Algorithmica (2013) 67:384–417
DOI 10.1007/s00453-013-9754-7

Improved Constructions for Non-adaptive Threshold Group Testing

Mahdi Cheraghchi

Received: 30 December 2011 / Accepted: 2 February 2013 / Published online: 12 February 2013

Abstract The basic goal in combinatorial group testing is to identify a set of up to d defective items within a large population of size $n \gg d$ using a pooling strategy. Namely, the items can be grouped together in pools, and a single measurement would reveal whether there are one or more defectives in the pool. The threshold model is a generalization of this idea where a measurement returns positive if the number of defectives in the pool reaches a fixed threshold $u > 0$, negative if this number is no more than a fixed lower threshold $\ell < u$, and may behave arbitrarily otherwise. We study non-adaptive threshold group testing (in a possibly noisy setting) and show that, for this problem, $O(d^{g+2}(\log d)\log(n/d))$ measurements (where $g := u - \ell - 1$ and u is any fixed constant) suffice to identify the defectives, and also present almost matching lower bounds. This significantly improves the previously known (non-constructive) upper bound $O(d^{u+1}\log(n/d))$. Moreover, we obtain a framework for explicit construction of measurement schemes using lossless condensers. The number of measurements resulting from this scheme is ideally bounded by $O(d^{g+3}(\log d)\log n)$. Using state-of-the-art constructions of lossless condensers, however, we obtain explicit testing schemes with $O(d^{g+3}(\log d)\mathsf{quasipoly}(\log n))$ and $O(d^{g+3+\beta}\mathsf{poly}(\log n))$ measurements, for arbitrary constant $\beta > 0$.

Keywords Threshold group testing · Lossless expanders · Combinatorial designs · Explicit constructions

Part of work was done while the author was with the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Research was supported in part by the ERC Advanced investigator grant 228021 of A. Shokrollahi, and the Swiss National Science Foundation research grant PBELP2-133367. A preliminary summary of this work appears (under the same title) in proceedings of the 37th International Colloquium on Automata, Languages and Programming (ICALP 2010), Bordeaux, France, July 2010 [11].

M. Cheraghchi (✉)
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: cheraghchi@cmu.edu

1 Introduction

Combinatorial group testing is a classical problem that deals with identification of sparse Boolean vectors using disjunctive queries. Suppose that among a large set of n items it is suspected that, for some *sparsity parameter* $d \ll n$, up to d items might be "defective". In technical terms, defective items are known as *positives* and the rest are called *negatives*. In a *pooling strategy*, the items may be arbitrarily grouped in pools, and a single "measurement" reveals whether there is one or more positives within the chosen pool. The basic goal in group testing to design the pools in such a way that the set of positives can be identified from a number of measurements that is substantially less than n.

Since its introduction in 1940's [16], group testing and its variations have been extensively studied and have found surprisingly many applications in seemingly unrelated areas. In particular, we mention applications in molecular biology and DNA library screening (cf. [3, 24, 31, 33, 38, 44, 45] and the references therein), multiaccess communication [43], data compression [28], pattern matching [12], streaming algorithms [13], software testing [2], compressed sensing [14], and secure key distribution [6], among others. We refer the reader to [17, 18] for an extensive review of the major results in this area.

Formally, in classical group testing one aims to learn an unknown Boolean vector $(x_1, \dots, x_n) \in \{0, 1\}^n$ which is known to be d-sparse (that is, contains at most d nonzero entries) using a set of m measurements, where each measurement is defined by a subset of the coordinates $\mathcal{I} \subseteq [n]$ and outputs the logical "or" $\bigvee_{i \in \mathcal{I}} x_i$. The goal is then to design the measurements in such a way that all d sparse vectors become uniquely identifiable using as few number of measurements as possible.

A natural generalization of classical group testing (that we call *threshold testing*), introduced by Damaschke [15], considers the case where the measurement outcomes are determined by a *threshold predicate* instead of the logical or. Namely, this model is characterized by two integer parameters ℓ, u such that $0 \geq \ell < u$ (that are considered fixed constants), and each measurement outputs positive if the number of positives within the corresponding pool is at least u. On the other hand, if the number of positives is less than or equal to[1] ℓ, the test returns negative, and otherwise the outcome can be arbitrary (that is, either 0 or 1 in any arbitrary way). In this view, classical group testing corresponds to the special case where $\ell = 0$ and $u = 1$. In addition to being of theoretical interest, the threshold model is interesting for applications, in particular in biology, where the measurements have reduced or unpredictable sensitivity or may depend on various factors that must be simultaneously present in the sample to result in a positive outcome.

The difference $g := u - \ell - 1$ is known as the *gap* parameter. As shown by Damaschke [15], in threshold group testing identification of the set of positives is only possible when the number of positives is at least u. Moreover, regardless of the number

[1] The proceedings version of this paper [11] and also the author's Ph.D. thesis [10] use a slightly different notation where the test returns negative if the number of positives in the group is strictly less than ℓ. Accordingly in those versions the gap parameter is defined to be $u - \ell$ rather than $u - \ell - 1$. A revised notation is used in this version to make the exposition consistent with the original paper of Damaschke [15].

of measurements, in general the set of positives can be only approximately identified within up to g false positives and g false negatives (thus, unique identification can only be guaranteed when $\ell = u - 1$). Additionally, Damaschke constructed a scheme for identification of the positives in the threshold model. For the gap-free case where $g = 0$, the number of measurements in this scheme is $O(d \log n)$, which is nearly optimal (within constant factors). However, when $g > 0$, the number of measurements becomes $O(dn^b + d^u)$, for an arbitrary constant $b > 0$, if up to $g + (u - 1)/b$ misclassifications are allowed.

A drawback of the scheme presented by Damaschke is that the measurements are *adaptive*; i.e., the group chosen by each measurement can depend on the outcomes of the previous ones. For numerous applications (in particular, in molecular biology), adaptive measurements are infeasible and must be avoided. In a *non-adaptive* setting, all measurements must be specified before their outcomes are revealed. This makes it convenient to think of the measurements in a matrix form. Specifically, a non-adaptive *measurement matrix* is an $m \times n$ Boolean matrix whose ith row is the characteristic vector of the set of items participating in the ith pool, and the goal would be to design a suitable measurement matrix.

More recently, non-adaptive threshold testing has been considered by Chen and Fu [5]. They observe that a generalization of the standard notion of disjunct matrices, the latter being extensively used in the literature of classical group testing, is suitable for the threshold model. Throughout this work, we refer to this generalized notion as *strongly disjunct* matrices and to the standard notion as *classical* disjunct matrices. Using strongly disjunct matrices, they show that $O(ed^{u+1} \log(n/d))$ non-adaptive measurements suffices to identify the set of positives (within g false positives/negatives) even if up to e erroneous measurements are allowed in the model. This number of measurements almost matches (up to constant factors) the known lower bounds on the number of rows of strongly disjunct matrices. However, the dependence on the sparsity parameter is d^{u+1}, which might be prohibitive for an interesting range of parameters, when the thresholds are not too small (e.g., $\ell + 1 = u = 10$) and the sparsity parameter is rather large (e.g., $d = n^{1/10}$).

In this work, we consider the non-adaptive threshold model in a possibly noisy setting, where a number of measurement outcomes (specified by an *error parameter* $e \geq 0$) may be incorrect. Our first observation is that, a new variation of classical disjunct matrices (that is in general strictly weaker than strongly disjunct matrices) suffices for the purpose of threshold group testing. Using a randomness-efficient probabilistic construction (that requires $\mathsf{poly}(d, \log n)$ bits of randomness), we construct generalized disjunct matrices with $O(d^{g+2}(\log d)\log(n/d))$ rows. Thus, we bring the exponent of d in the asymptotic number of measurements from $u + 1$ (that is optimal for strongly disjunct matrices) down to $g + 2$, which is *independent* of the actual choice of the thresholds and only depends on the *gap* between them. We also show that this tradeoff is essentially optimal for our notion of disjunct matrices. In the gap-free case, we furthermore show that this tradeoff is in fact the best to hope for (up to a $\log d$ term) for any threshold testing design, and thus our notion of disjunct matrices is indeed optimal (Corollary 13). For the positive-gap case, we show that the dependence d^{g+2}, up to poly-logarithmic factors, is necessary for any threshold testing design, and thus our notion obtains the correct exponent (Corollary 28).

We proceed to define a new auxiliary object, namely the notion of *regular* matrices, that turns out to be the key combinatorial object in our explicit constructions. Intuitively, given a gap $g \geq 0$, a suitable regular matrix M_1 can be used to take any measurement matrix M_2 designed for the threshold model with lower threshold $\ell = 0$ and higher threshold $u = g + 1$ and "lift" it up to matrix that works for any arbitrary lower threshold $\ell' > 0$ and the same gap g. Therefore, for instance, in order to address the gap-free model, it would suffice to have a non-adaptive scheme for the classical group testing model with $\ell + 1 = u = 1$. This transformation is accomplished using a simple product that increases the height of the original matrix M_2 by a multiplicative factor equal to the height of the regular matrix M_1, while preserving the "low-threshold" distinguishing properties of the original matrix M_2.

Next, we introduce a framework for construction of regular matrices using *strong lossless condensers* that are fundamental objects in derandomization theory, and more generally, theoretical computer science. We show that, by using an optimal condenser, it is possible to construct regular matrices with only $O(d(\log d)\log n)$ rows. This almost matches the upper bound achieved by a probabilistic construction that we also present in this work. To this date, no explicit construction of such optimal lossless condensers is known (though probabilistic constructions are easy to obtain). However, using state of the art in explicit condensers [4, 27], we will obtain two explicit constructions of regular matrices with incomparable parameters. Namely, one with $O(d(\log d)\mathsf{quasipoly}(\log n))$ rows and another with $O(d^{1+\beta}\mathsf{poly}(\log n))$, where $\beta > 0$ is any arbitrary constant and the exponent of the term $\mathsf{poly}(\log n)$ depends on the choice of β. By combining regular matrices with strongly disjunct ones (designed for the lowered thresholds $\ell' = 0$ and $u' = g + 1$), we obtain our threshold testing schemes. The bounds obtained by our final schemes are summarized in Table 1. When the lower threshold ℓ is not too small, our explicit constructions (rows M8 and M9 of Table 1) significantly improve what was previously known to be achievable even using non-constructive proofs.

The rest of the paper is organized as follows. In Sect. 1.1 we introduce preliminary notions and fix some notation. In Sect. 2 we formalize the notion of threshold testing designs. Moreover, we review the notion of strongly disjunct matrices and introduces our weaker notion of threshold disjunct matrices (for the gap-free case $g = 0$), in addition to the notion of regular matrices and its properties. We will also prove lower bounds on the number of rows of such matrices. In Sect. 3 we obtain matching probabilistic upper bounds on the number of rows using the probabilistic method. Furthermore, we develop our construction of regular matrices from lossless condensers, and instantiate the parameters in Sect. 3.1. This in particular leads to our explicit threshold testing schemes. In Sect. 4 we extend all our results to the case with nonzero gap. In Sect. 5, we obtain explicit constructions of strongly disjunct matrices from error-correcting codes, by extending the classical technique initiated by Kautz and Singleton. Finally, in Sect. 6 we discuss the future directions.

1.1 Preliminaries

For a matrix M, we denote by $M[i, j]$ the entry of M at the ith row and the jth column. Similarly, we denote the ith entry of a vector v by $v(i)$. The *support* a vector

Table 1 Summary of the parameters achieved by various constructions of threshold disjunct matrices. The noise parameter $p \in [0, 1)$ is arbitrary, and thresholds $\ell, u = \ell + g + 1$ are fixed constants. "Exp" and "Rnd" respectively indicate explicit and randomized constructions. "KS" refers to the construction of strongly disjunct matrices based on Kautz-Singleton superimposed codes [29], as described later in Sect. 5 (the bounds in rows M1–M5 are obtained by strongly disjunct matrices)

	Number of rows	Tolerable errors	Remarks
M1	$O\big(d^{u+1}\frac{\log(n/d)}{(1-p)^2}\big)$	$\Omega\big(pd\frac{\log(n/d)}{(1-p)^2}\big)$	Rnd: Random strongly disjunct matrices.
M2	$O\big((\frac{d}{1-p})^{u+1}\log n\big)$	$\Omega\big(pd\frac{\log n}{1-p}\big)$	Exp: KS using codes on the GV bound.
M3	$O\big((\frac{d\log n}{1-p})^{u+1}\big)$	$\Omega\big(pd\frac{\log n}{1-p}\big)$	Exp: KS using Reed-Solomon codes.
M4	$O\big((\frac{d}{1-p})^{2u+1}\log n\big)$	$\Omega\big(pd\frac{\log n}{1-p}\big)$	Exp: KS using Algebraic Geometric codes.
M5	$O\big((\frac{d\sqrt{\log n}}{1-p})^{u+3/2}\big)$	$\Omega\big(p(\frac{d\sqrt{\log n}}{1-p})^{3/2}\big)$	Exp: KS using Hermitian codes ($d \gg \sqrt{\log n}$).
M6	$O\big(d^{g+2}\frac{(\log d)\log(n/d)}{(1-p)^2}\big)$	$\Omega\big(pd\frac{\log(n/d)}{(1-p)^2}\big)$	Rnd: Construction 2.
M7	$O\big(d^{g+3}\frac{(\log d)\log^2 n}{(1-p)^2}\big)$	$\Omega\big(pd^2\frac{\log^2 n}{(1-p)^2}\big)$	Constructions 4 and 1 combined, assuming optimal condensers and strongly disjunct matrices.
M8	$O\big(d^{g+3}\frac{(\log d)T_2\log n}{(1-p)^{g+2}}\big)$	$\Omega\big(pd^2\frac{T_2\log n}{1-p}\big)$	Exp: Constructions 4 and 1 combined using Theorem 16 and M2, where $T_2 = \exp(O(\log^3\log n)) = \mathsf{quasipoly}(\log n)$.
M9	$O\big(d^{g+3+\beta}\frac{T_3^{\ell}\log n}{(1-p)^{g+2}}\big)$	$\Omega\big(pd^{2-\beta}\frac{\log n}{1-p}\big)$	Exp: Constructions 4 and 1 combined using Theorem 17 and M2, where $\beta > 0$ is any arbitrary constant and $T_3 = ((\log n)(\log d))^{1+u/\beta} = \mathsf{poly}(\log n, \log d)$.
	$\Omega(d^{g+2}\log_d n + ed^{g+1})$	e	Lower bound (Theorem 25).

$x \in \{0, 1\}^n$, denoted by $\mathsf{supp}(x)$, is a subset of $[n] := \{1, \ldots, n\}$ such that $i \in \mathsf{supp}(x)$ if and only if $x(i) = 1$. The Hamming weight of x, denoted by $\mathsf{wgt}(x)$ is defined as $|\mathsf{supp}(x)|$. The Hamming distance between vectors $x, x' \in \{0, 1\}^n$ is denoted by $\mathsf{dist}(x, x')$.

For an $m \times n$ Boolean matrix M and $S \subseteq [n]$, we denote by $M|_S$ the $m \times |S|$ submatrix of M formed by restricting M to the columns picked by S. Moreover, for a vector $x \in \{0, 1\}^n$, we use $M[x]_{\ell,u}$ to denote the set of all possible outcomes of measuring x in the threshold model with lower and upper thresholds ℓ and u and using the measurement matrix M. Formally, for any $y \in M[x]_{\ell,u}$ we have $y(i) = 1$ if $|\mathsf{supp}(M(i)) \cap \mathsf{supp}(x)| \geq u$, and $y(i) = 0$ if $|\mathsf{supp}(M(i)) \cap \mathsf{supp}(x)| \leq \ell$, where here $M(i)$ indicates the ith row of M. In the gap-free case, the measurement outcome is uniquely defined (since there is no ambiguity in the measurement process), and thus the set $M[x]_{\ell,u}$ only contains a single element that we denote by $M[x]_u$.

The *min-entropy* of a distribution $\mathcal{X}$ with finite support Ω is given by

$$H_\infty(\mathcal{X}) := \min_{x \in \Omega}\{-\log \mathcal{X}(x)\},$$

where $\mathcal{X}(x)$ is the probability that $\mathcal{X}$ assigns to the outcome x and logarithm is taken to base 2. A *flat* distribution is one that is uniform on its support. For such a distribution $\mathcal{X}$, we have $H_\infty(\mathcal{X}) = \log(|\mathsf{supp}(\mathcal{X})|)$. The *statistical distance* between two distributions $\mathcal{X}$ and $\mathcal{Y}$ defined on the same finite space Ω is given by $\frac{1}{2}\sum_{s\in\Omega} |\mathcal{X}(s) - \mathcal{Y}(s)|$, which is half the ℓ_1 distance of the two distributions when regarded as vectors of probabilities over Ω. Two distributions $\mathcal{X}$ and $\mathcal{Y}$ are said to be ϵ-close if their statistical distance is at most ϵ. We will use the shorthand $\mathcal{U}_n$ for the uniform distribution on $\{0,1\}^n$, and $X \sim \mathcal{X}$ for a random variable X drawn from a distribution $\mathcal{X}$.

The main technical tool that we use in our explicit constructions is the notion of *lossless condensers*, defined below.

Definition 1 A function $f\colon \{0,1\}^{\tilde{n}} \times \{0,1\}^t \to \{0,1\}^{\tilde{\ell}}$ is a strong *lossless condenser* for entropy k and with *error* ϵ (in short, (k,ϵ)-condenser) if for every distribution $\mathcal{X}$ on $\{0,1\}^{\tilde{n}}$ with min-entropy at least k, random variable $X \sim \mathcal{X}$ and a *seed* $Y \sim \mathcal{U}_t$, the distribution of $(Y, f(X,Y))$ is ϵ-close to some distribution $(\mathcal{U}_t, \mathcal{Z})$ with min-entropy at least $t + k$. A condenser is *explicit* if it is polynomial-time computable.

We will use the following "almost-injectivity" property of lossless condensers in our proofs.

Proposition 2 *Let $\mathcal{X}$ be a flat distribution with min-entropy* $\log K$ *over a finite sample space Ω and $f\colon \Omega \to \Gamma$ be a mapping to a finite set Γ. If $f(\mathcal{X})$ is ϵ-close to having min-entropy* $\log K$, *then there is a set $T \subseteq \Gamma$ of size at least* $(1-4\epsilon)K$ *such that*

$$(\forall y \in T) \quad f(x) = y \wedge f(x') = y \quad \Rightarrow \quad x = x'.$$

Proof Suppose that $\mathcal{X}$ is uniformly supported on a set $S \subseteq \Omega$ of size K. For each $y \in \Gamma$, define $n_y := |\{x \in \Omega\colon f(x) = y\}|$. Denote by μ the distribution $f(\mathcal{X})$ over Γ and by μ' a distribution on Γ with min-entropy $\log K$ that is ϵ-close to μ, which is guaranteed to exist by the assumption. Define $T := \{y \in \Gamma\colon n_y = 1\}$, and similarly, $T' := \{y \in \Gamma\colon n_y \geq 2\}$. Observe that for each $y \in \Gamma$ we have $\mu(y) = n_i/K$, and also $\mathsf{supp}(\mu) = T \cup T'$. Thus,

$$|T| + \sum_{y\in T'} n_y = K. \tag{1}$$

The fact that μ and μ' are ϵ-close implies that

$$\sum_{y\in T'} |\mu(y) - \mu'(y)| \leq 2\epsilon \quad \Rightarrow \quad \sum_{y\in T'} (n_y - 1) \leq 2\epsilon K.$$

In particular, this means that $|T'| \leq 2\epsilon K$ (since by the choice of T', for each $y \in T'$ we have $n_y \geq 2$). Furthermore,

$$\sum_{y\in T'} (n_y - 1) \leq 2\epsilon K \quad \Rightarrow \quad \sum_{y\in T'} n_y \leq 2\epsilon K + |T'| \leq 4\epsilon K.$$

This combined with (1) gives

$$|T| = K - \sum_{y \in T'} n_y \geq (1 - 4\epsilon)K,$$

as desired. □

2 Variations of Disjunct Matrices

The combinatorial structure used by Chen and Fu in their non-adaptive scheme is the following generalization of the standard notion of disjunct matrices that we refer to as *strongly disjunct* matrices throughout this work.

Definition 3 A matrix (with at least $d + u$ columns) is said to be strongly $(d, e; u)$-disjunct if for every choice of $d + u$ columns $C_1, \ldots, C_u, C'_1, \ldots, C'_d$, all distinct, we have

$$\left| \bigcap_{i=1}^{u} \mathsf{supp}(C_i) \setminus \bigcup_{i=1}^{d} \mathsf{supp}(C'_i) \right| > e.$$

Observe that, strongly $(d, e; u)$-disjunct matrices are, in particular, strongly $(d', e'; u')$-disjunct for any $d' \leq d$, $e' \leq e$, and $u' \leq u$. Moreover, *classical* (d, e)-disjunct matrices that are extensively used in group testing literature (see [17, Chap. 7]) are equivalent to strongly $(d, e; 1)$-disjunct matrices.

To make the main ideas more transparent, until Sect. 4 we will focus on the gap-free case where $\ell = u - 1$. The extension to nonzero gaps is straightforward and will be discussed in Sect. 4. Moreover, often we will implicitly assume that the Hamming weight of the Boolean vector that is to be identified is at least u (since otherwise, we know from the work of Damaschke [15] that confusions cannot be avoided), and will take the thresholds ℓ, u to be fixed constants.

The notion of strongly disjunct matrices, in its general form, has been studied in the literature under different names and equivalent formulations, e.g., superimposed (u, d)-designs/codes, superimposed distance codes, and (u, d) cover-free families (see [6, 7, 20, 23, 30, 39, 40] and the references therein). An important motivation for the study of this notion is the following *hidden hypergraph-learning problem* (cf. [17, Chap. 12]), itself being motivated by the so-called *complex model* in computational biology [6]: Suppose that G is a u-hypergraph; that is, a hypergraph where each edge is a set of u vertices. on a vertex set V of size n, and denote by $\mathcal{V}(G)$ the set of vertices induced by the hyper-edge set of G; i.e., $v \in \mathcal{V}(G)$ if and only if G has a hyper-edge incident to v. Then assuming that $|\mathcal{V}(G)| \leq d$ for a *sparsity parameter* d, the aim is to learn G using as few (non-adaptive) queries of the following type as possible: Each query specifies a set $Q \subseteq V$, and its corresponding answer is a Boolean value which is 1 if and only if G has a hyperedge contained in Q. It is known that [6, 25], in the hypergraph-learning problem, any suitable grouping strategy defines a strongly disjunct matrix (whose rows are characteristic vectors of individual queries Q), and conversely, any strongly disjunct matrix can be used as the incidence matrix of the set of queries. The parameter e determines "noise tolerance" of the measurement scheme. Namely, a strongly $(d, e; u)$-disjunct matrix can uniquely distinguish between d-sparse hypergraphs even in presence of up to $\lfloor e/2 \rfloor$ erroneous query outcomes.

For gap-free threshold group testing, the successful strategy needed for distinguishing between d-sparse Boolean vectors can trivially be captured by the following definition.

Definition 4 Let $n \geq d \geq u > 0$ and $e \geq 0$ be integer parameters. A Boolean matrix M with n columns is said to be a $(d, e; u)$-*threshold design* if for every d-sparse $x, x' \in \{0,1\}^n$ of Hamming weight u or more such that $x \neq x'$, we have $\mathsf{dist}(M[x]_u, M[x']_u) > e$.

The key observation made by Chen and Fu [5] is that threshold group testing corresponds to the special case of the hypergraph learning problem where the hidden graph G is known to be a u-clique.[2] In this case, the unknown Boolean vector in the corresponding threshold testing problem would be the characteristic vector of $\mathcal{V}(G)$. It follows that strongly disjunct matrices are threshold designs as defined in Definition 4 Specifically,

Theorem 5 [5] *Let M be a Boolean matrix with n columns that is strongly $(d, e; u)$-disjunct. Then, M is a $(d, e; u)$-threshold design.*

Nonconstructively, a probabilistic argument akin to the standard argument for the case of classical disjunct matrices (see [17, Chap. 7]) can be used to show that strongly $(d, e; u)$-disjunct matrices exist with $m = O(d^{u+1}(\log(n/d))/(1-p)^2)$ rows and error tolerance $e = \Omega(pd \log(n/d)/(1-p)^2)$, for any noise parameter $p \in [0,1)$. On the negative side, however, several concrete lower bounds are known for the number of rows of such matrices [23, 39, 40]. In asymptotic terms, these results show that one must have $m = \Omega(d^{u+1} \log_d n + ed^u)$, and thus, the probabilistic upper bound is essentially optimal.

For the underlying strongly disjunct matrix, Chen and Fu [5] use a greedy construction [7] that achieves, for any $e \geq 0$, $O((e+1)d^{u+1} \log(n/d))$ rows, but may take exponential time in the size of the resulting matrix. Nevertheless, as observed by several researchers [6, 23, 25, 30], a classical explicit construction of combinatorial designs due to Kautz and Singleton [29] can be extended to construct strongly disjunct matrices. This concatenation-based construction transforms any error-correcting code having large distance into a disjunct matrix. While the original construction uses Reed-Solomon codes and achieves nice bounds, it is possible to use other families of codes. In particular, as recently shown by Porat and Rothschild [34], codes on the Gilbert-Varshamov bound (cf. [32]) result in nearly optimal disjunct matrices. Moreover, for a suitable range of parameters, they give a *deterministic* construction

[2] As standard in graph theory, a u-clique on the vertex set V is a u-hypergraph (V, E) such that, for some $V' \subseteq V$, E is the set of all subsets of V' of size u.

of such codes that runs in polynomial time in the size of the resulting disjunct matrix (albeit exponential in the dimension of the code)[3]. We will elaborate on details of this class of constructions in Sect. 5, and will additionally consider a family of algebraic-geometric codes and Hermitian codes which give incomparable bounds, as summarized in Table 1 (rows M2–M5).

2.1 Threshold Disjunct and Regular Matrices

Even though, as discussed above, the general notion of strongly $(d,e;u)$-disjunct matrices is sufficient for threshold group testing with upper threshold u, in this section we show that a new, weaker, notion of disjunct matrices defined below (which, as we show later, turns out to be *strictly* weaker when $u>1$), would also suffice. We also define an auxiliary notion of *regular* matrices.

Definition 6 A Boolean matrix M with n columns is called $(d,e;u)$-regular if for every subset of columns $S \subseteq [n]$ (called the *critical set*) and every $Z \subseteq [n]$ (called the *zero set*) such that $u \leq |S| \leq d$, $|Z| \leq |S|$, $S \cap Z = \emptyset$, there are more than e rows of M at which $M|_S$ has weight exactly u and (at the same rows) $M|_Z$ has weight zero. Any such row is said to *u-satisfy* S and Z. If, in addition, for every *distinguished column* $i \in S$, more than e rows of M both u-satisfy S and Z and have a 1 at the ith column, the matrix is called threshold $(d,e;u)$-disjunct (and the corresponding "good" rows are said to u-satisfy i, S, and Z).

To distinguish between the above variant of disjunct matrices and strongly disjunct matrices or classical disjunct matrices, we will refer to our variant as *threshold disjunct* matrices throughout the paper.

It is easy to verify that (assuming $2d \leq n$) the classical notion of $(2d-1,e)$-disjunct matrices is equivalent to strongly $(2d-1,e;1)$-disjunct and threshold $(d,e;1)$-disjunct. Moreover, any threshold $(d,e;u)$-disjunct matrix is $(d,e;u)$-regular, $(d-1,e;u-1)$-regular, and classical (d,e)-disjunct (but the reverse implications do not in general hold). Therefore, the known lower bound of $m = \Omega(d^2 \log_d n + ed)$ that applies for (d,e)-disjunct matrices holds for threshold $(d,e;u)$-disjunct matrices as well (see Theorem 10). Below we show that our notion of disjunct matrices suffices for threshold designs.

Lemma 7 *Let M be an $m \times n$ Boolean matrix that is threshold $(d,e;u)$-disjunct. Then for every distinct d-sparse vectors $x, x' \in \{0,1\}^n$ such that* $\mathsf{supp}(x) \not\subseteq \mathsf{supp}(x')$, $\mathsf{wgt}(x) \geq |\mathsf{supp}(x') \setminus \mathsf{supp}(x)|$ *and* $\mathsf{wgt}(x) \geq u$, *we have*

$$|\mathsf{supp}(M[x]_u) \setminus \mathsf{supp}(M[x']_u)| > e. \tag{2}$$

Moreover, M is a $(d,e;u)$-threshold design. Conversely, if M satisfies (2) for every choice of x and x' as above, it must be threshold $(\lfloor d/2 \rfloor, e; u)$-disjunct.

[3] In this regard, this construction of disjunct matrices can be considered *weakly explicit* in that, contrary to fully explicit constructions, it is not clear if each individual entry of the matrix can be computed in time $\mathsf{poly}(d, \log n)$.

Proof First, suppose that M is threshold $(d, e; u)$-disjunct, and let $y := M[x]_u$ and $y' := M[x']_u$. Take any $i \in \mathsf{supp}(x) \setminus \mathsf{supp}(x')$, and let $S := \mathsf{supp}(x)$ and $Z := \mathsf{supp}(x') \setminus \mathsf{supp}(x)$. Note that $|S| \leq d$ and by assumption, we have $|Z| \leq |S|$. Now, Definition 6 implies that there is a set E of more than e rows of M that u-satisfy i as the distinguished column, S as the critical set and Z as the zero set. Thus for every $j \in E$, the jth row of M restricted to the columns chosen by $\mathsf{supp}(x)$ must have weight exactly u, while its weight on $\mathsf{supp}(x')$ is less than u. Therefore, $y(j) = 1$ and $y'(j) = 0$ for more than e choices of j.

The claim that M is a $(d, e; u)$-threshold design follows from the above argument combined with the observation that at least one of the two possible orderings of any two distinct d-sparse vectors, at least one having weight u or more, satisfies the conditions required by the lemma.

For the converse, consider any choice of a distinguished column $i \in [n]$, a critical set $S \subseteq [n]$ containing i (such that $|S| \geq u$), and a zero set $Z \subseteq [n]$ where $|Z| \leq |S|$. Define d-sparse Boolean vectors $x, x' \in \{0, 1\}^n$ so that $\mathsf{supp}(x) := S$ and $\mathsf{supp}(x') := Z \cup (S \setminus \{i\})$. Let $y := M[x]_u$ and $y' := M[x']_u$ and $E := \mathsf{supp}(y) \setminus \mathsf{supp}(y')$. By assumption we know that $|E| > e$. Take any $j \in E$. Since $y(j) = 1$ and $y'(j) = 0$, we get that the jth row of M restricted to the columns picked by $Z \cup (S \setminus \{i\})$ must have weight at most $u - 1$, whereas it must have weight at least u when restricted to S. As the sets $\{i\}$, $S \setminus \{i\}$, and Z are disjoint, this can hold only if $M[j, i] = 1$, and moreover, the jth row of M restricted to the columns picked by S (resp., Z) has weight exactly u (resp., zero). Hence, this row (as well as all the rows of M picked by E) must u-satisfy i, S, and Z, confirming that M is threshold $(\lfloor d/2 \rfloor, e; u)$-disjunct. □

We point out that Lemma 7 proves a matching converse, suggesting that the notion of threshold disjunct matrices might be "close" to a characterization of threshold designs (Definition 4), up to a constant factor in the sparsity parameter. However, this does not imply a precise characterization since the assumptions of Lemma 7 consider a particular ordering on the sparse vectors x and x', which must be consistent with the ordering in (2). However, as we show in Sect. 2.3, threshold designs (Definition 4) and threshold disjunct matrices (Definition 6) satisfy the same asymptotic lower bounds on the number of rows, which nearly matches the upper bounds that we prove by probabilistic arguments (Lemma 14), assuming that the threshold parameter is an absolute constant. Thus, quantitatively, our notion of threshold disjunct matrices essentially provides an optimal way of constructing threshold group testing designs.

2.2 Direct Product of Matrices

We will use regular matrices as intermediate building blocks in our constructions of disjunct matrices to follow. The connection with disjunct matrices is made apparent through a direct product of matrices defined in Construction 1. Intuitively, using this product, regular matrices can be used to transform any measurement matrix suitable for the standard group testing model to one with comparable properties in the threshold model. The following lemma formalizes the idea.

- *Given:* Boolean matrices M_1 and M_2 that are $m_1 \times n$ and $m_2 \times n$, respectively.
- *Output:* An $m \times n$ Boolean matrix $M_1 \odot M_2$, where $m := m_1 m_2$.
- *Construction:* Let the rows of $M := M_1 \odot M_2$ be indexed by the set $[m_1] \times [m_2]$. Then the row corresponding to (i, j) is defined as the bit-wise or of the ith row of M_1 and the jth row of M_2.

Construction 1: Direct product of measurement matrices

Lemma 8 *Let M_1 and M_2 be Boolean matrices with n columns, such that M_1 is $(d-1, e_1; u-1)$-regular. Let $M := M_1 \odot M_2$, and suppose that for d-sparse Boolean vectors $x, x' \in \{0,1\}^n$ such that $\mathsf{wgt}(x) \geq \mathsf{wgt}(x')$, we have*

$$|\mathsf{supp}(M_2[x]_1) \setminus \mathsf{supp}(M_2[x']_1)| \geq e_2.$$

Then, $|\mathsf{supp}(M[x]_u) \setminus \mathsf{supp}(M[x']_u)| \geq (e_1 + 1)e_2$.

Proof First we consider the case where $u > 1$. Let $y := M_2[x]_1 \in \{0,1\}^{m_2}$, $y' := M_2[x']_1 \in \{0,1\}^{m_2}$, where m_2 is the number of rows of M_2, and let $E := \mathsf{supp}(y) \setminus \mathsf{supp}(y')$. By assumption, $|E| \geq e_2$. Fix any $i \in E$ so that $y(i) = 1$ and $y'(i) = 0$. Therefore, the ith row of M_2 must have all zeros at positions corresponding to $\mathsf{supp}(x')$ and there is a $j \in \mathsf{supp}(x) \setminus \mathsf{supp}(x')$ such that $M_2[i, j] = 1$. Define $S := \mathsf{supp}(x) \setminus \{j\}$, $Z := \mathsf{supp}(x') \setminus \mathsf{supp}(x)$, $z := M[x]_u$ and $z' := M[x']_u$.

As $\mathsf{wgt}(x) \geq \mathsf{wgt}(x')$, we know that $|Z| \leq |S| + 1$. The extreme case $|Z| = |S| + 1$ only happens when x and x' have disjoint supports, in which case one can remove an arbitrary element of Z to ensure that $|Z| \leq |S|$ and the following argument (considering the assumption $u > 1$) still goes through.

By the definition of regularity, there is a set E_1 consisting of at least $e_1 + 1$ rows of M_1 that $(u-1)$-satisfy the critical set S and the zero set Z. Pick any $k \in E_1$, and observe that z must have a 1 at position (k, i). This is because the row of M indexed by (k, i) has a 1 at the jth position (since the kth row of M_2 does), and at least $u - 1$ more 1's at positions corresponding to $\mathsf{supp}(x) \setminus \{j\}$ (due to regularity of M_1). On the other hand, note that the kth row of M_1 has at most $u - 1$ ones at positions corresponding to $\mathsf{supp}(x')$ (because $\mathsf{supp}(x') \subseteq S \cup Z$), and the ith row of M_2 has all zeros at those positions (because $y'(i) = 0$). This means that the row of M indexed by (k, i) (which is the bit-wise or of the kth row of M_1 and the ith row of M_2) must have less than u ones at positions corresponding to $\mathsf{supp}(x')$, and thus, z' must be 0 at position (k, i). Therefore, z and z' differ at position (k, i).

Since there are at least e_2 choices for i, and for each choice of i, at least $e_1 + 1$ choices for k, we conclude that in at least $(e_1 + 1)e_2$ positions, z has a one while z' has a zero.

The argument for $u = 1$ is similar, in which case it suffices to take $S := \mathsf{supp}(x)$ and $Z := \mathsf{supp}(x') \setminus \mathsf{supp}(x)$. □

As a corollary it follows that, when M_1 is a $(d-1, e_1; u-1)$-regular and M_2 is a classical (d, e_2)-disjunct matrix, the product $M := M_1 \odot M_2$ will distinguish between any two distinct d-sparse vectors (of weight at least u) in at least $(e_1 + 1)(e_2 + 1)$

positions of the measurement outcomes. This combined with Lemma 7 would imply that M is, in particular, threshold $(\lfloor d/2 \rfloor, (e_1+1)(e_2+1)-1; u)$-disjunct. However, using a direct argument similar to the above lemma it is possible to obtain a slightly better result, given by Lemma 9.

Lemma 9 *Suppose that M_1 is a $(d, e_1; u-1)$-regular and M_2 is a classical $(2d, e_2)$-disjunct matrix. Then $M_1 \odot M_2$ is a threshold $(d, (e_1+1)(e_2+1)-1; u)$-disjunct matrix.*

As a particular example of where Lemma 8 can be used, we remark that the measurement matrices constructed in [9] that are not necessarily disjunct but allow approximation of sparse vectors in highly noisy settings of the standard group testing model (as well as those used in adaptive two-stage schemes; cf. [8] and the references therein), can be combined with regular matrices to offer the same qualities in the threshold model. In the same way, numerous existing results in group testing can be ported to the threshold model by using Lemma 8.

2.3 Lower Bounds

In this section, we show that the known asymptotic lower bounds on the number of rows of classical disjunct matrices apply to threshold designs (Definition 4) and our notion of threshold disjunct matrices (6) as well. It is immediate from the definitions that, assuming $2d \leq n$, a threshold $(d, e; u)$-disjunct matrix is in particular a classical (d, e)-disjunct matrix. Thus the latter lower bound is straightforward.

Theorem 10 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ threshold $(d, e; u)$-disjunct matrix. Then,*

$$m = \Omega(d^2 \log_d n + de).$$

Proof Immediate from the known bounds on the number of rows of classical disjunct matrices (e.g., Theorem 2.19 of [39]). □

Now, in order to show that any $(d, e; u)$-threshold design must satisfy essentially the same lower bound as in Theorem 10, we first observe the following combinatorial property of such matrices.

Lemma 11 *Let M be a $(d+1, e; \ell+1)$-threshold design. Then it satisfies the following property:*

"For every $S \subseteq [n]$ such that $|S| = d$ and every $i \in [n] \setminus S$, there are more than e rows of M at which the ith column of M contains a 1 and moreover in those rows, $M|_S$ has weight exactly ℓ."

Proof This is a special case of Lemma 26 that will be proved later (it suffices to set $u = \ell + 1$ and $g = 1$ in Lemma 26). □

Theorem 12 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ Boolean matrix that satisfies the property quoted in Lemma 11. Then,*

$$m = \Omega\left(\left(\frac{d}{\ell+1}\right)^2 \log_d n + \frac{de}{(\ell+1)^2}\right).$$

Proof We reduce the matrix to a classical disjunct matrix, and use the existing lower bounds. Let $d' := \lfloor d/(\ell+1) \rfloor$ and $e' := e/(\ell+1)$. We define the following notation: For a set $S \subseteq [n]$ and $i \in [n] \setminus S$, a vector $v \in \{0,1\}^n$ is said to *satisfy* (i, S) if $v(i) = 1$ and $v(j) = 0$ for all $j \in S$.

For each $i \in [n]$, we create a set $T(i) \subseteq [n]$ according to the following greedy algorithm:

1. Initialize $T(i)$ with the empty set.
2. Let $S \subseteq [n] \setminus (T(i) \cup \{i\})$ be any set of size at most d' such that the number of rows of M that satisfy (i, S) is at most e'. If there is no such S, terminate.
3. Set $T(i) := T(i) \cup S$, and go to step 2.

First, we argue that the above algorithm always terminates after looping at most ℓ times. Suppose for the sake of contradiction that the algorithm loops more and let $S_1, \ldots, S_{\ell+1}$ be the disjoint sets S obtained in the first $\ell+1$ iterations of the loop. Let M' be the matrix obtained from M by removing all the rows where the ith column has a zero, and define $T' := S_1 \cup \cdots \cup S_{\ell+1}$.

By the way the algorithm chooses the sets S_j, we know for each S_j that all but at most e' rows of $M'|_{S_j}$ have nonzero weights. Therefore, all but at most $e'(\ell+1) = e$ rows of $M'|_T$ have weights at least $\ell+1$ (i.e., at least one nonzero entry for the range of each S_j).

On the other hand, since $|S_j| \leq d'$ for all j, we have $|T'| \leq d'(\ell+1) \leq d$. So, the property of Lemma 11 implies that there are more than e rows of M' where $M'|_{T'}$ has weight exactly ℓ. This is a contradiction. Therefore, we conclude, for every i, that $|T(i)| \leq \ell d' < d$.

Now, define an undirected graph $G = (V, E)$ where $V := [n]$ and $\{i, j\} \in E$ iff either $j \in T(i)$ or $i \in T(j)$. We know from the upper bound on the size of every $T(i)$ that the maximum degree of this graph is less than $2d$. Therefore, the graph has an independent set $V' \subseteq V$ of size at least $n/(2d)$. Let $M'' := M|_{V'}$, with columns indexed by the elements of V'.

Now, consider any $i \in V'$ and any set $S \subseteq V' \setminus i$ where $|S| = d'$. Since V' is an independent set of G, we know that $T(i) \cap V' = \emptyset$. Since the greedy algorithm, given input i, has terminated at step 2, we know that there are more than e' rows of M'' that satisfy (i, S) (otherwise the algorithm would add S to $T(i)$ and loop another time). Since this holds for every choice of (i, S), we conclude that the matrix M'' must be a classical (d', e')-disjunct matrix.

Let n' be the number of columns of M'', so we know that $n' \geq n/(2d)$. Now it suffices to apply the known asymptotic lower bounds for the number of rows of classical disjunct matrices [19, 37, 39] on M''. In particular, Theorem 2.19 of [39] implies that, for some absolute constant $c > 0$, and whenever n is sufficiently large for the given parameter d,

$$m \geq 0.7c \frac{(d'+1)^2}{\log(d'+1)} \log n' + 0.5c(d'+1)e'$$
$$= \Omega\left(\frac{d^2(\log n - \log d - 1)}{(\ell+1)^2 \log d} + \frac{de}{(\ell+1)^2}\right),$$

which implies the claimed bound assuming n is large enough. □

Corollary 13 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ Boolean matrix that is a $(d, e; u)$*-threshold design, *for some constant $u > 0$. Then,*

$$m = \Omega_u(d^2 \log_d n + de).$$

Proof Immediate from Lemma 11 and Theorem 12. □

3 Constructions

In this section, we obtain several construction of regular and disjunct matrices. Our first construction, described in Construction 2, is a randomness-efficient probabilistic construction that can be analyzed using standard techniques from the probabilistic method. The bounds obtained by this construction are given in Lemma 14 below. The amount of random bits required by this construction is polynomially bounded in d and $\log n$, which is significantly smaller than it would be had we picked the entries of M fully independently.

Lemma 14 *For every $p \in [0, 1)$ and integer parameter $u > 0$, Construction 2 with $m' = O_u(d \log(n/d)/(1-p)^2)$ (resp., $m' = O_u(d^2 \log(n/d)/(1-p)^2)$) outputs a $(d, \Omega_u(pm'); u)$-regular (resp., threshold $(d, \Omega_u(pm'/d); u)$-disjunct) matrix with probability $1 - o(1)$.*

Proof We show the claim for regular matrices, the proof for disjunct matrices is similar. Consider any particular choice of a critical set $S \subseteq [n]$ and a zero set $Z \subseteq [n]$ such that $u \leq |S| \leq d$ and $|Z| \leq |S|$. Choose an integer i so that $2^{i-1}u \leq |S| \leq 2^i u$, and take any $j \in [m']$. Denote the (i, j)th row of M by the random variable $\boldsymbol{w} \in \{0, 1\}^n$, and by q the "success" probability that $\boldsymbol{w}|_S$ has weight exactly u and $\boldsymbol{w}|_Z$ is all zeros.

- *Given:* Integer parameters n, m', d, u.
- *Output:* An $m \times n$ Boolean matrix M, where $m := m'\lceil\log(d/u)\rceil$.
- *Construction:* Let $r := \lceil\log(d/u)\rceil$. Index the rows of M by $[r] \times [m']$. Sample the (i, j)th row of M independently from a $(u+1)$-wise independent distribution on n bit vectors, where each individual bit has probability $1/(2^{i+2}u)$ of being 1.

Construction 2: Probabilistic construction of regular and disjunct matrices

For an integer $r > 0$, we will use the shorthand 1^r (resp., 0^r) for the all-ones (resp., all-zeros) vector of length r. We have

$$\begin{aligned} q &= \sum_{\substack{R\subseteq[S]\\ |R|=u}} \Pr[(\boldsymbol{w}|_R) = 1^u \wedge (\boldsymbol{w}|_{Z\cup(S\setminus R)}) = 0^{|S|+|Z|-u}] \\ &= \sum_R \Pr[(\boldsymbol{w}|_R) = 1^u] \cdot \Pr[(\boldsymbol{w}|_{Z\cup(S\setminus R)}) = 0^{|S|+|Z|-u} \mid (\boldsymbol{w}|_R) = 1^u] \\ &\overset{\text{(a)}}{=} \sum_R (1/(2^{i+2}u))^u \cdot (1 - \Pr[(\boldsymbol{w}|_{Z\cup(S\setminus R)}) \neq 0^{|S|+|Z|-u} \mid (\boldsymbol{w}|_R) = 1^u]) \\ &\overset{\text{(b)}}{\geq} \sum_R (1/(2^{i+2}u))^u \cdot (1 - (|S| + |Z| - u)/(2^{i+2}u)) \\ &\geq \frac{1}{2}\binom{|S|}{u}(1/(2^{i+2}u))^u \geq \frac{1}{2}\left(\frac{|S|}{u}\right)^u \cdot (1/(2^{i+2}u))^u \geq \frac{1}{2^{3u+1}\cdot u^u} =: c, \end{aligned} \tag{3}$$

where (a) and (b) use the fact that the entries of $\boldsymbol{w}$ are $(u+1)$-wise independent, and (b) uses an additional union bound. Here the lower bound $c > 0$ is a constant that only depends on u. Now, let $e := m'pq$. using Chernoff bounds, and independence of the rows, the probability that there are at most e rows (among $(i,1),\ldots,(i,m')$) whose restrictions to S and Z have weights u and 0, respectively, becomes upper bounded by

$$\exp(-(m'q-e)^2/(2m'q)) = \exp(-(1-p)^2 m'q/2) \leq \exp(-(1-p)^2 m'c/2).$$

Now take a union bound on all the choices of S and Z to conclude that the probability that the resulting matrix is not $(d,e;u)$-regular is at most

$$\left(\sum_{s=u}^{d}\binom{n}{s}\sum_{z=0}^{s}\binom{n-s}{z}\right)\exp(-(1-p)^2 m'c/2),$$

which can be made $o(1)$ by choosing $m' = O_u(d\log(n/d)/(1-p)^2)$.

The proof of the claim for disjunct matrices follows along the same lines, except that we additionally need the vector $\boldsymbol{w}$ to be 1 at the position corresponding to the distinguished column i. Under this additional requirement, the lower bound on q would become $\Omega_u(1/d)$, and this only increases the number of rows by a factor $O_u(d)$. □

A significant part of this work is a construction of regular matrices using strong lossless condensers. Details of the construction are described in Construction 4 that assumes a family of lossless condensers with different entropy requirements,[4] and in turn, uses Construction 3 as a building block. The theorem below analyzes the obtained parameters without specifying any particular choice for the underlying family of condensers.

[4] We have assumed that all the functions in the family have the same seed length t. If this is not the case, one can trivially set t to be the largest seed length in the family.

- *Given:* A strong lossless (k, ϵ)-condenser $f\colon \{0,1\}^{\tilde{n}} \times \{0,1\}^t \to \{0,1\}^{\tilde{\ell}}$, integer parameter $u \geq 1$ and real parameter $p \in [0,1)$ such that $\epsilon < (1-p)/32$.
- *Output:* An $m \times n$ Boolean matrix M, where $n := 2^{\tilde{n}}$ and $m = 2^{t+k} O_u(2^{u(\tilde{\ell}-k)})$.
- *Construction:* Let $G_1 = (\{0,1\}^{\tilde{\ell}}, \{0,1\}^k, E_1)$ be any bipartite bi-regular graph with left vertex set $\{0,1\}^{\tilde{\ell}}$, right vertex set $\{0,1\}^k$, edge set E_1, left degree $d_\ell := 8u$, and right degree $d_r := 8u2^{\tilde{\ell}-k}$. Replace each right vertex v of G_1 with $\binom{d_r}{u}$ vertices, one for each subset of size u of the vertices on the neighborhood of v, and connect them to the vertices in the corresponding subsets. Denote the resulting graph by $G_2 = (\{0,1\}^{\tilde{\ell}}, V_2, E_2)$, where $|V_2| = 2^k \binom{d_r}{u}$ and E_2 is the edge set of the graph. Define the bipartite graph $G_3 = (\{0,1\}^n, V_3, E_3)$, where $V_3 := \{0,1\}^t \times V_2$ is the set of right vertices, as follows: Each left vertex $x \in \{0,1\}^n$ is connected to $(y, \Gamma_2(f(x,y)))$, for each $y \in \{0,1\}^t$, where $\Gamma_2(\cdot)$ denotes the neighborhood function of G_2 (i.e., $\Gamma_2(v)$ denotes the set of vertices adjacent to v in G_2). The output matrix M is the bipartite adjacency matrix of G_3 with columns indexed by the left vertices of row indexed by the right vertices of the graph.

Construction 3: A building block for construction of regular matrices

- *Given:* Integer parameters $d \geq u \geq 1$, real parameter $p \in [0,1)$, and a family $f_0, \dots, f_r$ of strong lossless condensers, where $r := \lceil \log(d/u') \rceil$ and u' is the smallest power of two such that $u' \geq u$. Each $f_i\colon \{0,1\}^{\tilde{n}} \times \{0,1\}^t \to \{0,1\}^{\tilde{\ell}(i)}$ is assumed to be a strong lossless $(k(i), \epsilon)$-condenser, where $k(i) := \log u' + i + 1$ and $\epsilon < (1-p)/32$.
- *Output:* An $m \times n$ Boolean matrix M, where $n := 2^{\tilde{n}}$ and $m = 2^t d \sum_{i=0}^{r} O_u(2^{u(\tilde{\ell}(i)-k(i))})$.
- *Construction:* For each $i \in \{0, \dots, r\}$, denote by M_i the output matrix of Construction 3 when instantiated with f_i as the underlying condenser, and by m_i its number of rows. Define $r_i := 2^{r-i}$ and let M_i' denote the matrix obtained from M_i by repeating each row r_i times. Construct the output matrix M by stacking $M_0', \dots, M_r'$ on top of one another.

Construction 4: Regular matrices from strong lossless condensers

Theorem 15 *The $m \times n$ matrix M output by Construction 4 is $(d, p\gamma 2^t; u)$-regular, where $\gamma = \max\{1, \Omega_u(d \cdot \min\{2^{k(i)-\tilde{\ell}(i)} \colon i = 0, \dots, r\})\}$.*

Proof As a first step, we verify the upper bound on the number of measurements m. Each matrix M_i has $m_i = 2^{t+k(i)} O_u(2^{u(\tilde{\ell}(i)-k(i))})$ rows, and M_i' has $m_i r_i$ rows, where $r_i = 2^{r-i}$. Therefore, the number of rows of M is

$$\sum_{i=0}^{r} r_i m_i = \sum_{i=0}^{r} 2^{t+\log u' + r + 1} m_i = 2^t d \sum_{i=0}^{r} O_u(2^{u(\tilde{\ell}(i)-k(i))}).$$

Let $S, Z \subseteq \{0,1\}^{\tilde{n}}$ respectively denote any choice of a critical set and zero set of size at most d, where $|Z| \le |S|$, and choose an integer $i \ge 0$ so that $2^{i-1}u' \le |S| \le 2^i u'$. Arbitrarily grow the two sets S and Z to possibly larger, and disjoint, sets $S' \supseteq S$ and $Z' \supseteq Z$ such that $|S'| = |Z'| = 2^i u'$ (for simplicity we have assumed that $d \le n/2$). Our goal is to show that there are "many" rows of the matrix M_i (in Construction 4) that u-satisfy S and Z.

Let $k := k(i) = \log u' + i + 1$, $\tilde{\ell} := \tilde{\ell}(i)$, and denote by G_1, G_2, G_3 the bipartite graphs used by the instantiation of Construction 3 that outputs M_i. Thus we need to show that "many" right vertices of G_3 are each connected to exactly u of the vertices in S and none of those in Z.

Consider the uniform distribution $\mathcal{X}$ on the set $S' \cup Z'$, which has min-entropy $\log u' + i + 1$. By an averaging argument, since the condenser f_i is strong, for more than a p fraction of the choices of the seed $y \in \{0,1\}^t$ (call them *good seeds*), the distribution $\mathcal{Z}_y := f_i(\mathcal{X}, y)$ is $\epsilon/(1-p)$-close (in particular, $(1/32)$-close) to a distribution with min-entropy $\log u' + i + 1$.

Fix any good seed $y \in \{0,1\}^t$. Let $G = (\{0,1\}^{\tilde{n}}, \{0,1\}^{\tilde{\ell}}, E)$ denote a bipartite graph representation of f_i, where each left vertex $x \in \{0,1\}^{\tilde{n}}$ is connected to $f_i(x,y)$ on the right. Denote by $\Gamma_y(S' \cup Z')$ the right vertices of G corresponding to the neighborhood of the set of left vertices picked by $S' \cup Z'$. Note that $\Gamma_y(S' \cup Z') = \mathsf{supp}(\mathcal{Z}_y)$. Using Proposition 2, we see that since $\mathcal{Z}_y$ is $(1/32)$-close to having min-entropy $\log(|S' \cup Z'|)$, there are at least $(7/8)|S' \cup Z'|$ vertices in $\Gamma(S' \cup Z')$ that are each connected to exactly one left vertex in $S' \cup Z'$. Since $|S| \ge |S' \cup Z'|/4$, this implies that at least $|S' \cup Z'|/8$ vertices in $\Gamma(S' \cup Z')$ (call them Γ'_y) are connected to exactly one left vertex in S and no other vertex in $S' \cup Z'$. In particular we get that $|\Gamma'_y| \ge 2^{k-3}$.

Now, in G_1, let T_y be the set of left vertices corresponding to Γ'_y (regarding the left vertices of G_1 in one-to-one correspondence with the right vertices of G). The number of edges going out of T_y in G_1 is $d_\ell |T_y| \ge u2^k$. Therefore, as the number of the right vertices of G_1 is 2^k, there must be at least one right vertex that is connected to at least u vertices in T_y. Moreover, a counting argument shows that the number of right vertices connected to u or more vertices in T_y is at least $2^{k-\tilde{\ell}}2^k/(10u)$.

Observe that in construction of G_2 from G_1, any right vertex of G_1 is replicated $\binom{d_r}{u}$ times, one for each u-subset of its neighbors. Therefore, for a right vertex of G_1 that is connected to *at least* u left vertices in T_y, one or more of its copies in G_2 must be connected to *exactly* u vertex in T_y (among the left vertices of G_2) and no other vertex (since the right degree of G_2 is equal to u).

Define $\gamma' := \max\{1, 2^{k-\tilde{\ell}}2^k/(10u)\}$. From the previous argument we know that, looking at T_y as a set of left vertices of G_2, there are at least γ' right vertices on the neighborhood of T_y in G_2 that are connected to exactly u of the vertices in T_y and none of the left vertices outside T_y. Letting v_y be any such vertex, this implies that the vertex $(y, v_y) \in V_3$ on the right part of G_3 is connected to exactly u of the vertices in S, and none of the vertices in Z. Since the argument holds for every good seed y, the number of such vertices is at least the number of good seeds, which is more than $p\gamma' 2^t$. Since the rows of the matrix m_i are repeated $r_i = 2^{r-i}$ times in M, we conclude that M has at least $p\gamma' 2^{t+r-i} \ge p\gamma 2^t$ rows that u-satisfy S and Z, and the claim follows. □

3.1 Instantiations

We now instantiate the result obtained in Theorem 15 by various choices of the family of lossless condensers. The crucial factors that influence the number of measurements are the seed length and the output length of the condenser.

Non-constructively, it can be shown that strong (k,ϵ) lossless condensers with input length $\tilde{n}$, seed length $t=\log\tilde{n}+\log(1/\epsilon)+O(1)$, and output length $\tilde{\ell}=k+\log(1/\epsilon)+O(1)$ exist, and moreover, almost matching lower bounds are known [4]. In fact, the optimal parameters can be achieved by a random function with overwhelming probability. In this work, we consider two important explicit constructions of lossless condensers. Namely, one based on "zig-zag products" due to Capalbo et al. [4] and another, coding theoretic, construction due to Guruswami et al. [27].

Theorem 16 [4] *For every $k\le\tilde{n}\in\mathbb{N}$, $\epsilon>0$ there is an explicit lossless (k,ϵ) condenser with seed length $O(\log^3(\tilde{n}/\epsilon))$ and output length $k+\log(1/\epsilon)+O(1)$.*

Theorem 17 [27] *For all constants $\alpha\in(0,1)$ and every $k\le\tilde{n}\in\mathbb{N}$, $\epsilon>0$ there is an explicit strong lossless (k,ϵ) condenser with seed length $t=(1+1/\alpha)\log(\tilde{n}k/\epsilon)+O(1)$ and output length $\tilde{\ell}=t+(1+\alpha)k$.*

As a result, we use Theorem 15 with the above condensers to obtain the following.

Theorem 18 *Let $u>0$ be fixed, and $p\in[0,1)$ be a real parameter. Then for integer parameters $d,n\in\mathbb{N}$ where $u\le d\le n$,*

1. *Using an optimal lossless condenser in Construction 4 results in an $m_1\times n$ matrix M_1 that is $(d,e_1;u)$-regular, where $m_1=O(d(\log n)(\log d)/(1-p)^{u+1})$ and $e_1=\Omega(pd\log n)$.*
2. *Using the lossless condenser of Theorem 16 in Construction 4 results in an $m_2\times n$ matrix M_2 that is $(d,e_2;u)$-regular, where $m_2=O(T_2d(\log d)/(1-p)^u)$ for some $T_2=\exp(O(\log^3((\log n)/(1-p))))=\mathsf{quasipoly}(\log n)$, and $e_2=\Omega(pdT_2(1-p))$.*
3. *Let $\beta>0$ be any fixed constant. Then Construction 4 can be instantiated using the lossless condenser of Theorem 17 so that we obtain an $m_3\times n$ matrix M_3 that is $(d,e_3;u)$-regular, where $m_3=O(T_3^{1+u}d^{1+\beta}(\log d))$ for $T_3:=((\log n)(\log d)/(1-p))^{1+u/\beta}=\mathsf{poly}(\log n,\log d)$, and $e_3=\Omega(p\max\{T_3,d^{1-\beta/u}\})$.*

Proof First we show the claim for M_1. In this case, we take each f_i in Construction 4 to be an optimal lossless condenser satisfying the (non-constructive) bounds obtained in[5] [4]. Thus we have that $2^t=O(\tilde{n}/\epsilon)=O(\log n/\epsilon)$, and for every $i=0,\dots,r$, we have $2^{\tilde{\ell}(i)-k(i)}=O(1/\epsilon)$, where $\epsilon=O(1-p)$. Now we apply Theorem 15 to obtain the desired bounds (and in particular, $\gamma=\Omega(\epsilon d)$).

[5] This result is similar in spirit to the probabilistic argument used in [35] for showing the existence of good extractors.

Similarly, for the construction of M_2 we set up each f_i with the explicit construction of condensers in Theorem 16 for min-entropy $k(i)$. In this case, the maximum required seed length is $t = O(\log^3(\tilde{n}/\epsilon))$, and we let $T_2 := 2^t = \exp(O(\log^3((\log n)/(1-p))))$. Moreover, for every $i = 0, \dots, r$, we have $2^{\tilde{\ell}(i)-k(i)} = O(1/\epsilon)$. Plugging these parameters in Theorem 15 gives $\gamma = \Omega(\epsilon d)$ and the bounds on m_2 and e_2 follow.

Finally, for M_3 we use Theorem 17 with $\alpha := \beta/u$. Thus the maximum seed length becomes $t = (1 + u/\beta)\log(\tilde{n}(\log d)/(1-p)) + O(1)$, and for every $i = 0, \dots, r$, we have $\tilde{\ell}(i) - k(i) = O(t + \beta(\log d)/u)$. Clearly, $T_3 = \Theta(2^t)$, and thus (using Theorem 15) the number of measurements becomes $m_3 = T^{1+u}d^{1+\beta}(\log d)$. Moreover, we get $\gamma = \max\{1, \Omega(d^{1-\beta/u}/T)\}$, which gives $e_3 = \Omega(pT\gamma) = p\max\{T, d^{1-\beta/u}\}$, as claimed. □

By combining this result with Lemma 9 using any explicit construction of classical disjunct matrices, we obtain threshold $(d, e; u)$-disjunct matrices that can be used in the threshold model with any fixed threshold, sparsity d, and error tolerance $\lfloor e/2 \rfloor$. In particular, using the coding-theoretic explicit construction of nearly optimal classical disjunct matrices from codes on the Gilbert-Varshamov bound [34] (Theorem 30 in the appendix), we obtain threshold $(d, e; u)$-disjunct matrices with $m = O(m'd^2(\log n)/(1-p)^2)$ rows and error tolerance $e = \Omega(e'pd(\log n)/(1-p))$, where m' and e' are respectively the number of rows and error tolerance of any of the regular matrices obtained in Theorem 18. We note that in all cases, the final dependence on the sparsity parameter d is, roughly, $O(d^3)$ which has an exponent independent of the threshold u. Rows M7–M9 of Table 1 summarize the obtained parameters for the general case (with arbitrary gaps). We see that, when d is not negligibly small (e.g., $d = n^{1/10}$), the bounds obtained by our explicit constructions are significantly better than those offered by strongly disjunct matrices.

4 The Case with Positive Gaps

In preceding sections we have focused on the case where $g = 0$. However, in this section we observe that all the techniques that we have developed in this work can be extended to the positive-gap case in a straightforward way. The main observations are listed below. Recall from [15] that in the positive-gap case, we can only hope to distinguish between distinct d-sparse vectors x and x' where at least one has support size u or more and either $|\mathsf{supp}(x) \setminus \mathsf{supp}(x')| > g$ or $|\mathsf{supp}(x') \setminus \mathsf{supp}(x)| > g$. We will call any pair of such vectors *distinguishable*. Moreover, we naturally extend the Definition 4 of threshold designs to the positive-gap case as follows.

Definition 19 (Definition 4, generalized) Let $n \geq d \geq u > 0$ and $g \in [0, u)$, and $e \geq 0$ be integer parameters, and define $\ell := u - g - 1$. A Boolean matrix M with n columns is said to be a $(d, e; u, g)$-*threshold design* if for every d-sparse $x, x' \in \{0,1\}^n$ of Hamming weight u or more such that $|\mathsf{supp}(x) \setminus \mathsf{supp}(x')| > g$, every $y \in M[x]_{\ell,u}$ and every $y' \in M[x']_{\ell,u}$, we have $\mathsf{dist}(y, y') > e$.

4.1 Generalized Threshold Disjunct Matrices

For the positive-gap case, Definition 6 of threshold disjunct matrices can be adapted to allow more than one distinguished column in disjunct matrices. In particular, in general we may require the matrix M to have more than e rows that u-satisfy every choice of a critical set S, a zero set Z, and any set of $g+1$ designated columns $I \subseteq S$ (at which all entries of the corresponding rows must be 1). Denote this generalized notion by threshold $(d, e; u, g)$-disjunct matrices. It is straightforward to extend the arguments of Lemma 7 to show that the generalized notion of threshold $(d, e; u, g)$-disjunct matrices suffices to capture non-adaptive threshold group testing with upper threshold u and gap g. More precisely, the generalized definitions of threshold disjunct and regular matrices are as follows.

Definition 20 (Definition 6, generalized) Let n, d, e, u, g be non-negative integers where $g < u \leq d \leq n$. A Boolean matrix M with n columns is called threshold $(d, e; u, g)$-disjunct if for every subset of columns $S \subseteq [n]$ (called the *critical set*), every $Z \subseteq [n]$ (called the *zero set*) such that $u \leq |S| \leq d$, $|Z| \leq |S|$, $S \cap Z = \emptyset$, and every set $I \subseteq S$ of $g+1$ distinguished columns ($|I| = g+1$), there are more than e rows of M that u-satisfy S and Z and moreover, $M|_I$ has all ones at those columns. Moreover, M is called $(d, e; u, g)$-regular if for every choice of the critical and zero sets $S, Z \subseteq [n]$ with $|Z| \leq |S| + g$, there is a set of more than e rows of M that $(u-g)$-satisfy S and Z.

Note the slight difference between the notion of regular matrices above compared to Definition 6, namely, that the zero set Z can now be slightly larger than the critical set S (by at most u), and that the matrix is now required to $(u-g)$-satisfy (as opposed to u-satisfy) every choice of S and Z. The two notions coincide for $g = 0$. In general, the difference between the two notions of regular matrices is negligible as long as the parameter g remains small. In particular, it is straightforward to verify that all our results about the construction of regular matrices in the gap-free case (Constructions 2 and 4) as well as the obtained bounds (Lemma 14, Theorems 15 and 18) hold for the generalized notion of regular matrices with only a slight effect on the hidden terms that only depend on the threshold parameter u. We will see, however, that the generalized notion of threshold disjunct matrices is stronger than Definition 6 and the extra requirements may substantially affect the bounds (but not the construction techniques).

Below we show that the generalized notion of threshold disjunct matrices suffices for construction of threshold designs for the positive-gap case.

Lemma 21 (Lemma 7, generalized) *Let M be an $m \times n$ Boolean matrix that is threshold $(d, e; u, g)$-disjunct, and define $\ell := u - g - 1$. Then for every distinguishable d-sparse vectors $x, x' \in \{0,1\}^n$, each having support size u or more and such that $|\mathsf{supp}(x) \setminus \mathsf{supp}(x')| > g$ and $\mathsf{wgt}(x) \geq |\mathsf{supp}(x') \setminus \mathsf{supp}(x)|$, the following holds. Let $y \in M[x]_{\ell,u}$ and $y' \in M[x']_{\ell,u}$. Then,*

$$|\mathsf{supp}(y) \setminus \mathsf{supp}(y')| > e. \tag{4}$$

Moreover, M is a $(d, e; u, g)$-threshold design. Conversely, if M satisfies (4) for every choice of x, x', y, y' as above, it must be threshold $(\lfloor d/2 \rfloor, e; u, g)$-disjunct (assuming $n > d + g$).

Proof First, suppose that M is threshold $(d, e; u, g)$-disjunct, and let $y \in M[x]_{\ell,u}$ and $y' \in M[x']_{\ell,u}$ be arbitrarily chosen. Take any $I \subseteq \mathsf{supp}(x) \setminus \mathsf{supp}(x')$ of size $g + 1$, and let $S := \mathsf{supp}(x)$ and $Z := \mathsf{supp}(x') \setminus \mathsf{supp}(x)$. Note that $|S| \leq d$ and by assumption, we have $|Z| \leq |S|$. Now, Definition 20 implies that there is a set E of more than e rows of M that u-satisfy I as the set of distinguished columns, S as the critical set and Z as the zero set. Thus for every $j \in E$, the jth row of M restricted to the columns chosen by $\mathsf{supp}(x)$ must have weight exactly u, while its weight on $\mathsf{supp}(x')$ is at most $u - g - 1 = \ell$. Therefore, $y(j) = 1$ and $y'(j) = 0$ for more than e choices of j.

The claim that M is a $(d, e; u, g)$-threshold design follows from the above argument combined with the observation that, given any two d-sparse distinguishable vectors, having Hamming weight u or more, at least one of their two possible orderings satisfies the conditions required by the lemma.

For the converse, consider any choice of a set of distinguished columns $I \subseteq [n]$ with $|I| = g + 1$, a critical set $S \subseteq [n]$ containing I (such that $|S| \geq u$), and a zero set $Z \subseteq [n]$ where $|Z| \leq |S|$. Define d-sparse Boolean vectors $x, x' \in \{0, 1\}^n$ so that $\mathsf{supp}(x) := S$ and $\mathsf{supp}(x') := Z \cup (S \setminus I)$. We note that $\mathsf{wgt}(x) = |\mathsf{supp}(x)| \geq u$ and also, without loss of generality, $\mathsf{wgt}(x') \geq u$ (if the latter is not the case, we can simply enlarge Z by arbitrarily adding up to $g + 1$ elements outside the support of S to it and observe that is suffices to show the claim for the larger Z).

Let $y := M[x]_{\ell+1}$ and $y' := M[x']_u$, and observe that $y, y' \in M[x]_{\ell,u}$. Moreover, let $E := \mathsf{supp}(y) \setminus \mathsf{supp}(y')$. By assumption we know that $|E| > e$. Take any $j \in E$. Since $y(j) = 1$ and $y'(j) = 0$, we get that the jth row of M restricted to the columns picked by $Z \cup (S \setminus I)$ must have weight at most $\ell = u - (g + 1)$, whereas it must have weight at least u when restricted to S. As the sets I, $S \setminus I$, and Z are disjoint and $|I| = g + 1$, this can hold only if the jth row of M restricted to the columns picked by S, Z, and I has weights exactly u, 0, and $g + 1$, respectively. Hence, this row (as well as all the rows of M picked by E) must u-satisfy I, S, and Z, confirming that M is threshold $(\lfloor d/2 \rfloor, e; u, g)$-disjunct. □

4.2 Strongly Disjunct Versus Threshold Disjunct Matrices

The following proposition directly follows from the definitions, and relates strongly disjunct matrices to generalized threshold disjunct matrices.

Proposition 22 *Let n, d, e, u, g be non-negative integers where $g < u \leq d \leq n - (d + g + 1)$. Suppose that M and M' are binary $m \times n$ matrices, where M is threshold $(d, e; u, g)$-disjunct and M' is strongly $(2d, e; u)$-disjunct. Then, M is strongly $(d, e; g + 1)$-disjunct and M' is threshold $(d, e; u, g)$-disjunct.*

Proof First, we verify the conditions of Definition 3 for M. Consider any pair of disjoint sets $I, Z \subseteq [n]$ where $|I| = g + 1$ and $|Z| \leq d$. Let $S \subseteq [n]$ be any set of size

d containing I and disjoint from Z. Note that $|Z| \leq |S|$. From Definition 20 (with the critical set S, zero set Z, and distinguished set I), there is a set of more than e rows of M at which $M|_Z$ is all zeros and $M|_I$ is all ones. In other words, denoting the ith column of M by C_i, we have that

$$\left|\bigcap_{i \in I} \mathsf{supp}(C_i) \setminus \bigcup_{i \in Z} \mathsf{supp}(C_i)\right| > e,$$

as required by Definition 3.

Now consider the matrix M' and any choice of a I, S, Z as in Definition 20. Let $J \subseteq S$ be any subset of S of size u that contains I, and $S' := Z \cup (S \setminus J)$. Note that $|S'| \leq |S| + |Z| \leq 2d$. Now from Definition 3 of strongly disjunct matrices, we know that

$$\left|\bigcap_{i \in J} \mathsf{supp}(C_i) \setminus \bigcup_{i \in S'} \mathsf{supp}(C_i)\right| > e.$$

In other words, there is a set of more than e rows of M' at which $M'|_I$ is all ones, $M'|_S$ has weight exactly u, and $M'|_Z$ is all zeros, as required by Definition 20. □

The special case $u = g + 1$ in the above proposition is particularly interesting. A chain of reductions between strongly disjunct and threshold disjunct matrices in this case implied by the above result is schematically shown below.

$$\begin{array}{c}\text{threshold-}(2d, e; g+1, g)\text{-disjunct}\\ \downarrow \\ \text{strongly } (2d, e; g+1)\text{-disjunct}\\ \downarrow \\ \text{threshold-}(d, e; g+1, g)\text{-disjunct}\\ \downarrow \\ \text{strongly } (d, e; g+1)\text{-disjunct}\end{array}$$

Therefore, when the upper threshold u is more than the gap parameter g by one (equivalently, when the lower threshold ℓ is zero), the two notions of threshold disjunct matrices and strongly disjunct matrices become equivalent up to a multiplicative factor in the sparsity parameter d. As discussed in Sect. 2, almost matching lower bounds and upper bounds on the number of rows m achievable by a strongly $(d, e; g+1)$-disjunct matrix are known. Asymptotically, the number of rows must always satisfy $m = \Omega(d^{g+2} \log_d n + ed^{g+1})$ and moreover, a probabilistic construction achieves $m = O(d^{g+2} \log(n/d))$ and $e = \Omega(d \log(n/d))$ with probability $1 - o(1)$ (see Table 1). As a result, the upper and lower bounds on the number of rows of strongly disjunct and threshold disjunct matrices become equivalent up multiplicative constants when the lower threshold ℓ is zero.

Proposition 22 asserts that the notion of strongly disjunct matrices is in general stronger than threshold disjunct matrices. As we will see below, the former becomes strictly stronger when $\ell > 0$. As the lower threshold ℓ becomes larger, the discrepancy between the number of rows achievable by threshold disjunct matrices and strongly disjunct matrices becomes more significant (see Table 1).

4.3 Probabilistic Upper Bounds

As pointed out after Definition 20, the generalized definition of regular matrices may affect the bounds obtained by our probabilistic and explicit constructions (Constructions 2 and 4) only by hidden factors depending on u (essentially without any change in the proofs). For the case of generalized disjunct matrices, however, the bounds may substantially change depending on the gap parameter g.

Below we generalize Lemma 14 for the case of threshold disjunct matrices and show that Construction 2 results in a threshold $(d, \Omega_u(pd\log(n/d)/(1-p)^2); u, g)$-disjunct matrix (with probability $1 - o(1)$) if the number of measurements is increased by a factor $O(d^g)$. More precisely, we can now show the following lemma.

Lemma 23 (Lemma 14, generalized) *For every $p \in [0, 1)$ and integer parameters $u > g \geq 0$, Construction 2 with $m' = O_u(d^{g+2}\log(n/d)/(1-p)^2)$ outputs a threshold $(d, \Omega_u(pm'/d^{g+1}); u, g)$-disjunct matrix with probability $1 - o(1)$.*

Proof The proof essentially follows along the same lines as the proof of Lemma 14. The difference, compared to the case $g = 0$ covered by Lemma 14, is that we have a set I of distinguished columns $I \subseteq [n]$ in Definition 20 where $|I| = g + 1$ and the random vector $\boldsymbol{w}$ in the proof of Lemma 14 must have ones at all positions picked by I. With this requirement, the lower bound on the success probability q in (3) becomes $c = \Omega_u(1/d^{g+1})$. The rest of the proof remains unchanged except for the new lower bound on c, which makes the error tolerance parameter e in the proof lower bounded by $\Omega(pm'/d^{g+1})$, while increasing the parameter m' to a quantity upper bounded by $O_u(d^{g+2}\log(n/d)/(1-p)^2)$. □

4.4 The Direct Product

Lemma 9 can be extended to positive gaps as follows.

Lemma 24 (Lemma 9, generalized) *Suppose that M_1 is a $(d, e_1; u, g + 1)$-regular and M_2 is a strongly $(2d, e_2; g + 1)$-disjunct matrix. Then $M_1 \odot M_2$ is a threshold $(d, (e_1 + 1)(e_2 + 1) - 1; u, g)$-disjunct matrix.*

Proof Let $\ell := u - g - 1$ and $M := M_1 \odot M_2$. Towards verifying that M satisfies the requirements of Definition 20, consider a set $I \subseteq [n]$ of distinguished columns of M, where n is the number of columns of the matrices and $|I| = g + 1$, in addition to critical and zero sets $S, Z \subseteq [n]$ as in Definition 20 satisfying $|Z| \leq |S|$. Index the rows of M naturally by the elements of $[m_1] \times [m_2]$, where m_1 and m_2 are the number of rows of M_1 and M_2, respectively, and the (i, j)th row of M is the bitwise disjunction of the ith row of M_1 and the jth row of M_2.

Let $S' := S \setminus I$ and $Z' := Z \cup (S \setminus I)$. Observe that $|Z| \leq |S| \leq |S'| + g + 1 = |S| + |I|$ and $|Z'| \leq 2d$. From Definition 20, there is a set $E_1 \subseteq [m_1]$ of size more than e_1 such that $M_1|_{S'}$ has weight exactly $\ell = u - |I|$, and $M_2|_Z$ is all zeros. Moreover, there is a set $E_2 \subseteq [m_2]$ of size more than e_2 at which $M_2|_I$ has all ones and $M_2|_{Z'}$ has all zeros. This means that, at all rows corresponding to $E_1 \times E_2$, the product matrix

M has weight exactly $\ell + |I| = u$ at positions corresponding to S and all zeros at positions corresponding to Z. Therefore, M indeed u-satisfies any choice of the sets I, S, Z at more than $(e_1 + 1)(e_2 + 1) - 1$ rows. □

Consequently, using the coding-theoretic construction of strongly disjunct matrices described in Sect. 5, our explicit constructions of threshold $(d, e; u)$-disjunct matrices obtained in Sect. 3 can be extended to the positive gap model at the cost of a factor $O(d^g)$ increase in the number of measurements. The results from combining the above lemma with various constructions of regular and strongly disjunct matrices are summarized in Table 1.

4.5 Lower Bounds

We now extend the lower bounds proved in Sect. 2.3 to the positive-gap case, and show that the optimal exponent of d in the number of measurements is $g + 1$.

The lower bound on the number of rows of threshold disjunct matrices is an immediate consequence of Proposition 22.

Theorem 25 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ threshold $(d, e; u, g)$-disjunct matrix. Then, for some absolute constant $c > 0$,*

$$\begin{aligned} m &\geq 0.7c \frac{\binom{g+d+1}{g+1}(g+d+1)}{\log \binom{g+d+1}{g+1}} \log n + 0.5c \binom{g+d+1}{g+1} e \\ &= \Omega\big((d/g)^{g+2} \log_d n + (d/g)^{g+1} e\big). \end{aligned}$$

Proof Immediate from combining Proposition 22 and Theorem 2.19 of [39] that proves a lower bound on the number of rows of strongly disjunct matrices. The asymptotic simplification is straightforward. □

In order to lower bound the number of measurements in a threshold design, we first generalize Lemma 11 as follows.

Lemma 26 *Let M be a $(d + g, e; u, g - 1)$-threshold design, and $\ell := u - (g - 1) - 1 = u - g$. be the lower threshold Then M satisfies the following property:*

"For every $S \subseteq [n]$ such that $|S| = d$ and every $T \subseteq [n] \setminus S$ such that $|T| = g$, there are more than e rows of M at which $M|_T$ consists of all ones and moreover in those rows, $M|_S$ has weight exactly ℓ."

Proof Let $D := d + g$ be the sparsity parameter in the threshold model that M is designed for. In order to verify the claimed property for a given choice of S and T, consider the D-sparse vectors $x, x' \in \{0, 1\}^n$ such that $\mathsf{supp}(x) = S \cup T$ and $\mathsf{supp}(x') = S$. Let $y := M[x]_{\ell+1}$ and $y' := M[x']_u$, and observe that $y, y' \in M[x]_{\ell,u}$. Also, since x' is point-wise less than or equal to x, or in symbols $x' \preceq x$, by monotonicity it also follows that $y' \preceq y$.

Thus, we know from the assumption that there are more than e positions at which y and y' are different. Let j be any such position. In particular, we know that $y(j) = 1$ and $y'(j) = 0$. Therefore, by Definition 19 and the way that the threshold model is defined, the submatrix $M|_{\mathsf{supp}(x')}$ must have weight at most ℓ at the jth row and $M|_{\mathsf{supp}(x)}$ must have weight at least u at the jth row. Since the support of x is chosen to be larger than the support of x' at exactly g positions, and $g = u - \ell$, the only possibility is to have $M|_{\mathsf{supp}(x)}$ (that is, $M|_{S\cup T}$) with weight *exactly* u at the jth row and $M|_{\mathsf{supp}(x)\setminus\mathsf{supp}(x')}$ (that is, $M|_T$) having all ones at the jth row. In turn, this implies that $M|_{\mathsf{supp}(x')}$ (that is, $M|_S$) must have weight exactly ℓ at that row.

This concludes proof, since the argument holds for every possible choice of the distinguishing entry j. □

The following theorem is the analogous version of Theorem 12 for the positive-gap case.

Theorem 27 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ Boolean matrix that satisfies the property quoted in Lemma 26. Then,*

$$m = \Omega\left(\frac{m'}{\log m'} + \left(\frac{d}{\ell+1}\right)^2 \log_d n + \frac{de}{(\ell+1)^2}\right), \quad \textit{where}$$
$$m' := \frac{(d/\ell g)^{g+1}\log_d n + (d/\ell g)^g e}{(g+\ell)\log n}.$$

In particular, when ℓ, g are absolute constants, we have

$$m = \Omega_{\ell,g}\left(\frac{1}{\log d + \log(e+2)} \cdot \left(\frac{d^{g+1}}{\log d} + \frac{d^g e}{\log n}\right) + d^2 \log_d n + de\right).$$

Proof First, observe that the property quoted in Lemma 26 is stronger than the property quote in Lemma 11. Thus, the lower bound of Theorem 12 holds; namely, we have

$$m = \Omega\left(\left(\frac{d}{\ell+1}\right)^2 \log_d n + \frac{de}{(\ell+1)^2}\right).$$

Therefore, it suffices to show that $m = \Omega(m'/\log m')$. Given the matrix M, we will use a "random resampling method" to create a strongly disjunct matrix out of M, and will then use the known lower bounds related to strongly disjunct matrices.

Given a vector $v \in \{0,1\}^n$, a *resampling* of v is a random vector $v' \in \{0,1\}^n$ defined in the following way: For ever $i \in [n]$, if $v(i) = 0$, then we set $v'(i) = 0$. Otherwise, we independently set $v'(i) = 1$ with probability $1/\ell$ and zero with the remaining probability.

Let $t > 0$ be an integer parameter to be determined later. Let $M_1, \ldots, M_t$ be random $m \times n$ Boolean matrices, such that each M_i is obtained from M by independently resampling each row of the matrix. Also, define the $mt \times n$ Boolean matrix M' by stacking $M_1, \ldots, M_t$ on top of one another. We will argue that, if t is chosen sufficiently large, there is a nonzero probability that M' becomes a strongly disjunct matrix. Thus, there is a fixing of the resampling randomness that indeed makes M' strongly disjunct, which then allows us to obtain the desired lower bound.

Consider any triple (j, T, W) where $j \in [m]$, $T, W \subseteq [n]$ such that $|T| = g$, $|W| = \ell$, $T \cap W = \emptyset$, and moreover, the jth row of M has all-ones at the columns picked by T and W. We say that the triple *survives* in M_i when on the jth row of M_i, the columns picked by T all contain ones and those picked by W all contain zeros. The probability of this happening is exactly

$$p = (1 - 1/\ell)^{\ell} \cdot 1/\ell^{g} \geq c/\ell^{g},$$

for some absolute constant $c > 0$. The probability that the triple does not survive in any of $M_1, \dots, M_t$ is

$$(1 - p)^t \leq (1 - c/\ell^g)^t \leq C^{t/\ell^g},$$

for some absolute constant $C \in (0, 1)$. Combined with a union bound on all choices of (j, T, W), we deduce that the probability that some triple does not survive in any of $M_1, \dots, M_t$ is at most

$$mn^{g+\ell} C^{t/\ell^g} = 2^{\log m + (g+\ell)\log n - (t/\ell^g)\log(1/C)},$$

which is strictly less than 1 for some large enough choice of t, namely, for $t \geq t_0$ such that

$$t_0 = O(\ell^g((g + \ell)\log n + \log m)).$$

Now, pick $t := t_0$ and fix the resampling randomness so that all triples (j, T, W) survive. We claim that the matrix M' is strongly $(d, e; g)$-disjunct.

In order to verify the disjunctness property, consider any choice of sets $S, T \subseteq [n]$ such that $|S| = d$ and $|T| = g$. Let J be the set of rows of M where $M|_T$ has all-ones and $M|_S$ has Hamming weight ℓ. By the property quoted in Lemma 26, we know that $|J| > e$.

For any $j \in J$, we know that the jth row of $M|_S$ is supported on some set $W \subseteq [n]$ of size ℓ. We know, on the other hand, that the triplet (j, T, W) survives in some M_i. Clearly, by the way we defined the survival property, this implies that jth row of M_i (and thus, the corresponding row in M') contains all-ones at columns picked by T and all-zeros at columns picked by S. Since this argument holds for any choice of S, T, and j, we conclude that M' is strongly $(d, e; g)$-disjunct.

The number of rows of M' is mt. We can now apply the known lower bounds on the number of rows of strongly disjunct matrices in order to lower bound the number of rows of M'. In particular, Theorem 2.19 of [39] implies that, for some absolute constant $c' > 0$, and whenever n is sufficiently large for the given parameter d,

$$\begin{aligned} mt &\geq 0.7c' \frac{\binom{g+d}{g}(g + d)}{\log \binom{g+d}{g}} \log n + 0.5c' \binom{g + d}{g} e \\ &\geq 0.7c' \frac{(g + d)^{g+1} \log n}{\log(g + d) g^{g+1}} \log n + 0.5c' \frac{(g + d)^g e}{g^g} \\ &= \Omega\big((d/g)^{g+1} \log_d n + (d/g)^g e\big). \end{aligned}$$

Now we substitute the chosen value of t in the above bound to obtain

$$m = \Omega\left(\frac{(d/\ell g)^{g+1}\log_d n + (d/\ell g)^g e}{(g+\ell)\log n + \log m}\right).$$

Now if n is sufficiently large for the given d, we can ensure that the conditions of Proposition 31 in the appendix are satisfied, and the above bound implies that

$$m = \Omega(m'/\log m'), \quad \text{where } m' := \frac{(d/\ell g)^{g+1}\log_d n + (d/\ell g)^g e}{(g+\ell)\log n},$$

as claimed. The simplification when ℓ and g are absolute constants is straightforward. □

Theorem 27 combined with Lemma 26 implies the desired lower bound on the number of measurements of a threshold design. The following corollary summarizes the simplified bounds for the case $e = 0$.

Corollary 28 *For every integer $d > 0$ there is an $n_0 > 0$ such that the following holds. For any $n \geq n_0$, let M be an $m \times n$ Boolean matrix that is a $(d, 0; u, g)$-*threshold design, *for constants $u > g \geq 0$. Then,*

$$m = \Omega_u\left(\frac{d^{g+2}}{\log^2 d} + \frac{d^2 \log n}{\log d}\right).$$

5 Strongly Disjunct Matrices from Codes

A well known coding-theoretic construction of combinatorial designs, and classical disjunct matrices is due to Kautz and Singleton [29], which was further refined in several subsequent works (such as [21, 22]).

In this section we describe a construction of strongly disjunct matrices (as in Definition 3) which is a straightforward extension of the original construction of Kautz and Singleton. Construction 5 explains the idea, which is analyzed in Lemma 29 below. In this section we use standard tools from the theory of error-correcting codes.[6] The interested reader is referred the standard texts in coding theory (e.g., the books by MacWilliams and Sloane [32], van Lint [42], and Roth [36]) for background.

Lemma 29 *Construction 5 outputs a strongly $(d, e; u)$-disjunct matrix for every $d < (\tilde{n} - e)/((\tilde{n} - \tilde{d})u)$.*

Proof Let $C := \{c_1, \ldots, c_u\} \subseteq [n]$ and $C' := \{c'_1, \ldots, c'_d\} \subseteq [n]$ be disjoint subsets of column indices. We wish to show that, for more than e rows of M, the entries at positions picked by C are all-ones while those picked by C' are all-zeros. For each

[6] We use the standard coding-theoretic notation of $(\tilde{n}, k, \tilde{d})_q$ code for a q-ary code of length $\tilde{n}$, size q^k, and minimum distance at least $\tilde{d}$.

- *Given:* An $(\tilde{n}, k, \tilde{d})_q$ error-correcting code $\mathcal{C} \subseteq [q]^{\tilde{n}}$, and integer parameter $u > 0$.
- *Output:* An $m \times n$ Boolean matrix M, where $n = q^k$, and $m = \tilde{n} q^u$.
- *Construction:* First, consider the mapping $\varphi \colon [q] \to \{0, 1\}^{q^u}$ from q-ary symbols to column vectors of length q^u defined as follows. Index the coordinates of the output vector by the u-tuples from the set $[q]^u$. Then $\varphi(x)$ has a 1 at position $(a_1, \ldots, a_u)$ if and only if there is an $i \in [u]$ such that $a_i = x$. Arrange all codewords of $\mathcal{C}$ as columns of an $\tilde{n} \times q^k$ matrix M' with entries from $[q]$. Then replace each entry x of M' with $\varphi(x)$ to obtain the output $m \times n$ matrix M.

Construction 5: Extension of Kautz-Singleton's method [29]

$j \in [n]$, denote the jth column of M' by $M'(j)$, and let $M'(C) := \{M'(c_j) \colon j \in [u]\}$, and $M'(C') := \{M'(c'_j) \colon j \in [d]\}$.

From the minimum distance of $\mathcal{C}$, we know that every two distinct columns of M' agree in at most $\tilde{n} - \tilde{d}$ positions. By a union bound, for each $i \in [d]$, the number of positions where $M'(c'_i)$ agrees with one or more of the codewords in $M'(C)$ is at most $u(\tilde{n} - \tilde{d})$, and the number of positions where some vector in $M'(C')$ agrees with one or more of those in $M'(C)$ is at most $du(\tilde{n} - \tilde{d})$.

By assumption, we have $\tilde{n} - du(\tilde{n} - \tilde{d}) > e$, and thus, for a set $E \subseteq [\tilde{n}]$ of size greater than e, at positions picked by E none of the codewords in $M'(C')$ agree with any of the codewords in $M'(C)$.

Now let $w \in [q]^n$ be any of the rows of M' picked by E, and consider the $q^u \times n$ Boolean matrix W formed by applying the mapping $\varphi(\cdot)$ on each entry of w. We know that $\{w(c_j) \colon j \in [u]\} \cap \{w(c'_j) \colon j \in [d]\} = \emptyset$. Thus we observe that the particular row of W indexed by $(w(c_1), \ldots, w(c_u))$ (and in fact, any of its permutations) must have all-ones at positions picked by C and all-zeros at those picked by C'. As any such row is a distinct row of M, it follows that M is strongly $(d, e; u)$-disjunct. □

Here we mention a few specific instantiations of the above construction. Namely, we will first consider the family of Reed-Solomon codes, that are also used in the original work of Kautz and Singleton [29], and then move on to the family of algebraic geometric (AG) codes on the Tsfasman-Vlăduţ-Zink (TVZ) bound, Hermitian codes, and finally, codes on the Gilbert-Varshamov (GV) bound.

5.1 Reed-Solomon Codes

Let $p \in [0, 1)$ be an arbitrary "noise" parameter. If we take $\mathcal{C}$ to be an $[\tilde{n}, k, \tilde{d}]_{\tilde{n}}$ Reed-Solomon code over an alphabet of size $\tilde{n}$ (which we assume to be a prime power), where $\tilde{d} = \tilde{n} - k + 1$, we get a strongly disjunct $(d, e; u)$-matrix with $m = O(du \log n/(1 - p))^{u+1}$ rows and $e = p\tilde{n} = \Omega(pdu(\log n)/(1 - p))$.

5.2 Algebraic Geometric Codes on the TVZ Bound

Another interesting family for the code $\mathcal{C}$ is the family of algebraic geometric codes that attain the Tsfasman-Vlăduţ-Zink bound (cf. [26, 41]). This family is defined over any alphabet size $q \geq 49$ that is a square prime power, and achieves a minimum distance $\tilde{d} \geq \tilde{n} - k - \tilde{n}/(\sqrt{q} - 1)$. Let $e := pn$, for a noise parameter $p \in [0, 1)$. By Lemma 29, the underlying code $\mathcal{C}$ needs to have minimum distance at least $\tilde{n}(1 - (1 - p)/(du))$. Thus in order to be able to use the above-mentioned family of AG codes, we need to have $q \gg (du/(1 - p))^2 =: q_0$. Let us take an appropriate $q \in [2q_0, 8q_0]$, and following Lemma 29, $\tilde{n} - \tilde{d} = \lceil \tilde{n}(1 - p)/(du) \rceil$. Thus, the dimension of $\mathcal{C}$ becomes at least

$$k \geq \tilde{n} - \tilde{d} - \frac{\tilde{n}}{\sqrt{q} - 1} = \Omega\left(\frac{\tilde{n}(1 - p)}{du}\right) = \Omega(\tilde{n}/\sqrt{q_0}),$$

and subsequently[7] we get that $\log n = k \log q \geq k = \Omega(\tilde{n}/\sqrt{q_0})$. Now, noting that $m = q^u \tilde{n}$, we conclude that

$$m = q^u \tilde{n} = O(q_0^{u+1/2} \log n) = O\left(\frac{du}{1 - p}\right)^{2u+1} \log n,$$

and $e = \Omega(pdu(\log n)/(1 - p))$.

We see that the dependence of the number of measurements on the sparsity parameter d is worse for AG codes than Reed-Solomon codes by a factor d^u, but the construction from AG codes benefits from a linear dependence on $\log n$, compared to $\log^{u+1} n$ for Reed-Solomon codes. Thus, AG codes become more favorable only when the sparsity is substantially low; namely, when $d \ll \log n$.

5.3 Hermitian Codes

A particularly nice family of AG codes arises from the Hermitian function field. Let q' be a prime power and $q := q'^2$. Then the Hermitian function field over $\mathbb{F}_q$ is a finite extension of the rational function field $\mathbb{F}_q(x)$, denoted by $\mathbb{F}_q(x, y)$, where we have $y^{q'} + y = x^{q'+1}$. The structure of this function field is relatively well understood and the family of Goppa codes defined over the rational points of the Hermitian function field is known as Hermitian codes. This family is recently used by Ben-Aroya and Ta-Shma [1] for construction of small-bias sets. Below we quote some parameters of Hermitian codes from their work.

The number of rational points of the Hermitian function field is equal to $q'^3 + 1$, which includes a common pole Q_∞ of x and y. The genus of the function field is $\tilde{g} = q'(q' - 1)/2$. For some integer parameter r, we take $G := rQ_\infty$ as the divisor defining the Riemann-Roch space $\mathcal{L}(G)$ of the code $\mathcal{C}$, and the set of rational points except Q_∞ as the evaluation points of the code. Thus the length of $\mathcal{C}$ becomes $\tilde{n} = q'^3$. Moreover, the minimum distance of the code is $\tilde{d} = n - \deg(G) = n - r$. When $r \geq 2\tilde{g} - 1$, the dimension of the code is given by the Riemann-Roch theorem, which is equal to $r - \tilde{g} + 1$. For the low-degree regime where $r < 2\tilde{g} - 1$, the dimension k of the code is the size of the Wirestrauss semigroup of G, which turns out to be the set $W = \{(i, j) \in \mathbb{N}^2 : j \leq q' - 1 \wedge iq' + j(q' + 1) \leq r\}$.

[7] Note that, given the parameters p, d, n, the choice of q depends on p, d, as explained above, and then one can choose the code length $\tilde{n}$ to be the smallest integer for which we have $q^k \geq n$. But for the sake of clarity we have assumed that $q^k = n$, which does not affect the asymptotic bounds.

Now, given parameters d, p of the disjunct matrix, define $\rho := (1-p)/((d+1)u)$, take the alphabet size q as a square prime power, and set $r := \rho q^{3/2}$. First we consider the case where $r < 2\tilde{g} - 1 = 2q - 2\sqrt{q} - 1$. In this case, the dimension of the Hermitian code becomes $k = |W| = \Omega(r^2/q) = \Omega(\rho^2 q^2)$. The distance $\tilde{d}$ of the code satisfies $\tilde{d} = \tilde{n} - r \geq \tilde{n}(1-\rho)$ and thus, for $e := p\tilde{n}$, conditions of Lemma 29 are satisfied. The number of the rows of the resulting measurement matrix becomes $m = q^{u+3/2}$, and we have $n = q^k$. Therefore,

$$\begin{aligned}\log n = k \log q \geq k = \Omega(\rho^2 q^2) \quad &\Rightarrow \quad q = O(\sqrt{\log n}/\rho)\\ &\Rightarrow \quad m = O\left(\left(\frac{d\sqrt{\log n}}{1-p}\right)^{u+3/2}\right),\end{aligned}$$

and in order to ensure that $r < 2\tilde{g} - 1$, we need to have $du/(1-p) \gg \sqrt{\log n}$. On the other hand, when $du/(1-p) \ll \sqrt{\log n}$, we are in the high-degree regime, in which case the dimension of the code becomes $k = r - \tilde{g} + 1 = \Omega(r) = \Omega(\rho q^{3/2})$, and we will thus have

$$q = O((\log n/\rho)^{2/3}) \quad \Rightarrow \quad m = O\left(\left(\frac{d \log n}{1-p}\right)^{1+2u/3}\right)$$

Altogether, we conclude that Construction 5 with Hermitian codes results in a strongly $(d, e; u)$-disjunct matrix with

$$m = O\left(\left(\frac{d\sqrt{\log n}}{1-p} + \left(\frac{d \log n}{1-p}\right)^{2/3}\right)^{u+3/2}\right)$$

rows, where $e = p \cdot \Omega(d(\log n)/(1-p) + (d\sqrt{\log n}/(1-p))^{3/2})$. Compared to the Reed-Solomon codes, the number of measurements has a slightly worse dependence on d, but a much better dependence on n. Compared to AG codes on the TVZ bound, the dependence on d is better while the dependence on n is inferior.

5.4 Codes on the Gilbert-Varshamov Bound

A q-ary $(\tilde{n}, k, \tilde{d})$-code (of sufficiently large length) is said to be on the Gilbert-Varshamov bound if it satisfies $k \geq \tilde{n}(1 - h_q(\tilde{d}/\tilde{n}))$, where $h_q(\cdot)$ is the q-ary entropy function defined as

$$h_q(x) := x \log_q(q-1) - x \log_q(x) - (1-x)\log_q(1-x).$$

It is well known that a random linear code achieves the bound with overwhelming probability (cf. [32]). Now we apply Lemma 29 on a code on the GV bound, and calculate the resulting parameters. Let $\rho := (1-p)/(4du)$, choose any alphabet size $q \in [1/\rho, 2/\rho]$, and let $\mathcal{C}$ be any q-ary code of length $\tilde{n}$ on the GV bound, with minimum distance $\tilde{d} \geq \tilde{n}(1 - 2/q)$. By the Taylor expansion of the function $h_q(x)$ around $x = 1 - 1/q$, we see that the dimension of $\mathcal{C}$ asymptotically behaves as $k =$

$\Theta(\tilde{n}/(q \log q))$. Thus, the number of columns of the resulting measurement matrix becomes $n = q^k = 2^{\Omega(\tilde{n}/q)}$. Moreover, the number m of its rows becomes

$$m = q^u \tilde{n} = O(q^{u+1} \log n) = O((d/(1-p))^{u+1} \log n),$$

and the matrix becomes strongly $(d, e; u)$-disjunct for $e = p\tilde{n} = \Omega(pd(\log n)/(1-p))$.

We remark that for the range of parameters that we are interested in, Porat and Rothschild [34] have obtained a deterministic construction of linear codes on the GV bound that runs in time $\mathsf{poly}(q^k)$ (and thus, polynomial in the size of the resulting measurement matrix).

Their construction is based on a derandomization of the probabilistic argument for random linear codes using the method of conditional expectations, and as such, can be considered *weakly explicit* (in the sense that, the entire measurement matrix can be computed in polynomial time in its length; whereas for a fully explicit construction one must ideally be able to deterministically compute any single entry of the measurement matrix in time $\mathsf{poly}(d, \log n)$, which is not the case for this construction). Altogether, we obtain the following result.

Theorem 30 *There is an algorithm that, given integer parameters $d \leq n$ and $u > 0$ and real parameter $p \in [0, 1)$ outputs an $m \times n$ binary matrix which is strongly $(d, e; u)$-disjunct. The parameters m and e satisfy the bounds $m = O((d/(1-p))^{u+1} \log n)$ and $e = \Omega(pd(\log n)/(1-p))$. Moreover, the running time of the algorithm is polynomial in mn.*

Using a standard probabilistic argument it is easy to see that a random $m \times n$ matrix, where each entry is an independent Bernoulli random variable with probability $1/d$ of being 1, is with overwhelming probability strongly $(d, e; u)$-disjunct for $e = \Omega(pd \log(n/d)/(1-p)^2)$ and $m = O(d^{u+1}(\log(n/d))/(1-p)^2)$ (the proof is very similar to the proof of Lemma 14). Thus we see that, for a fixed p, Construction 5 when using codes on the GV bound almost matches these parameters. Moreover, the explicit construction based on Reed-Solomon codes possesses the “right” dependence on the sparsity d, while AG codes on the TVZ bound have a matching dependence on the vector length n with random measurement matrices, and finally, the trade-off offered by the construction based on Hermitian codes lies in between the one for Reed-Solomon codes and AG codes.

6 Concluding Remarks

In this work we have introduced the combinatorial notion of regular binary matrices, that is used as an intermediate tool towards obtaining threshold testing designs.

Even though our construction, assuming an optimal lossless condenser, matches the probabilistic upper bound for regular matrices, the number of measurements in the resulting threshold testing scheme (obtained from the simple direct product in Construction 1) becomes larger than the probabilistic upper bound by a factor of $\Omega(d \log n)$. Thus, an outstanding question is directly constructing threshold disjunct matrices that match the probabilistic upper bound. Despite this, the notion of regular matrices may be of independent interest, and an interesting question is to obtain (nontrivial) concrete lower bounds on the number of rows of such matrices in terms of the parameters n, d, e, u (and the gap parameter g in the generalized definition of Sect. 4).

Moreover, in this work we have assumed the upper threshold u to be a fixed constant, allowing the constants hidden in asymptotic notions to have a poor dependence on u. An outstanding question is whether the number of measurements can be reasonably controlled when the upper threshold u and possibly the gap parameter g become large; e.g., $g, u = \Omega(d)$.

The lower bound proved in Corollary 28 on the number of rows of threshold designs shows an exponent $g+2$ for the sparsity parameter, which matches the upper bounds obtained using the probabilistic method. We conjecture that this bound can be improved to $\Omega_u(d^{g+2}\log_d n)$ and more generally when $e > 2$, to $\Omega_u(d^{g+2}\log_d n + d^{g+1}e)$. In other words, for fixed thresholds, we suspect that the asymptotic bounds for $(d, e; u, g)$-threshold designs and strongly $(d, e; g+1)$-disjunct matrices should nearly be the same.

Another interesting problem is decoding. While our constructions can combinatorially guarantee identification of sparse vectors, for applications it is important to have an efficient reconstruction algorithm as well. Contrary to the case of strongly disjunct matrices that allow a straightforward decoding procedure (cf. [5]), it is not clear whether in general our notion of disjunct matrices allow efficient decoding, and thus it becomes important to look for constructions that are equipped with efficient reconstruction algorithms.

Finally, for clarity of the exposition, in this work we have only focused on asymptotic trade-offs, and it would be nice to obtain good, finite length, estimates on the obtained bounds that are useful for applications.

Acknowledgements The author thanks Christian Deppe, Arkadii D'yachkov, and Vyacheslav Rykov for fruitful discussions, and anonymous referees for their feedback on earlier drafts of this work.

Appendix: A Technical Lemma

The following simple proposition is used in the proof of Theorem 27.

Proposition 31 *Suppose for some values of $a > 0$, $b, m \geq 2$, and $c \geq 2b/a$, we have $m \geq c \cdot \frac{a}{b+\log m}$, where the logarithm is to base 2. Then,*

$$m \geq \frac{(ac/b)}{\log(ac/b)}.$$

Proof We can write

$$m \geq c \cdot \frac{a}{b+\log m} \geq \frac{ca}{b\log m} \quad \Rightarrow \quad m\log m \geq ac/b,$$

where the second inequality is from the assumption that $b, m \geq 2$.

Since $m \log m$ is an increasing and convex function of m, we know that $m \geq m_0$, where m_0 is the solution to the equation $m_0 \log m_0 = ac/b$. Thus it suffices to lower bound m_0.

Since $ac/b \geq 2$ by assumption, it follows that $m_0 \leq m_0 \log m_0 = ac/b$, and thus,

$$m_0 \log(ac/b) \geq m_0 \log m_0 = ac/b \quad \Rightarrow \quad m_0 \geq \frac{(ac/b)}{\log(ac/b)},$$

and the claim follows. □

References

1. Ben-Aroya, A., Ta-Shma, A.: Constructing small-bias sets from algebraic-geometric codes. In: Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS) (2009)
2. Blass, A., Gurevich, Y., testing, P.: Bull. Eur. Assoc. Theor. Comput. Sci. **78**, 100–132 (2002)
3. Bruno, W.J., Knill, E., Balding, D.J., Bruce, D.C., Doggett, N.A., Sawhill, W.W., Stallings, R.L., Whittaker, C.C., Torney, D.C.: Efficient pooling designs for library screening. Genomics **26**(1), 21–30 (1995)
4. Capalbo, M., Reingold, O., Vadhan, S., Wigderson, A.: Randomness conductors and constant-degree expansion beyond the degree/2 barrier. In: Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC), pp. 659–668 (2002)
5. Chen, H.-B., Fu, H.-L.: Nonadaptive algorithms for threshold group testing. Discrete Appl. Math. **157**, 1581–1585 (2009)
6. Chen, H.-B., Du, D.-Z., Hwang, F.-K.: An unexpected meeting of four seemingly unrelated problems: graph testing, DNA complex screening, superimposed codes and secure key distribution. J. Comb. Optim. **14**(2–3), 121–129 (2007)
7. Chen, H.-B., Fu, H.-L., Hwang, F.-K.: An upper bound of the number of tests in pooling designs for the error-tolerant complex model. Optim. Lett. **2**(3), 425–431 (2008)
8. Cheng, Y., Du, D.-Z.: New constructions of one- and two-stage pooling designs. J. Comput. Biol. **15**(2), 195–205 (2008)
9. Cheraghchi, M.: Noise-resilient group testing: limitations and constructions. In: Proceedings of the 17th International Symposium on Fundamentals of Computation Theory (FCT). Lecture Notes in Computer Science, vol. 5699, pp. 62–73 (2009)
10. Cheraghchi, M.: Applications of derandomization theory in coding. Ph.D. Thesis, EPFL, Lausanne, Switzerland (2010). Available online at http://eccc.hpi-web.de/static/books/Applications_of_Derandomization_Theory_in_Coding/
11. Cheraghchi, M.: Improved constructions for non-adaptive threshold group testing. In: Proceedings of the 37th International Colloquium on Automata, Languages and Programming (ICALP) (2010). arXiv:1002.2244v3 [cs.DM]
12. Clifford, R., Efremenko, K., Porat, E., Rothschild, A.: k-mismatch with don't cares. In: Proceedings of the 15th European Symposium on Algorithm (ESA). Lecture Notes in Computer Science, vol. 4698, pp. 151–162 (2007)
13. Cormode, G., Muthukrishnan, S.: What's hot and what's not: tracking most frequent items dynamically. ACM Trans. Database Syst. **30**(1), 249–278 (2005)
14. Cormode, G., Muthukrishnan, S.: Combinatorial algorithms for compressed sensing. In: Proceedings of Information Sciences and Systems, pp. 198–201 (2006)
15. Damaschke, P.: Threshold group testing. In: General Theory of Information Transfer and Combinatorics. Lecture Notes in Computer Science, vol. 4123, pp. 707–718. Springer, Berlin (2006)
16. Dorfman, R.: The detection of defective members of large populations. Ann. Math. Stat. **14**, 436–440 (1943)
17. Du, D.-Z., Hwang, F.: Combinatorial Group Testing and Its Applications, 2nd edn. World Scientific, Singapore (2000)
18. Du, D.-Z., Hwang, F.-K.: Pooling Designs and Nonadaptive Group Testing. World Scientific, Singapore (2006)

19. D'yachkov, A.G., Rykov, V.V.: Bounds of the length of disjunct codes. Probl. Control Inf. Theory **11**, 7–13 (1982)
20. D'yachkov, A.G., Rykov, V.V., Rashad, A.M.: Superimposed distance codes. Probl. Control Inf. Theory **18**(4), 237–250 (1989)
21. D'yachkov, A.J., an Macula, A.G., Rykov, V.V.: New applications and results of superimposed code theory arising from the potentialities of molecular biology. In: Numbers, Information and Complexity, pp. 265–282 (2000)
22. D'yachkov, A.J., an Macula, A.G., Rykov, V.V.: New constructions of superimposed codes. IEEE Trans. Inf. Theory **46**(1), 284–290 (2000)
23. D'yachkov, A., Vilenkin, P., Macula, A., Torney, D.: Families of finite sets in which no intersection of ℓ sets is covered by the union of s others. J. Comb. Theory, Ser. A **99**, 195–218 (2002)
24. Farach, M., Kannan, S., Knill, E., Muthukrishnan, S.: Group testing problems with sequences in experimental molecular biology. In: Proceedings of Compression and Complexity of Sequences, pp. 357–367 (1997)
25. Gao, H., Hwang, F.K., Thai, M.T., Wu, W., Znati, T.: Construction of $d(H)$-disjunct matrix for group testing in hypergraphs. J. Comb. Optim. **12**, 297–301 (2006)
26. Garcia, A., Stichtenoth, H.: A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vlăduţ bound. Invent. Math. **121**, 211–222 (1995)
27. Guruswami, V., Umans, C., Vadhan, S.: Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. In: Proceedings of the 22nd IEEE Conference on Computational Complexity (CCC) (2007)
28. Hong, E.-S., Ladner, R.E.: Group testing for image compression. In: Data Compression Conference, pp. 3–12 (2000)
29. Kautz, W.H., Singleton, R.C.: Nonrandom binary superimposed codes. IEEE Trans. Inf. Theory **10**, 363–377 (1964)
30. Kim, H.K., Lebedev, V.: On optimal superimposed codes. J. Comb. Des. **12**, 79–91 (2004)
31. Macula, A.J.: Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. Ann. Comb. **3**(1), 61–69 (1999)
32. MacWilliams, F.J., Sloane, N.J.: The Theory of Error-Correcting Codes. North Holand, Amsterdam (1977)
33. Ngo, H.-Q., Du, D.-Z.: A survey on combinatorial group testing algorithms with applications to DNA library screening. In: DIMACS Series on Discrete Math. and Theoretical Computer Science, vol. 55, pp. 171–182 (2000)
34. Porat, E., Rothschild, A.: Explicit non-adaptive combinatorial group testing schemes. In: Proceedings of the 35th International Colloquium on Automata, Languages and Programming (ICALP). Lecture Notes in Computer Science, vol. 5125, pp. 748–759 (2008)
35. Radhakrishan, J., Ta-Shma, A.: Tight bounds for depth-two superconcentrators. In: Proceedings of the 38th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp. 585–594 (1997)
36. Roth, R.M.: Introduction to Coding Theory. Cambridge University Press, Cambridge (2006)
37. Ruszinkó: On the upper bound of the size of the r-cover-free families. J. Comb. Theory, Ser. A **66**, 302–310 (1994)
38. Schliep, A., Torney, D., Rahmann, S.: Group testing with DNA chips: Generating designs and decoding experiments. In: Proceedings of Computational Systems Bioinformatics (2003)
39. Stinson, D.R., Wei, R.: Generalized cover-free families. Discrete Math. **279**, 463–477 (2004)
40. Stinson, D.R., Wei, R., Zhu, L.: Some new bounds for cover-free families. J. Comb. Theory, Ser. A **90**, 224–234 (2000)
41. Tsfasman, M.A., Vlăduţ, S.G., Zink, Th.: Modular curves, Shimura curves, and Goppa codes better than the Varshamov-Gilbert bound. Math. Nachr. **109**, 21–28 (1982)
42. van Lint, J.H.: Introduction to Coding Theory, 3rd edn. Graduate Texts in Mathematics, p. 86. Springer, Berlin (1998)
43. Wolf, J.: Born-again group testing: multiaccess communications. IEEE Trans. Inf. Theory **31**, 185–191 (1985)
44. Wu, W., Huang, Y., Huang, X., Li, Y.: On error-tolerant DNA screening. Discrete Appl. Math. **154**(12), 1753–1758 (2006)
45. Wu, W., Li, Y., Huang, C.H., Du, D.Z.: Molecular biology and pooling design. Data Min. Biomed. **7**, 133–139 (2008)