)
\]

\[
= \log^2 \left( \frac{1}{(1-p)^i} \right)
\]

\[
= i^2 \log^2(1 - p). \tag{6.20}
\]

Since \( P(y = 0, t_2 = 0 | t_1) = P(y = 0 | t_1, t_2 = 0) P(t_2 = 0) \), one has

\[
P(y = 0, t_2 = 0 | t_1) = \begin{cases} 
0 & \text{if } t_1 \neq 0 \\
P(t_2 = 0) & \text{if } t_1 = 0.
\end{cases}
\]

Consequently,

\[
g_\rho(y = 0, t_2 = 0) = \left( \sum_{t_1} P(t_1) P(y = 0, t_2 = 0 | t_1)^{1/(1+\rho)} \right)^{(1+\rho)}
\]

\[
= \left( P(t_1 = 0) P(t_2 = 0)^{1/(1+\rho)} \right)^{(1+\rho)}
\]

\[
= P(t_1 = 0)^{(1+\rho)} P(t_2 = 0). \tag{6.21}
\]

Note that (6.21) implies that \( g_\rho(y = 0, t_2 = 0) \) is a non-increasing function of \( \rho \).

**Case 2:** Let \( y = 1 \) and \( t_2 = 0 \). Then, we have

\[
P(y = 1 | t_1, t_2 = 0) = \begin{cases} 
1 & \text{if } t_1 \neq 0 \\
0 & \text{if } t_1 = 0,
\end{cases}
\]

which implies that

\[
u_\rho = \begin{cases} 
\frac{1}{1-(1-p)^i} & \text{if } t_1 \neq 0 \\
0 & \text{if } t_1 = 0.
\end{cases}
\]
As a result,

$$
E_{T_1}[u_\rho \log^2(u_\rho)] = \frac{1-(1-p)^i}{1-(1-p)^i} \log^2 \left( \frac{1}{1-(1-p)^i} \right)
= \log^2 (1-(1-p)^i).
$$

(6.22)

Since \( P(y=1, t_2=0|t_1) = P(y=1|t_1, t_2=0) P(t_2=0) \), one has

$$
P(y=1, t_2=0|t_1) = \begin{cases} P(t_2=0) & \text{if } t_1 \neq 0 \\ 0 & \text{if } t_1 = 0. \end{cases}
$$

Consequently,

$$
g_\rho(y=1, t_2=0) = \left( \sum_{t_1} P(t_1) P(y=1, t_2=0|t_1) \right)^{(1+\rho)}
= (P(t_1 \neq 0) P(t_2=0))^{1/(1+\rho)}
= P(t_1 \neq 0)^{(1+\rho)} P(t_2=0).
$$

(6.23)

Note that (6.23) implies that \( g_\rho(y=1, t_2=0) \) is a non-increasing function of \( \rho \).

**Case 3:** Let \( y = 0 \) and \( t_2 \neq 0 \). Then, we have

$$
P(y=0|t_1, t_2 \neq 0) = 0.
$$

Consequently, \( u_\rho = 0 \) and

$$
E_{T_1}[u_\rho \log^2(u_\rho)] = 0.
$$

(6.24)

**Case 4:** Let \( y = 1 \) and \( t_2 \neq 0 \). Then, we have

$$
P(y=1|t_1, t_2 \neq 0) = 1
$$

Consequently \( u_\rho = 1 \), and

$$
E_{T_1}[u_\rho \log^2(u_\rho)] = 0.
$$

(6.25)

In all these four cases, \( E_{T_1}[u_\rho \log u_\rho] \) is independent on \( \rho \) and therefore a
non-decreasing function of $\rho$. Substituting (6.20)-(6.25) into (6.19) gives

$$
\left| \frac{\partial^2 E_o(\rho, i, d, n)}{\partial \rho^2} \right| \leq \left| \sum_{T_2} \sum_Y g_\rho \mathbb{E}_{T_1} \left[ u_\rho \log^2(u_\rho) \right] \right|
$$

$$
= |g_\rho(y = 0, t_2 = 0) i^2 \log^2(1 - p) + g_\rho(y = 1, t_2 = 0) \log^2(1 - (1 - p)^i)|
$$

$$
= g_\rho(y = 0, t_2 = 0) i^2 \log^2(1 - p) + g_\rho(y = 1, t_2 = 0) \log^2(1 - (1 - p)^i),
$$

(6.26)

where the last equality follows since $g_\rho$ is non-negative. Since $g_\rho(y = 0, t_2 = 0)$ and $g_\rho(y = 1, t_2 = 0)$ are non-increasing functions in $\rho$, we have

$$
\max_{\rho \leq \rho < 1} g_\rho(y = 0, t_2 = 0) = g_0(y = 0, t_2 = 0)
$$

$$
= P(t_1 = 0)P(t_2 = 0)
$$

$$
= (1 - p)^d,
$$

(6.27)

and

$$
\max_{\rho \leq \rho < 1} g_\rho(y = 1, t_2 = 0) = g_0(y = 1, t_2 = 0)
$$

$$
= P(t_1 = 0)P(t_2 = 0)
$$

$$
= (1 - p)^d - (1 - (1 - p)^i)
$$

$$
= (1 - p)^d \frac{1 - (1 - p)^i}{(1 - p)^i}
$$

$$
= (1 - p)^d (1 - (1 - p)^i - 1)
$$

$$
= (1 - p)^d p(1 + o(1)).
$$

(6.28)

Consequently, using (6.26)-(6.28), it can be shown that

$$
\max_{\rho \leq \rho < 1} \left| \frac{\partial^2 E_o(\rho, i, d, n)}{\partial \rho^2} \right|
$$

$$
\leq \max_{\rho \leq \rho < 1} g_\rho(y = 0, t_2 = 0) i^2 \log^2(1 - p)
$$

$$
+ \max_{\rho \leq \rho < 1} g_\rho(y = 1, t_2 = 0) \log^2(1 - (1 - p)^i)
$$

$$
= i^2 (1 - p)^d \log^2(1 - p) + (1 - p)^d - (1 - (1 - p)^i) \log^2(1 - (1 - p)^i)
$$

$$
= (1 - p)^d (i^2 p^2 (1 + o(1)) + ip \log^2(1 - (1 - p)^i)(1 + o(1)))
$$

$$
= (1 - p)^d ip \log^2(ip) (1 + o(1)).
$$

(6.29)

Next, note that the mutual information in (6.18) can be bounded according
\[ I(T_1; Y|T_2) = H(Y|T_2) - H(Y|T_1, T_2) = 
= H(Y|T_2) = (1 - p)^{d-i} h((1 - p)^i) 
\geq (1 - p)^{d-i}(1 - p)^i \log(1 - p)^{-i} 
= (1 - p)^{d-i} p(1 + o(1)). \] (6.30)

Substituting (6.29) and (6.30) into (6.18) yields

\[ E_o(\rho, i, d, n) \geq \rho I(T_1; Y|T_2) - \frac{\rho^2}{2} \max_{\rho: 0 < \rho < 1} |E_o''(\rho, i, d, n)| \]
\[ \geq \rho(1 - p)^{d-i}p \left( 1 - \frac{\rho}{2} \log^2(ip) + o(1) \right). \]

We would like to point out that Lemma 12 is a generalization of a lower bound on \( E_o(\rho, i, d, n) \) from [79], as the bound in [79] does not apply directly to the Poisson PGT model.

In order to find the number of tests that guarantee \( P(\mathcal{E}') = o(1) \), we consider separately two asymptotic regimes for \( \lambda(n) \): Theorem 24 presents the results for the asymptotic regime \( \lambda(n) = o(n) \) and \( \lim_{n \to \infty} \lambda(n) = \infty \); similarly, Theorem 25 presents the results for the regime where \( \lambda(n) = o(n) \), but \( 0 < \lim_{n \to \infty} \lambda(n) < \infty \). Note that the case of constant \( \lambda \) is covered by the latter scenario.

**Theorem 24.** Assume that \( D \) follows the right-truncated Poisson distribution, with \( \lambda(n) = o(n) \) and \( \lim_{n \to \infty} \lambda(n) = \infty \). Construct a test matrix by choosing each entry according to a Bernoulli(\( p \)) distribution, where \( p = [\lambda(n)^{(1+\epsilon)}]^{-1+\gamma} \), for some fixed arbitrarily small scalars \( \epsilon > 0 \) and \( 0 < \gamma < 1 \). Under ML decoding, one can identify the set of defectives using \( m = (2 + \delta)\lambda(n)^{2+\alpha} \log n \) tests so that \( \lim_{n \to \infty} P(\mathcal{E}') = 0 \), for any fixed arbitrarily small scalars \( \delta > 0 \) and \( \alpha > 0 \).

**Proof.** Since \( \lambda(n) = o(n) \), there exists a fixed \( \epsilon > 0 \) small enough such that \( h(n) \leq \lambda(n)^{1+\epsilon} = o(n) \). Choose \( p = [h(n)]^{-1+\gamma} \), for some \( 0 < \gamma < 1 \). The probability of error given by formula (6.16) can be rewritten as \( P(\mathcal{E}') \leq \)
where

\[ P_{e_1} = \sum_{d=1}^{[h(n)]} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} P(\mathcal{E}'_i), \]

\[ P_{e_2} = \sum_{d=[h(n)]+1}^{n} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} P(\mathcal{E}'_i). \]

The idea is to bound these probabilities by finding a tight upper bound on \( P(\mathcal{E}'_i) \), independent of \( i \) and \( d \), for \( 1 \leq d \leq [h(n)] \), while using the upper bound \( P(\mathcal{E}'_i) \leq 1 \) for \([h(n)]+1 \leq d \leq n\). Since \( P(\mathcal{E}'_i) \leq 1 \), one has

\[ P_{e_2} \leq \sum_{d=[h(n)]+1}^{n} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \]

\[ \leq \sum_{d=[h(n)]+1}^{\infty} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \]

\[ = c\lambda(n) \sum_{d=[\lambda(n)^{1+\epsilon}]}^{\infty} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)}. \] (6.31)

The Chernoff bound for standard Poisson distributions ensures that for any \( a \geq 0 \),

\[ \sum_{d=[\lambda(n)^{1+\epsilon}]+a}^{\infty} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \leq \exp\left(-\left(\lambda(n) + a\right) \log \frac{\lambda(n) + a}{\lambda(n)} + a\right). \] (6.32)

Substituting \( a=[\lambda(n)^{1+\epsilon}]-\lambda(n) \) for \( \epsilon > 0 \) yields

\[ \sum_{d=[\lambda(n)^{1+\epsilon}]+a}^{\infty} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \leq \exp\left(-\left[\lambda(n)^{1+\epsilon}\right] \log \lambda(n)^\epsilon + a\right) \]

\[ = \exp\left(-\epsilon\left[\lambda(n)^{1+\epsilon}\right] \log \lambda(n) \left(1 + o(1)\right)\right). \] (6.33)

Consequently, substituting (6.33) in (6.31) yields \( P_{e_2} = o(1) \).

The goal is to find the smallest value of \( m \) such that \( P_{e_1} = o(1) \). Since we have chosen \( p = [h(n)]^{-(1+\gamma)} \), for some \( 0 < \gamma < 1 \), using Lemma 12 one can
show that $\forall i, d, 1 \leq i \leq d \leq [h(n)]$ and $\forall \rho$, $0 < \rho < 1$,

$$E_o(\rho, i, d, n) \geq \rho (1-p)^{d_i} \left(1 - \frac{\rho}{2} \log^2(ip) + o(1)\right)$$

$$\geq \rho p (1-p)^{[h(n)]} \left(1 - \frac{\rho}{2} \log^2(p) + o(1)\right)$$

$$= \rho p e^{-p [h(n)]} \left(1 - \frac{\rho}{2} \log^2(p) + O(1)\right)$$

$$\geq \rho \left(1 + o(1)\right) \left(1 - \frac{\rho}{2} (1+\gamma)^2 \log^2([h(n)])\right)$$

$$\geq \rho \frac{\left(1 + o(1)\right)}{1 + o(1)} \left(1 - \frac{\rho}{2} (1+\gamma)^2 \log^2(h(n)+1)\right).$$

By choosing $\rho = \frac{1}{(1+\gamma)^2 \log^2(h(n)+1)}$, one arrives at

$$E_o(\rho, i, d, n) \geq \frac{1}{2} \rho \left(1 + o(1)\right),$$

for any $i, d$ such that $1 \leq i \leq d \leq [h(n)]$. In addition, using the inequality $\binom{d}{i} \leq \left(\frac{d}{i}\right)^i$, it can be easily shown that for $1 \leq i, d \leq [h(n)]$, it holds that

$$\log \binom{n-d}{i} \leq i \log \left(\frac{d(n-d) e^2}{i^2}\right)$$

$$\leq i \log (dn e^2)$$

$$\leq h(n) \log n (1 + o(1)).$$

As a result, if

$$m > \frac{\rho h(n) \log n}{\frac{1}{2} \rho (h(n)+1)^{1+\gamma}} (1 + o(1)) = 2h(n)^{2+\gamma} \log n (1 + o(1)),$$

then $\left(E_o(\rho, i, d, n) - \frac{o \log \binom{n-d}{i}}{m}\right)$ is positive. Therefore, using $m \geq (2+\delta) h(n)^{2+\gamma} \log n$
for any fixed $\delta > 0$, we may write

$$P(\mathcal{E}'_i) \leq 2^{-m \left( E_o(\rho, i, d, n) - \frac{\rho \log \left( \frac{n-d}{m} \right)}{m} \right)}$$

$$\leq 2^{-m \rho \left( \frac{h(n)+1}{h(n)+1+\gamma} - \frac{\log n}{m} \right)}$$

$$= 2^{-m \rho \left( \frac{\log h(n)+1}{2m(h(n)+1+\gamma)} \right)}$$

$$= 2^{-\rho \left( \log h(n) + 1 + o(1) \right)}$$

$$= 2 \cdot 2^{-\rho \left( \log h(n) + 1 + o(1) \right)} = P_1(n).$$

Since

$$P_{e_1} = \sum_{d=1}^{[h(n)]} \sum_{i=1}^{d} c(n) \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} P(\mathcal{E}'_i)$$

$$\leq c(n) P_1(n) \sum_{d=1}^{[h(n)]} d \frac{\lambda(n)^d}{d!} e^{-\lambda(n)}$$

$$\leq c(n) P_1(n) \sum_{d=1}^{\infty} d \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} = c(n) P_1(n) \lambda(n),$$

it follows that

$$P_{e_1} \leq c(n) \lambda(n) P_1(n)$$

$$= 2^{-\rho \left( \log h(n) + 1 + o(1) \right)}$$

$$= 2 \cdot 2^{-\rho \left( \log h(n) + 1 + o(1) \right)} = o(1). \quad (6.34)$$

Consequently, the probability of error converges to zero, i.e., $P(\mathcal{E}') = o(1)$ if $m = (2 + \delta)h(n)^{2+\gamma} \log n$, for any fixed $\delta > 0$ and $\gamma > 0$. Substituting $h(n) = \lambda(n)^{1+\epsilon}$ in the previous expression, and performing some straightforward simplifications yield $m = (2 + \delta)\lambda(n)^{2+\alpha} \log n$, for $\delta > 0$ and $\alpha > 0$. \qed

**Theorem 25.** Assume that $D$ follows the right-truncated Poisson distribution, with $\lambda(n) = o(n)$ and $0 < \lim_{n \to \infty} \lambda(n) < \infty$. Let $\beta(n) = \log^{(K)} n$, for some finite $K > 1$. Construct a test matrix by choosing each entry according to a Bernoulli($p$) distribution, where $p = \left[ \beta(n) \lambda(n) \right]^{-1+\gamma}$. for some fixed arbitrarily small scalar $\gamma > 0$. Using ML decoding, one can identify defectives using $m = (2 + \delta)\lambda(n)^{2+\gamma} \log n$ tests such that $\lim_{n \to \infty} P(\mathcal{E}') = 0$, for any fixed arbitrarily small scalar $\delta > 0$. 

112
Proof. Let \( h(n) \equiv \beta(n) \lambda(n) \). Similar to the proof of Theorem 24, we write \( P(\mathcal{E}') \leq P_{e_1} + P_{e_2} \), where

\[
P_{e_1} = \sum_{d=1}^{[h(n)]} \sum_{i=1}^{d} c \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)} P(\mathcal{E}_i'),
\]

\[
P_{e_2} = \sum_{|h(n)|+1}^{n} \sum_{i=1}^{d} c \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)} P(\mathcal{E}_i').
\]

Since \( P(\mathcal{E}_i') \leq 1 \), one has

\[
P_{e_2} \leq \sum_{d=|h(n)|+1}^{n} c d \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)}
\]

\[
\leq \sum_{d=|h(n)|+1}^{\infty} c d \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)}
\]

\[
= c \lambda(n) \sum_{d=|\beta(n) \lambda(n)|}^{\infty} \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)}. \tag{6.35}
\]

Since \( \lim_{n \to \infty} \beta(n) = \infty \), there exists an integer \( n' \) large enough such that \( \beta(n) > 1 \), for all \( n > n' \). Let \( n > n' \) and substitute \( a = [\beta(n) \lambda(n)] - \lambda(n) \) in (6.32). Then,

\[
\sum_{d=|\beta(n) \lambda(n)|}^{\infty} \frac{\lambda(n)^d}{d^!} e^{-\lambda(n)} \leq \exp (-[\beta(n) \lambda(n)] \log \beta(n) + [\beta(n) \lambda(n)] - \lambda(n)). \tag{6.36}
\]

As a result, substituting (6.36) in (6.35) yields \( P_{e_2} = o(1) \).

Now, we find \( m \) such that \( p_m = o(1) \). The ideas behind the proof are similar to those described in the proof of Theorem 24, except that in this case, we set \( h(n) = \beta(n) \lambda(n) \) and \( p = [h(n)]^{-(1+\gamma)} = [\beta(n) \lambda(n)]^{-(1+\gamma)} \), for some \( 0 < \gamma < 1 \). As a result, \( P(\mathcal{E}') = o(1) \), if \( m = (2 + \delta)h(n)^{2+\gamma} \log n \), for any fixed \( \delta > 0 \) and \( \gamma > 0 \). Substituting \( h(n) = \beta(n) \lambda(n) \) in the previous expression, one arrives at \( m = (2 + \delta)(\beta(n) \lambda(n))^{2+\gamma} \log n \), for \( \delta > 0 \) and \( \gamma > 0 \).

6.4 Semi-adaptive methods for Poisson PGT

An alternative to both adaptive and non-adaptive GT approaches is semi-adaptive testing. A semi-adaptive GT algorithm is an algorithm in which
tests are designed in several stages. The tests in each stage are constructed in a nonadaptive manner and therefore can be performed in parallel. However, the set of subjects on which the tests are performed changes from one stage to the next; in other words, the results obtained during one stage of testing may guide the choice of test subjects and potential defectives in the next stage. One of the best known semi-adaptive algorithms is the original 2-stage algorithm proposed by Dorfman [16].

In the absence of error, a semi-adaptive algorithm is expected to identify all defectives, even if no prior knowledge regarding the number of defectives is available. As a result, unlike the case of nonadaptive algorithms in which one seeks to find a number of tests \( m \) for which \( \lim_{n \to \infty} P(\mathcal{E}) = 0 \), in semi-adaptive framework one is interested in the expected number of tests \( \bar{m} \) that an algorithm performs in order to identify the defectives with zero probability of error, i.e., with \( P(\mathcal{E}) = 0 \). In what follows, we first find a lower bound on \( \bar{m} \) for any adaptive (and hence, semi-adaptive) algorithm for Poisson PGT using Huffman coding. Then, we devise a semi-adaptive algorithm and show that for this algorithm, \( \bar{m} \) is only a constant factor away from the lower bound.

### 6.4.1 Lower bound on the expected number of tests

Suppose that the number of defectives follows the truncated Poisson distribution; in addition, assume that for any fixed \( 1 \leq d \leq n \), all the sets of \( D = d \) defectives are equally likely.

In what follows, we show that one can use Huffman source coding [75] to find a lower bound on the expected number of adaptive tests required to identify the defectives. Let \( w \in \{0, 1\}^n \) be a binary random vector such that \( w(i) = 1 \) if the \( i^{th} \) subject is a defective, and \( w(i) = 0 \) otherwise. There are \( 2^n \) choices for \( w \), contained in a set denoted by \( \mathcal{W} \). An adaptive GT algorithm has to identify the true realization of \( w \), denoted by \( w_t \), using a set of tests. Each such test can be represented as a “yes/no” query of the form “is \( w_t \) a member of the set \( \mathcal{W}' \)?”, where the set \( \mathcal{W}' \subseteq \mathcal{W} \) is determined by the design of the test. For example for \( n = 5 \), the query corresponding to a test that
contains the first, the fourth and the fifth subjects asks if \( w_t \in W' \), where

\[
W' = \{ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \}.
\]

If the output of the test is 0, the answer to the query is “yes”, since none of the three subjects in the test are defective and therefore \( w_t \in W' \); otherwise the answer to the query is “no” which implies that \( w_t \in W \setminus W' \). On the other hand, it can be easily verified that not every possible subset query corresponds to a group test [7, 80]. As a result, the minimum expected number of subset queries required to identify \( w_t \) provides a lower bound on the minimum expected number of group tests required to identify \( w_t \) in an adaptive manner. One should note that the minimum expected number of queries of the form above is equal to the expected length of a Huffman code designed for a source with alphabet \( W \) and the corresponding probability distribution [75].

For a fixed \( 0 \leq d \leq n \), let \( w_{d,j}, j = 1, 2, \ldots, \binom{n}{d} \) be a realization of \( w \) with exactly \( d \) entries equal to 1. As a result, the alphabet of the source \( w \) is of the form \( W = \{ w_{d,j} \}, j = 1, 2, \ldots, \binom{n}{d}, d = 0, 1, \ldots, n \). It follows that for all \( 0 \leq d \leq n \) and for all \( 1 \leq j \leq \binom{n}{d} \),

\[
P(w = w_{d,j}) = \frac{1}{\binom{n}{d}} P(D = d) = \frac{c(n)}{n!} \binom{n - d}{d} ! \lambda(n)^d e^{-\lambda(n)} = P(w_d).
\]

**Theorem 26.** Let \( \lambda(n) = o(n) \). Then, the minimum expected number of tests in an adaptive (and semi-adaptive) group testing algorithm satisfies \( \bar{m} > \lambda(n) \log_2 \frac{n}{\lambda(n)} (1 + o(1)) - \log_2 e^{-\lambda(n)^d / n^d} \). In addition, if \( \lambda(n) = o((n^2 \log_2 n)^{1/3}) \), this lower bound simplifies to \( \bar{m} > \lambda(n) \log_2 \frac{n}{\lambda(n)} (1 + o(1)) \).

**Proof.** To prove this theorem, we note that the Shannon entropy [75] of the source, \( H(w) \), provides a lower bound on the average length of the optimum
Huffman code. Consequently, using (6.37), one has

\[
\bar{m} \geq H(w) = - \sum_{d=0}^{n} \sum_{j=1}^{(n)} P(w_{d,j}) \log_2 P(w_{d,j}) \\
= - \sum_{d=0}^{n} \binom{n}{d} P(w_d) \log_2 P(w_d) \\
= - \sum_{d=0}^{n} P(D = d) \log_2 P(w_d) \\
= \mathbb{E}_D \left[ \log_2 \frac{1}{P(w_d)} \right].
\]

By invoking (6.37), the previous expression may be rewritten as

\[
\bar{m} \geq \mathbb{E}_D \left[ \lambda(n) \log_2 e - \log_2 c(n) + \log_2 \frac{n!}{(n-d)!} - d \log_2 \lambda(n) \right] \\
\geq \lambda(n) \log_2 e - \log_2 c + \mathbb{E}_D \left[ d \log_2 (n-d+1) - d \log_2 \lambda(n) \right] \\
= \lambda \log_2 e - \log_2 c + \log_2 \frac{n}{\lambda(n)} \mathbb{E}[D] + \mathbb{E}_D \left[ d \log_2 \left( 1 - \frac{d-1}{n} \right) \right] \\
\geq \lambda(n) \log_2 e - \log_2 c + \log_2 \frac{n}{\lambda(n)} \mathbb{E}[D] - \log_2 e \left( \frac{d(d-1)}{n-d+1} \right), \quad (6.38)
\]

where the last inequality follows since \( \log(1 + x) \geq \frac{x}{1+x} \), for any \( x > -1 \). Next, note that

\[
\mathbb{E}_D \left[ \frac{d(d-1)}{n-d+1} \right] = c(n) \sum_{d=0}^{n} \frac{d(d-1)}{n-d+1} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \\
= \lambda(n)^2 c(n) \sum_{d=0}^{n-2} \frac{\lambda(n)^d}{d!(n-d-1)} e^{-\lambda(n)}. \quad (6.39)
\]
For any $d$ such that $0 \leq d \leq n - 2$, one has

\[
\frac{1}{n-d-1} = \frac{1}{n} \sum_{i=0}^{\infty} \left( \frac{d+1}{n} \right)^i \\
= \frac{1}{n} \left( 1 + \frac{d+1}{n} + \frac{(d+1)^2}{n^2} \sum_{i=0}^{\infty} \left( \frac{d+1}{n} \right)^i \right) \\
= \frac{1}{n} \left( 1 + \frac{d+1}{n} + \frac{(d+1)^2}{n(n-d-1)} \right) \\
\leq \frac{1}{n} \left( 1 + \frac{d+1}{n} + \frac{(d+1)^2}{n} \right) \\
= \frac{1}{n^2} ((n+2) + 4d + d(d-1)). \tag{6.40}
\]

Substituting (6.40) in (6.39) yields

\[
\mathbb{E}_D \left[ \frac{d(d-1)}{n - d + 1} \right] = c\lambda(n) \frac{2n^2}{n^2} + 2 \sum_{d=0}^{n-2} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \\
+ 4c \frac{\lambda(n)^3}{n^2} \sum_{d=0}^{n-3} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} + c \frac{\lambda(n)^4}{n^2} \sum_{d=0}^{n-4} \frac{\lambda(n)^d}{d!} e^{-\lambda(n)} \\
< \frac{\lambda(n)^2}{n^2} (n + 2 + 4\lambda(n) + \lambda(n)^2) \\
= \frac{\lambda(n)^2}{n} (1 + o(1)) + \frac{\lambda(n)^4}{n^2}. \tag{6.41}
\]

Substituting (6.3) and (6.41) in (6.38), and by recalling that $\lambda(n) = o(n)$, one arrives at

\[
\bar{m} \geq \lambda(n) \log_2 e^{-\log_2 c + \log_2 \frac{n}{\lambda(n)} \mathbb{E}[D] - \log_2 e \mathbb{E}_D \left[ \frac{d(d-1)}{n - d + 1} \right]} \\
> \lambda(n) \log_2 \frac{n}{\lambda(n)} (1 + o(1)) - \log_2 e \frac{\lambda(n)^4}{n^2}. 
\]

If $\lambda(n) = o((n^2 \log_2 n)^{1/3})$, this bound simplifies to $\bar{m} > \lambda(n) \log_2 \frac{n}{\lambda(n)} (1 + o(1))$. \hfill \Box

6.4.2 Constructive upper bound on expected number of tests using an $s$-stage algorithm

In [81], Li proposed an $s$-stage algorithm to identify $d$ defectives in a combinatorial group testing framework. In what follows, we modify his algorithm
and show that the expected number of tests performed by $s$-stage testing allows one to find all the defectives in a Poisson PGT model, while being only a constant away from the lower bound of Theorem 26.

Let $s = s(n, \lambda(n))$ denote the total number of stages. Also, let $\mathcal{S}_i$, $1 \leq i \leq s$, be the set of potential defectives at stage $i$ on which the group tests are performed. In the first stage, we set $\mathcal{S}_1 = \mathcal{S}$, where $\mathcal{S}$ is the set of all subjects, $|\mathcal{S}| = n$. Then, we randomly divide $\mathcal{S}_1$ into disjoint sets of size $k_1$, where $k_1 = k_1(\lambda(n), n)$. If $k_1$ does not divide $|\mathcal{S}_1|$, one set will contain fewer than $k_1$ entries, equal to the remainder of dividing $|\mathcal{S}_1|$ by $k_1$. A test is performed on each of these sets independently. In the second stage, $\mathcal{S}_2$ is formed by pooling all the subjects in sets with a positive test outcome in the first stage. Similarly, the set $\mathcal{S}_2$ is randomly divided into disjoint sets of size $k_2$. Again, one set may contain fewer subjects as compared to the other sets, and a test is performed on each set. The procedure continues in the same manner up to stage $s - 1$. In the last stage, $\mathcal{S}_s$ is formed by pooling all the subjects in sets with a positive test outcome at stage $s - 1$; then, each remaining subject is tested individually to determine if it is defective. The following theorem shows that proper choices of $s$ and $k_i$, $1 \leq i \leq s - 1$, may guarantee that the expected number of tests performed using this algorithm is upper bounded by a value only a constant away from the lower bound.

**Theorem 27.** Let $\lambda(n) = o(n)$ and let $\tilde{\lambda}(n) = \mathbb{E}_D[d]$. Then, by choosing $s_0 = \log \frac{n}{\lambda(n)}$, $s = \lceil s_0 \rceil$, and $k_i = \left( \frac{n}{\lambda(n)} \right)^{\frac{s_0 - 1}{s_0}}$, for $1 \leq i \leq s - 1$, the expected number of the proposed semi-adaptive group testing algorithm satisfies

$$m \leq \frac{e}{\log_2 e} \tilde{\lambda}(n) \log_2 \left( \frac{n}{\lambda(n)} \right) (1 + o(1))$$

$$= \frac{e}{\log_2 e} \lambda(n) \log_2 \left( \frac{n}{\lambda(n)} \right) (1 + o(1)),$$

where $\frac{e}{\log_2 e} \approx 1.884$.

**Proof.** Assume that $D = d$ is the number of defectives. In the first stage, divide the test subjects into disjoint groups of size $k_1$. This leads to $\lceil \frac{n}{k_1} \rceil$ tests. In the $i^{th}$ stage, $1 \leq i \leq s - 2$, at most $d$ tests are positive, with the upper bound achieved when each defective is in a different group; as a result, the number of remaining subjects and the number of tests in the $(i + 1)^{th}$ stage are at most $dk_i$ and $\lceil d \frac{k_i}{k_{i+1}} \rceil$, respectively. In the last stage, the number
of remaining subjects and the number of tests both equal to \( dk_{s-1} \). Hence, the total number of tests is bounded as

\[
m \leq \left\lfloor \frac{n}{k_1} \right\rfloor + \sum_{i=2}^{s-1} \left\lfloor d \frac{k_{i-1}}{k_i} \right\rfloor + d k_{s-1}.
\]

Consequently since \( s \) and \( k_i, 1 \leq i \leq s - 1 \), do not depend on \( d \), one has

\[
\bar{m} = \mathbb{E}_D[m] \leq \left\lfloor \frac{n}{k_1} \right\rfloor + \sum_{i=2}^{s-1} \mathbb{E}_D \left[ \left\lfloor d \frac{k_{i-1}}{k_i} \right\rfloor \right] + \tilde{\lambda}(n) k_{s-1}
\]

\[
\leq \left\lfloor \frac{n}{k_1} \right\rfloor + \sum_{i=2}^{s-1} \mathbb{E}_D \left[ d \frac{k_{i-1}}{k_i} + 1 \right] + \tilde{\lambda}(n) k_{s-1}
\]

\[
= \left\lfloor \frac{n}{k_1} \right\rfloor + \sum_{i=2}^{s-1} \tilde{\lambda}(n) \frac{k_{i-1}}{k_i} + \tilde{\lambda}(n) k_{s-1} + s - 2.
\]

Substituting \( s \) and \( k_i \) in the previous expressions, one obtains

\[
\bar{m} \leq \left\lfloor \frac{n}{k_1} \right\rfloor + \sum_{i=2}^{s-1} \tilde{\lambda}(n) \frac{k_{i-1}}{k_i} + \tilde{\lambda}(n) k_{s-1} + s - 2
\]

\[
\leq \tilde{\lambda}(n) \left( \frac{n}{\tilde{\lambda}(n)} \right)^{\frac{1}{2}} (1 + o(1)) + \sum_{i=2}^{s-1} \tilde{\lambda}(n) \left( \frac{n}{\tilde{\lambda}(n)} \right)^{\frac{1}{2}} (1 + o(1))
\]

\[
+ \tilde{\lambda}(n) \left( \frac{n}{\tilde{\lambda}(n)} \right)^{\frac{1}{2}} \log \frac{n}{\tilde{\lambda}(n)}(1 + o(1))
\]

\[
\leq e(s - 1) \tilde{\lambda}(n) (1 + o(1)) + \tilde{\lambda}(n) e^2 + \log \frac{n}{\tilde{\lambda}(n)}(1 + o(1))
\]

\[
\leq e \tilde{\lambda}(n) \log \frac{n}{\tilde{\lambda}(n)} (1 + o(1))
\]

\[
= \frac{e}{\log_2 e} \tilde{\lambda}(n) \log_2 \left( \frac{n}{\tilde{\lambda}(n)} \right) (1 + o(1)).
\]

\[\square\]

### 6.5 Summary of the results and discussion

In the previous sections, we introduced the Poisson probabilistic group testing framework for modeling the number of defectives according to a random variable following a right-truncated Poisson distribution. For the proposed model and under the assumption that \( \lambda(n) = o(n) \), we considered nonadaptive and semi-adaptive methods to identify the defectives. These methods
are based on generalization of combinatorial GT schemes, which to the best of our knowledge are used in the context of probabilistic GT for the first time.

In Section 6.3.1, we used information theoretic arguments to derive a lower bound on the number of tests (Thm. 17). In addition, we derived constructive upper bounds on the number of tests using practical testing schemes (Thms. 18-23) and information theoretic arguments (Thms. 24 and 25). The results under the assumption that the vector of test results is error-free are summarized in Table 6.1. In the table, \( \beta(n) \) is used to represent the slowly-growing function defined in (6.1), and \( \epsilon, \alpha, \delta, \) and \( \gamma \) are arbitrary small positive constants. In Thms. 20 and 21, we considered the case in which there are at most \( v(n) \) errors in the vector of test results and showed that

\[
m = (2e \beta(n) \lambda(n))^2 \log n + 4e v(n) \beta(n) \lambda(n)) (1 + o(1)) \text{ tests are sufficient}
\]

to identify the defectives using a decoder with computational complexity of

\[
O(mn) \text{ if } 0 < \lim_{n \to \infty} \lambda(n) < \infty.
\]

Similarly, we showed that if \( \lim_{n \to \infty} \lambda(n) = \infty \), the same decoder requires

\[
m = (2e \lambda(n)^2(1+\epsilon) \log n + 4e v(n) \lambda(n)^{1+\epsilon}) (1 + o(1)) \text{ tests.}
\]

The test constructions and decoding algorithms used in Thms. 18-21 rely on designing test matrices that can identify the defectives with zero error probability as long as \( D \leq \Delta \), for an appropriate choice of \( \Delta \). However, it is well-known that the minimum number of tests for these matrices satisfies\(^2\) [82]

\[
\frac{\Delta^2}{2 \log_2 \Delta} \log_2 n (1+o(1)) \leq m \leq \Delta^2 \log_2 e \log_2 n (1+o(1)).
\]

It is not difficult to show that \( \Delta \) must be larger than \( \lambda \) in order to have \( \lim_{n \to \infty} P(D > \Delta) = 0 \). As a result, by requiring \( P(\mathcal{E}|D \leq \Delta) = 0 \), one cannot obtain upper bounds on the number of tests for Poisson PGT that match the lower bound in Thm. 17. In order to overcome this problem, we instead used the less stringent condition \( \lim_{n \to \infty} P(\mathcal{E}|D \leq \Delta) = 0 \) in Thms. 22 and 23, and employed the results of [76] to obtain matching upper bounds on \( m \). One should note that there exist other test constructions and decoding algorithms that may be used in conjunction with \( \lim_{n \to \infty} P(\mathcal{E}|D \leq \Delta) = 0 \) to

---

\(^2\)One should note that these bounds correspond to disjunct matrices. One can relax the disjunct property and yet achieve zero-error probability in CGT using the so-called separable matrices, which lead to the same asymptotic behavior as disjunct matrices [7].
Table 6.1: Lower and upper bounds on the minimum number of measurements \( m \) using nonadaptive methods

<table>
<thead>
<tr>
<th>Theorem</th>
<th>Number of tests</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thm. 17</td>
<td>( m \geq (1-\epsilon)\lambda \log_2 n (1-o(1)) )</td>
<td>( \lambda = o(n) )</td>
</tr>
<tr>
<td>Thm. 18</td>
<td>( m \leq e \lambda^2 \epsilon \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty )</td>
</tr>
<tr>
<td>Thm. 19</td>
<td>( m \leq e^2 \lambda^2 \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty ), ( 0 &lt; \lim_{n \to \infty} \lambda &lt; \infty )</td>
</tr>
<tr>
<td>Thm. 22</td>
<td>( m \leq \frac{3}{\log_2 3} \lambda \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty )</td>
</tr>
<tr>
<td>Thm. 23</td>
<td>( m \leq \frac{3}{\log_2 3} \beta(n) \lambda \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty ), ( 0 &lt; \lim_{n \to \infty} \lambda &lt; \infty )</td>
</tr>
<tr>
<td>Thm. 24</td>
<td>( m \leq (2+\delta) \lambda^2 \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty )</td>
</tr>
<tr>
<td>Thm. 25</td>
<td>( m \leq (2+\delta)(\beta \lambda)^2 \log_2 n (1+o(1)) )</td>
<td>( \lambda = o(n) ), ( \lim_{n \to \infty} \lambda = \infty ), ( 0 &lt; \lim_{n \to \infty} \lambda &lt; \infty )</td>
</tr>
</tbody>
</table>

Table 6.2: Lower and upper bounds on the minimum expected number of measurements \( \bar{m} \) using semi-adaptive methods

<table>
<thead>
<tr>
<th>Theorem</th>
<th>Number of tests</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thm. 26</td>
<td>( \bar{m} &gt; \lambda \log_2 \frac{n^2}{\lambda} (1+o(1)) )</td>
<td>( \lambda = o(n) )</td>
</tr>
<tr>
<td>Thm. 26</td>
<td>( \bar{m} &gt; \lambda \log_2 \frac{n^2}{\lambda} (1+o(1)) )</td>
<td>( \lambda = o(n) )</td>
</tr>
<tr>
<td>Thm. 27</td>
<td>( \bar{m} \leq e \lambda \log_2 \frac{n^2}{\lambda} (1+o(1)) )</td>
<td>( \lambda = o(n) )</td>
</tr>
</tbody>
</table>

obtain matching upper bounds on \( m \) for Poisson PGT (see for example [83] and [84]); however, since an approach similar to the proof of Thms. 22 and 23 can be used in these cases as well, we choose not to repeat these arguments and results.

In the second part of our exposition (Sec. 6.4), we focused on the family of semi-adaptive algorithms. These algorithms are performed in sequential stages, allowing to design new tests based on the outcome of previous tests in order to decrease the expected number of tests; in addition, in each stage the tests are designed and performed simultaneously, allowing parallel testing. In Sec. 6.4, we used Huffman source coding to find a lower bound on the expected number of tests; in addition, we showed how Li’s stage-wise algorithm [81] developed for combinatorial GT can be modified for the Poisson
PGT model. These lower and upper bounds are listed in Table 6.2.

Recent work in the area of group testing has almost exclusively focused on combinatorial GT. The results derived in this chapter show that there exists a close connection between methods used for combinatorial GT and probabilistic GT.
CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we introduced semi-quantitative group testing, a novel framework that unifies and generalizes many group testing models in the literature and introduces many new interesting open problems with a wide range of applications. In this work, we showed that conventional group testing, quantitative group testing, threshold group testing, and many other models fall under the framework of SQGT, and analyzing these models in one unifying framework can improve upon the known pooling designs and decoding algorithms. Then, we focused on the most general SQGT model, which does not enforce any constraint on the choice of thresholds, i.e. their number and values can be arbitrary. For this general model, we introduced two families of test matrices, SQ-disjunct and SQ-separable, that are capable of identifying the defectives in the presence of errors. We described different explicit and probabilistic construction of these test matrices considering both binary and non-binary alphabets. Next, we focused on the important special case of SQGT with equidistant thresholds and described test matrix constructions and decoding algorithms based on message passing on factor graphs. Finally, we introduced the novel probabilistic group testing framework of Poisson group testing, applicable to dynamic testing with diminishing relative rates of defectives. For this new model, we described both nonadaptive and adaptive testing schemes and derived lower bounds and tight constructive upper bounds on the number of required tests.

In all these constructions, we assumed that the unknown vector $w$ is a binary vector representing which subject is defective and which subject is not. Although this model is sufficient in many applications, it would be useful to consider the case in which the unknown vector is non-binary but has a small alphabet size. While some of the constructions discussed in this dissertation may be applicable to this model, it is not clear if one can generalize all these constructions for the case of an unknown non-binary vector. As a result,
new methods may be necessary to construct proper test matrices and design efficient decoders.

The SQGT model is inspired by applications in which the result of a test only depends on the number of defective subjects present in that test. However, in many applications an effect called “dilution” happens, which implies that the result of a test not only depends on the number of defectives, but also depends on the number of subjects present in the test. Given a model for the effect of non-defective subjects on the result of each test, it would be interesting to design test matrices that are capable of identifying the defectives even in the presence of this dilatation effect.

Another interesting line of research corresponds to application of the SQGT model to machine learning and particularly learning interpretable rules for prediction [85]. The motivation in this line of research comes from the recent growing belief that in the face of high complexity and large datasets, checklists and other simple scorecards can significantly improve people’s performance on decision-making tasks [86]. For example, in medicine the clinical prediction rule is a simple decision-making rubric that helps physicians estimate the likelihood of a patient having or developing a particular condition in the future. Recent machine learning research has attempted to learn clinical prediction rules that generalize accurately from large-scale electronic health record data rather than relying on manual development [87, 88]. The key aspect of the problem is maintaining the simplicity and interpretability of the learned rule rather than a complicated, uninterpretable “black-box” model. Such transparency is critical for trust and adoption by users. In [85], the authors developed a method for learning interpretable clinical prediction rules using conventional group testing techniques. In this model, the authors developed a classifier in the form of a sparse AND-rule or OR-rule. However, in clinical prediction rule learning, one is interested in finding a sparse set of medical conditions or features with small integer coefficients that are added together to produce a score. Such a model is between the “1-of-N” and “N-of-N” forms implied by OR-rules and AND-rules and falls under the category of SQGT. In this application, the thresholds of the SQGT, in addition to the features which will form the rule sets, are learnt from training data. Some preliminary results corresponding to this application can be found in [89].
REFERENCES


128


