Sparse Combinatorial Group Testing

Huseyin A. Inan, Peter Kairouz, *Member, IEEE*, and Ayfer Özgür, *Senior Member, IEEE*

***Abstract*— In combinatorial group testing, the primary objective is to fully identify the set of at most d defective items from a pool of n items using as few tests as possible. The celebrated result for the combinatorial group testing problem is that the number of tests, denoted by t, can be made logarithmic in n when $d = O(\text{poly}(\log n))$. However, state-of-the-art group testing codes require the items to be tested $w = \Omega\left(\frac{d \log n}{\log d + \log\log n}\right)$ times and tests to include $\rho = \Omega\left(\frac{n}{d \log_d n}\right)$ items. In many emerging applications, items can only participate in a limited number of tests and tests are constrained to include a limited number of items. In this paper, we study the "sparse" regime for the group testing problem where we restrict the number of tests each item can participate in by $w_{\max}$ or the number of items each test can include by $\rho_{\max}$ in both noiseless and noisy settings. These constraints lead to a largely unexplored regime where t is a fractional power of n, rather than logarithmic in n as in the classical setting. Our results characterize the number of tests t needed in this regime as a function of $w_{\max}$ or $\rho_{\max}$ and show, for example, that t decreases drastically when $w_{\max}$ is increased beyond a bare minimum. In particular, in the noiseless case it can be shown that if $w_{\max} \leq d$, then we must have $t = n$, i.e., testing every item individually is optimal. We show that if $w_{\max} = d+1$, the number of tests decreases suddenly from $t = n$ to $t = \Theta(d\sqrt{n})$. The order-optimal construction is obtained via a modification of the classical Kautz-Singleton construction, which is known to be suboptimal for the classical group testing problem. For the more general case, when $w_{\max} = ld + 1$ for integer $l > 1$, the modified Kautz-Singleton construction requires $t = \Theta\left(dn^{\frac{1}{l+1}}\right)$ tests, which we prove to be near order-optimal. We also show that our constructions have a favorable encoding and decoding complexity, i.e. they can be decoded in $(\text{poly}(d) + O(t))$-time and each entry in any codeword can be computed in $\text{poly}(\log n)$ memory space. We finally discuss an application of our results to the construction of energy-limited random access schemes for Internet of Things networks, which provided the initial motivation for our work.**

***Index Terms*— Group testing, Internet of Things (IoT), machine-to-machine communications, error-correcting codes, sparse codes.**

I. Introduction

GROUP testing is a subfield of combinatorial mathematics that studies how to identify a set of d (or less) defective items from a large population of size n. For an unknown sequence $x \in \{0,1\}^n$ with at most d ones representing the defective items, we are allowed to test any subset $S \subseteq \{1,\ldots,n\}$ of the items. The result of a test S could either be positive, which happens when at least one item in S is defective (i.e., $\exists\, i \in S$ such that $x_i = 1$), or negative when all the items in S are not defective (i.e., $\forall i \in S$ we have $x_i = 0$). The goal is to design as few tests as possible (say t tests) so that we can recover the unknown sequence x. In this paper, we focus on the zero-error criterion, i.e., we require the exact identification of the unknown sequence x without any error. This problem is referred to as combinatorial group testing in the literature [1].

The original group testing framework was developed in 1943 by Dorfman [2]. At the time, group testing was devised to identify which WWII draftees were infected with syphilis – without having to test them individually. In Dorfman's application, items represented draftees and tests represented actual blood tests. Over the years, group testing has found numerous applications in an array of exciting fields spanning biology [3], medicine [4], machine learning [5], data analysis [6], computer science [7], and signal processing [8]. In addition, group testing has been extensively applied to various disciplines of wireless communication, such as multiple access control protocols [9]–[12] and neighborhood discovery [13].

The celebrated result for the combinatorial group testing problem is that t can be made logarithmic in n. One of the earliest explicit group testing constructions is due to Kautz and Singleton and requires $t = O(d^2 \log_d^2 n)$ tests [9]. This construction uses a Reed-Solomon code concatenated with a non-linear identity code and it matches the best known lower bound $\Omega(d^2 \log_d n)$ [14], [15] in the regime where $d = \Theta(n^\alpha)$ for some $\alpha \in (0,1)$. More recently, a different explicit construction achieving $t = O(d^2 \log n)$ was introduced by Porat and Rothschild in [16], which outperforms the Kautz-Singleton construction in the regime where $d = O(\text{poly}(\log n))$.

These results imply that group testing can provide drastic gains when $d \ll n$, say $d = O(\text{poly}(\log n))$, compared to the naive approach of testing every item individually, which results in $t = n$ total number of tests. However, it can be shown that these constructions require each item to participate in $w = \Omega\left(\frac{d \log n}{\log d + \log\log n}\right)$ tests and each test to include $\rho = \Omega\left(\frac{n}{d \log_d n}\right)$ items. In many applications, the total number of tests that can be performed on each item or the number of items each test can include can be limited due to different reasons. For example, the amount of blood or genetic material available from an individual can limit the number of tests that this individual can participate in. Similarly, equipment

Manuscript received January 25, 2019; revised August 3, 2019; accepted October 28, 2019. Date of publication November 5, 2019; date of current version April 21, 2020. This work was supported in part by NSF grant NeTS-1817205 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. This article was presented in part at Allerton 2017.

H. A. Inan and A. Özgür are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: hinan1@stanford.edu; aozgur@stanford.edu).

P. Kairouz is with Google AI, Mountain View, CA 94043 USA (e-mail: kairouz@google.com).

Communicated by P. Sadeghi, Associate Editor for Coding Techniques.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2951703

limitations and testing procedures can impose a maximum on the number of samples that can be simultaneously tested. For example, combining too many blood samples in one test often increases the misdetection probability of the targeted disease. Our interest in the group testing problem was mainly motivated by its applications in wireless communication, in particular the use of group testing codes for constructing random access schemes for the Internet of Things (IoT) networks, as we discuss in more detail in the last section of the paper. Here items represent sensor nodes and tests represent binary transmissions over a common channel, and the energy constraint at the physical layer can be translated to a constraint on the number of tests applied to each item in the group testing framework.

Motivated by these observations, in this paper we study the combinatorial group testing problem when the number of tests each item can participate in is restricted by $w_{\max}$ or when the number of items each test can include is restricted by $\rho_{\max}$. In the noiseless case, one can show that if $w_{\max} \leq d$ or if $\rho_{\max} \leq d+1$, then we need $t = n$ tests (c.f. Proposition 2 and Theorem 8), i.e., testing every item individually is optimal. A natural question then is: how does t decrease as we increase $w_{\max}$ and $\rho_{\max}$ beyond these bare minimums (up to their values in state-of-the-art constructions)? In particular, can we slightly increase $w_{\max}$ and $\rho_{\max}$ beyond d and $d+1$, respectively, and significantly reduce t, the number of tests needed? The answer turns out to be positive when we relax the constraint on $w_{\max}$, whereas the effect on t is less dramatic when the constraint on $\rho_{\max}$ is relaxed.

We show that when $w_{\max} = d+1$, the number of tests decreases drastically from $t = n$ to $t = (d+1)\sqrt{n}$. More generally, if $w_{\max} = ld+1$ for any positive integer l such that $ld+1 \leq n^{\frac{1}{l+1}}$, we can achieve

$$t = (ld+1)n^{\frac{1}{l+1}}.$$

This implies that the fractional power of n can be reduced drastically when $w_{\max}$ is increased as a multiple of d. Note that this result is most significant when $d = O(\text{poly}(\log n))$. We achieve this performance by introducing a simple modification of the Kautz-Singleton construction, which shows that the field size in this construction can be used to trade between t and w_{max}. We then prove a nearly matching lower bound, which shows that

$$t = \Omega\left(d^{\frac{2}{l+1}} n^{\frac{1}{l+1}}\right).$$

In particular when $w_{\max} = d+1$, this shows that the Kautz-Singleton construction is order-optimal (to an almost matching constant). This is somewhat surprising given that the Kautz-Singleton construction is strictly suboptimal in the classical group testing setting when $d = O(\text{poly}(\log n))$.

As opposed to the case with a constraint on $w_{\max}$, the reduction trend in t is much less dramatic with increasing ρ_{max}. We prove that for any $\rho_{\max}$,

$$t \geq \frac{(d+1)n}{\rho_{\max}},$$

in the noiseless case. In other words, $\rho_{\max}$ needs to be a fractional power of n, in order for t to scale sublinearly in n. When $\rho_{\max} = \Theta(n^{\alpha})$, we prove that there exists a construction with $t = \Theta(dn^{1-\alpha})$. We also show that when $\alpha = \frac{l}{l+1}$ for any fixed integer $l \geq 1$, the Kautz-Singleton construction can be used to achieve the order-optimal $t = (ld+1)n^{\frac{1}{l+1}}$ tests. For the special case $l = 1$, it is interesting to note that this construction matches the lower bound with matching constants, therefore, the Kautz-Singleton construction is exactly optimal. We further extend these results to the noisy case when a certain number of tests can have faulty outcomes. We also prove that our constructions can be decoded in $(\text{poly}(d)+O(t))$-time and each entry in any codeword can be computed in $\text{poly}(\log n)$ memory space. This shows that these constructions not only (nearly) achieve the fundamental lower bounds, but also have a favorable encoding and decoding complexity.

A. Comparison With Prior Work

To the best of our knowledge, our problem formulation has not been well explored in the combinatorial group testing literature. We discuss two particular prior works that are most relevant to this paper. The first one is a recent paper by Gandikota *et al.* [17], which focuses on the sparse group testing problem when the defective set can be recovered with high probability. In this framework, the defective set is chosen uniformly at random among all items and the exact identification of the defective set may fail with some arbitrarily small but positive $\epsilon > 0$ probability of error, where the error probability is taken over the randomness of the set of defectives. We refer readers interested in group testing designs with sparsity constraints under the small error criterion to [17]. We point out that we focus on the zero-error criterion and our approach is purely combinatorial, therefore, complementing the small error setting considered by Gandikota *et al.* [17].

An earlier work by Macula [18] presents an explicit construction with constant column and row weights under a zero-error reconstruction criterion. However, the construction in [18] is highly suboptimal as we discuss next.

When the weights of the columns are constrained with $w_{\max} = ld+1$ for some integer $l \geq 1$, we provide an explicit construction achieving $O\left(dn^{\frac{1}{l+1}}\right)$ scaling in the number of tests, whereas the explicit construction in [18] provides $O\left(dn^{\frac{1}{(ld+1)^{1/d}}}\right)$ scaling in the number of tests, which is significantly larger. When the weights of the rows are constrained with $\rho_{\max} = n^{\frac{l}{l+1}}$ for some integer $l \geq 1$, we provide an explicit construction that achieves $O\left(dn^{\frac{1}{l+1}}\right)$ scaling in the number of tests. On the other hand, the explicit construction in [18] provides $O\left(n^{\frac{1}{l+1}}(l+1)^d\right)$ scaling in the number of tests, which has an exponential term in d.

B. Paper Organization

The remainder of this paper is organized as follows. In Section II, we present the needed prerequisite material and describe two common combinatorial group testing constructions. The main results of our paper are formally presented and proved in Section III. In Section IV, we discuss the encoding and decoding complexities of the explicit constructions introduced in Section III. In Section V, we discuss an application of our results to wireless random access, which provided the

original motivation for our work. Finally, we conclude our paper in Section VI by noting a few interesting and nontrivial extensions.

II. Preliminaries

For a $t \times n$ binary matrix M, we use M_j to refer to its j'th column and M_{ij} to refer to its (i,j)'th entry. For an integer $m \geq 1$, we denote the set $\{1,\ldots,m\}$ by $[m]$. The support of a column M_j is denoted as $\text{supp}(M_j) \coloneqq \{i : M_{ij} = 1\}$. We say that a binary column M_i *covers* a binary column M_j if $M_i \vee M_j = M_i$, or equivalently $\text{supp}(M_j) \subseteq \text{supp}(M_i)$. The Hamming weight of a row or a column of M will be simply referred to as the *weight* of the row or column and w_j represents the weight of j'th column.

A. Non-Adaptive Combinatorial Group Testing

Our paper focuses on non-adaptive combinatorial group testing (CGT). Non-adaptive refers to the fact that the tests are designed and fixed a priori, in constrast to the adaptive case, where the tests are designed sequentially, meaning that the j^{th} test is a function of the outcomes of the $j-1$ previous tests. Combinatorial refers to the fact that we want our group testing schemes to recover the set of defective items with zero-error, in contrast to the probabilistic approach, which allows for a small probability of error. A non-adaptive CGT strategy can be represented by a $t \times n$ binary matrix M, where $M_{ij} = 1$ indicates that item j participates in test i. We will occasionally refer to M as a group testing code (or codebook) and its i^{th} column M_i as the i^{th} codeword. A necessary and sufficient condition for the design of a non-adaptive CGT strategy M is that of *separability*. A matrix M is *d-separable* if for any d-sparse vectors x_1 and x_2, $x_1 \neq x_2$, we have that $Mx_1 \neq Mx_2$. Unfortunately, the d-separability condition does not lead to tractable, explicit, and efficiently decodable constructions of M for an arbitrary value of n. To circumvent this issue, a stronger condition on M is needed. This condition is known as *d-disjunctiveness* [1]. We first revisit the definition of d-disjunctiveness [1].

Definition 1. *A $t \times n$ binary matrix M is called d-disjunct if any Boolean sum of up to d columns of M does not cover any other column not included in the sum.*

The d-disjunctiveness property ensures that we can recover up to d columns from their Boolean sum. This can be naively done using the *cover decoder*. The cover decoder simply scans through the columns of M, and checks whether or not the test results vector Y covers a particular column. If column i is covered by Y, then item i is declared defective. When M is d-disjunct, the cover decoder succeeds at identifying all the defective items, while achieving a zero false positive rate. Interestingly, one can also show that $(d+1)$-separability implies d-disjunctiveness [1]. Therefore, even though disjunctiveness is stronger than separability, the two conditions are essentially equivalent.

We define $t(d,n)$ to be the smallest t needed for a binary $t \times n$ matrix M to be d-disjunct. Notice that naturally, $t(d,n) \leq n$ because we can always use the identity matrix $M = I_n$ to identify any $1 \leq d \leq n$ defectives among n items. A classical result in the non-adaptive combinatorial group testing literature shows that $t(d,n) = \min\left\{n, \Omega(d^2 \log_d n)\right\}$ [14], [15]. Several explicit and randomized constructions of d-disjunct matrices have been developed over the past fifty years and the most efficient construction achieved $t = O(d^2 \log n)$ (when $d = O(\text{poly}(\log n))$) [1], [16], [19].

B. Relevant Lower Bounds

We now summarize two lower bounds introduced in the literature on the minimum number of tests. These bounds imply that individual testing is necessary whenever $d = \omega(n)$ or $w_{\max} \leq d$, where $w_{\max}$ is the maximum number of tests an item participates in (or equivalently, the maximum column weight).

Proposition 1. *For all n and d, the following bound on $t(d,n)$ holds*

$$t(d,n) \geq \min\left\{\binom{d+2}{2}, n\right\}. \tag{1}$$

Proposition 1 suggests that designing a d-disjunct matrix with $t < n$ is only possible when $d = O(\sqrt{n})$. Slightly overloading the notation, in the following proposition we denote $t(d,n,w_{\max})$ as the smallest t needed for a binary $t \times n$ matrix M to be d-disjunct with maximum column weight $w_{\max}$.

Proposition 2. *If $w_{\max} \leq d$, then $t(d,n,w_{\max}) = n$.*

The above proposition shows that one cannot do better than individual testing when the maximum number of tests an item can participate in is less than or equal to d.

The proofs of these propositions are due to D'yachkov and Rykov, and can be found in [14, Lemma 1, Remark 2], [1, Lemma 7.2.7, Theorem 7.2.9].

C. Disjunct Matrices via Error Correcting Codes

A q-ary error-correcting code is a code whose codewords consist of q basic symbols [20]. Binary codes are a special case of q-ary codes with $q = 2$. Consider a q-ary code with $n = q^k$ codewords of length $t = k+r$. Denoting the minimum distance between the codewords as $d_{\min}$, one can show that $d_{\min} \leq r+1$ from the following observation. Fix any k positions in the codewords. If any two codewords have the same symbols in these positions, then it must be the case that $d_{\min} \leq r$. Otherwise, we must observe all possible q^k sequences in the k fixed positions. In this case, some of the codewords will differ by only one position on the fixed k positions. Hence, $d_{\min} \leq r+1$. We state this formally in the following theorem [21].

Theorem 1. *A q-ary code with $n = q^k$ codewords of length $t = k + r$ must satisfy $d_{\min} \leq r + 1$.*

Codes with $d_{\min} = r+1$ and $n = q^k$ are called maximum distance separable (MDS) codes [21]. Reed-Solomon codes [22] are a known class of MDS codes with the constraint that $q \geq t$. When concatenated with a nonlinear code, Reed-Solomon codes lead to d-disjunct group testing codes. In what

follows, we will use the subscript q in the parameters of the Reed-Solomon codes to distinguish them from the parameters of the group testing codes that will be constructed shortly. To recap, Reed-Solomon codes achieve

$$d_{\min,q} = r_q + 1, \tag{2}$$

$$t_q = k_q + r_q, \tag{3}$$

$$n_q = q^{k_q}, \tag{4}$$

provided that $k_q \leq t_q \leq q$ and q is a prime power.

We can convert a Reed-Solomon code into a group testing code using the following method introduced by Kautz and Singleton in [9]. We replace each codeword symbol $i \in \{1, 2, \ldots, q\}$ by e_i, a length-q binary sequence with a single nonzero entry in the i^{th} position. Thus, a Reed-Solomon code is transformed into a binary code of length $t = qt_q$ by concatenating it with the "identity code". The minimum distance of the resultant binary code is double that of the Reed-Solomon code; i.e., $d_{\min} = 2\ d_{\min,q} = 2(r_q + 1)$. This is because any two distinct q-ary symbols will differ in two positions in their corresponding length-q binary sequences. Note that the number of codewords remains the same $n = n_q = q^{k_q}$. Since each symbol has a single nonzero entry in its binary expansion, the weight of the codewords are the same with $w = t_q$. Furthermore, since Reed-Solomon codes satisfy MDS property, any chosen $k_q \times q^{k_q}$ submatrix must include all q^{k_q} possible assignments of q-ary symbols in the columns. It follows that any row of a Reed-Solomon code must include every q-ary symbol an equal number of times. More precisely, each q-ary symbol must present q^{k_q-1} times in all rows. Therefore, the corresponding binary code has a constant row weight of $\rho = n/q$. This construction will be referred to as the Kautz-Singleton construction.

Consider a binary code M with minimum codeword weight of $w_{\min}$. We define $\lambda_{\max}$ to be the maximum number of overlapping ones between any two codewords in M. In the coding theory literature, $\lambda_{\max}$ is commonly referred to as the maximum overlap of M [9]. A central result in group testing demonstrates that M is d-disjunct as long as $\lambda_{\max} d + 1 \leq w_{\min}$. This can be seen from the following simple argument. Take any $d+1$ codewords and fix one codeword among them. The number of overlapping ones between the fixed codeword and the rest of the codewords is at most $d\lambda_{\max}$. Since the minimum weight satisfies $w_{\min} \geq d\lambda_{\max} + 1$, this codeword cannot be covered by the rest of the codewords. Thus, M must be at least d-disjunct. We state this formally in the following lemma.

Lemma 1. *A binary code M with codewords of minimum weight $w_{\min}$ and maximum overlap $\lambda_{\max}$ is $\left\lfloor \frac{w_{\min}-1}{\lambda_{\max}} \right\rfloor$-disjunct.*

Observe that in the Kautz-Singleton construction, we have

$$\lambda_{\max} = w - d_{\min}/2 = t_q - r_q - 1 = k_q - 1.$$

Therefore, the Kautz-Singleton construction provides us with a group testing code that is $\left\lfloor \frac{t_q-1}{k_q-1} \right\rfloor$-disjunct. In the following lemma, we summarize the properties of the Kautz-Singleton construction that has been discussed in this section.

Lemma 2. *Let q be a prime power and t_q and k_q be positive integers such that $k_q \leq t_q \leq q$. The Kautz-Singleton construction provides a binary $t \times n$ matrix that is d-disjunct with $t = qt_q$, $n = q^{k_q}$, and $d = \left\lfloor \frac{t_q-1}{k_q-1} \right\rfloor$. Furthermore, all columns have the same weight $w = t_q$ and all rows have the same weight $\rho = n/q$.*

In classical combinatorial group testing where there is no weight constraint, two famous constructions that provide the best achievability results are introduced by Kautz and Singleton in [9] and by Porat and Rothscheld in [16]. We next summarize the results of these constructions.

Theorem 2. *The Kautz-Singleton construction provides a $t \times n$ d-disjunct matrix where $t = O(d^2 \log_d^2\ n)$ with constant column weight $w = \Omega\left(\frac{d \log n}{\log d + \log\log n}\right)$ and constant row weight $\rho = \Omega\left(\frac{n}{d \log_d n}\right)$.*

Proof. To obtain a d-disjunct code using the Kautz-Singleton construction, let us set $t_q = q$, and choose q and k_q such that $d = \left\lfloor \frac{q-1}{k_q-1} \right\rfloor$. Note that $n = q^{k_q}$ and $q = \Theta(dk_q)$. Hence, $q = \Theta(d \log_q n)$ or $q \log q = \Theta(d \log n)$. Since $q \geq d$, we get that $q = O(d \log_d n)$. Note that $t = qt_q = q^2$, therefore $t = O(d^2 \log_d^2\ n)$. By Lemma 2, the corresponding binary code has constant column weights $w = t_q = q$ and constant row weights $\rho = n/q$. Note that $q \log q = \Theta(d \log n)$ is related to the famous Lambert W function [23] and using $W(x) \geq \log x - \log\log x$, we get $q = \Omega\left(\frac{d \log n}{\log(d \log n)}\right)$ or equivalently $w = \Omega\left(\frac{d \log n}{\log d + \log\log n}\right)$. Since $q = O(d \log_d n)$, it follows that $\rho = \Omega\left(\frac{n}{d \log_d n}\right)$. □

A different line of work introduced by Porat and Rothscheld in [16] constructs $t \times n$ d-disjunct matrices with $t = O(d^2 \log n)$. Their approach is based on q-ary codes that meet the Gilbert-Varshamov bound where the alphabet size is $q = \Theta(d)$. As in the Kautz-Singleton construction, their inner code is the identity code. The resulting binary code has the property that all the columns have the same weight of $w = \Theta(d \log n)$. Furthermore, the maximum row weight satisfies $\rho_{\max} = \Omega(n/d)$.

Theorem 3. *The explicit construction by Porat and Rothscheld in [16] achieves a $t \times n$ d-disjunct matrix where $t = O(d^2 \log n)$ with constant column weight $w = \Theta(d \log n)$ and maximum row weight $\rho_{\max} = \Omega(n/d)$.*

Compared to the Kautz-Singleton construction, one can observe that the Porat and Rothscheld's construction performs better in terms of scaling in the number of tests in the regime where $d = O(\text{poly}(\log n))$. However, the Kautz-Singleton construction performs better in terms of scaling in the number of tests when $d = \Theta(n^\alpha)$ for some constant $\alpha \in (0, 1/2)$. In this regime, the Kautz-Singleton construction meets the fundamental lower bound and is therefore order-optimal.

D. Noisy Test Outcomes

We have so far discussed the setting in which the test outcomes are always correct, i.e., there is no noise in the

measurement process. However, in many practical applications such as drug discovery and DNA library screening, testing errors are present [1], [24]. Naturally, the aforementioned definitions and techniques can be extended so that one can identify the defective items even with certain number of faulty test outcomes. The following definition extends the notion of disjunctiveness in the presence of error in the measurement process [1].

Definition 2. *A $t\times n$ binary matrix M is called (d,ν)-disjunct (ν-error detecting d-disjunct) if $|\text{supp}(M_i)\backslash\cup_{j\in S}\text{supp}(M_j)| > \nu$ for every set S of columns with $|S|\leq d$ and every $i\in[n]\backslash S$.*

We note that $(d,0)$-disjunct is simply d-disjunct and a (d,ν)-disjunct matrix can detect up to ν errors and can correct up to $\lfloor\nu/2\rfloor$ errors in the test measurements. The latter can be done by simply modifying the cover decoder to incorporate the noise as follows.

Proposition 3. *Let the cover decoder scan through the columns of M and eliminate all items belonging to at least $\lceil\nu/2\rceil+1$ negative tests and return the remaining items. The cover decoder correctly identifies all defective items without any error if M is (d,ν)-disjunct in the case when the test outcomes have up to $\lfloor\nu/2\rfloor$ errors.*

We similarly define $t(d,\nu,n)$ to be the smallest t needed for a binary $t\times n$ matrix M to be (d,ν)-disjunct.

III. Main Results

In this section, we formally present our results for both the sparse codeword and the sparse test settings. We begin with the sparse codeword setting.

A. Sparse Codeword

In the sparse codeword setting, we focus on a model where each item can participate in a limited number of tests. This is equivalent to restricting the codewords (columns of M) to have a limited number of "1"s. For ease of presentation, we begin with the noiseless case and then extend our results to the more general noisy case in what follows.

We recall from Proposition 2 that if the codewords have a Hamming weight that is bounded by d, one cannot do better than the identity matrix; i.e., $t=n$. Hence, we are interested in the regime where $w_{\max}>d$.

Given that it is impossible to achieve $t<n$ when $w_{\max}\leq d$, it is natural to ask: what happens when $w_{\max}=d+1$? We dedicate the following section answering this question.

1) The Case $w_{\max}=d+1$: We recall from Lemma 2 that the Kautz-Singleton construction provides a constant column weight ($w=t_q$) group testing code that is $\left\lfloor\frac{t_q-1}{k_q-1}\right\rfloor$-disjunct. By choosing $k_q=2$ and $t_q=d+1$ we get a d-disjunct matrix with $t=(d+1)\sqrt{n}$ tests and $w=d+1$ column weights when $q\geq t_q$ is satisfied. Therefore the natural question is how good this construction is in terms of the required number of tests for a d-disjunct matrix with $w_{\max}\leq d+1$. The following theorem presents our first result answering this question.

Theorem 4. *For all integers $d,n\geq 2$ such that $d+1\leq\sqrt{n}$ and $\sqrt{n}$ is a prime power, the Kautz-Singleton construction provides a $t\times n$ matrix that is d-disjunct with constant column weight $w=d+1$ and*

$$t=(d+1)\sqrt{n}.$$

On the other hand, for all integers $d,n\geq 2$, a $t\times n$ matrix that is d-disjunct with maximum column weight $w_{\max}\leq d+1$ must satisfy

$$t\geq\min\left\{\sqrt{nd(d+1)},n\right\}.$$

Proof. We begin with the lower bound. Let M be a $t\times n$ d-disjunct matrix with $w_{\max}\leq d+1$. We will separate the columns of M into two groups $\mathcal{N}_1,\mathcal{N}_2\subseteq[n]$ such that $\mathcal{N}_1\cup\mathcal{N}_2=[n]$ and $\mathcal{N}_1\cap\mathcal{N}_2=\emptyset$. We define a row $i\in[t]$ to be private for a column j, if j is the only column having row i in its support. If a column M_j has weight at most d, then it must have at least one private row, otherwise we can find at most d columns such that their union will span M_j, which contradicts with d-disjunctiveness. Now consider all columns that have weight equal to $d+1$. It is possible that some of them also have private rows. Hence, we construct the first set $\mathcal{N}_1$ consisting of two types of columns. The first one is the columns whose weight is less than or equal to d. The second one is the columns that have weight equal to $d+1$ and they have at least one private row. The second set $\mathcal{N}_2$ consists of the rest of the columns; i.e., the ones that have weight equal to $d+1$ and do not have any private row. Defining w_j to be weight of the column j for $1\leq j\leq n$, more formally we define

$$\begin{aligned}\mathcal{N}_1 &\coloneqq \{j\in[n]\mid w_j\leq d \text{ or}\\ &\qquad w_j=d+1 \text{ and } M_j \text{ has at least one private row}\},\\ \mathcal{N}_2 &\coloneqq \{j\in[n]\mid w_j=d+1 \text{ and } M_j \text{ has no private row}\}.\end{aligned}$$

Note that by construction, $\mathcal{N}_1\cup\mathcal{N}_2=[n]$ and $\mathcal{N}_1\cap\mathcal{N}_2=\emptyset$, hence $n=|\mathcal{N}_1|+|\mathcal{N}_2|$. In the following, we will bound the size of both sets $\mathcal{N}_1$ and $\mathcal{N}_2$.

We note that each column in the set $\mathcal{N}_1$ has at least one private row and by definition of the private row it cannot be shared by two distinct columns. Since there could be at most t private rows, we have $|\mathcal{N}_1|\leq t$.

We next consider the set $\mathcal{N}_2$. We generalize the definition of the private row to the private set as follows. A private set for a column is defined as a subset of its support such that no other column has this subset in its support. In other words, no other column has all ones in these positions. We claim that for any column in the set $\mathcal{N}_2$, all the subsets of its support with size 2 (i.e., all pairs of positions in its support) are private. We prove this by contradiction. Assume there exists a column in the set $\mathcal{N}_2$ such that it has a subset of size 2 in its support that is not private. This means that there exists another column that can cover the two positions in this subset. Note that any column in the set $\mathcal{N}_2$ has weight $d+1$ and has no private row, therefore, there are $d-1$ positions in the support except this pair and we can find at most $d-1$ columns that can cover all these positions. It follows that we can find at most d columns that can cover all $d+1$ positions in the support of this column, which contradicts with the d-disjunctiveness. Note that there

are $\binom{d+1}{2}$ number of pairs in the support of any column in $\mathcal{N}_2$ and by definition of private set it cannot be shared by two distinct columns. We further note that each column in the set $\mathcal{N}_1$ will have a private row and it must be the case that the columns in the set $\mathcal{N}_2$ must have a zero in these rows, therefore, there could be at most $\binom{t-|\mathcal{N}_1|}{2}$ number of private pairs. Hence, we have

$$|\mathcal{N}_2|\binom{d+1}{2} \leq \binom{t-|\mathcal{N}_1|}{2}.$$

Therefore,

$$|\mathcal{N}_2|d(d+1) \leq (t-|\mathcal{N}_1|)(t-|\mathcal{N}_1|-1) \leq (t-|\mathcal{N}_1|)^2.$$

Defining $n_1 \triangleq |\mathcal{N}_1|$, this gives us

$$t \geq n_1 + \sqrt{(n-n_1)d(d+1)}. \tag{5}$$

Note that $0 \leq n_1 \leq t \leq n$. One can take the second derivative of the right-hand side of (5) and observe that it is negative for $0 \leq n_1 \leq t \leq n$, which means it is a concave function of n_1 and it will be minimum at either $n_1 = 0$ or $n_1 = t$. Therefore,

$$t \geq \min\left\{\sqrt{nd(d+1)}, t + \sqrt{(n-t)d(d+1)}\right\}.$$

Noting that $t \geq t + \sqrt{(n-t)d(d+1)}$ only when $t = n$, one can observe that

$$t \geq \min\left\{\sqrt{nd(d+1)}, n\right\}.$$

For the achievability, we use the Kautz-Singleton construction in Lemma 2. We choose $w = t_q = d+1$ and $k_q = 2$, therefore, we have a d-disjunct matrix. Since $n = q^{k_q}$, we obtain $q = \sqrt{n}$ and therefore $t = (d+1)\sqrt{n}$. In order to satisfy the requirement $q \geq t_q$ where q is any prime power, we must ensure that $d+1 \leq \sqrt{n}$ and $q = \sqrt{n}$ must be a prime power. This completes the proof for the achievability. □

A few comments are in order. First, Theorem 4 shows that by increasing $w_{\max}$ from d to $d+1$, we suddenly get $t = \Theta(d\sqrt{n})$ instead of $t = n$. Second, the achievability result in Theorem 4 is obtained by changing the field size from $q = O(d \log_d n)$ to $q = \sqrt{n}$ in the Kautz-Singleton construction. The Kautz-Singleton construction is strictly suboptimal in the classical case when $d = O(\text{poly}(\log n))$. It is interesting that a simple modification of this well known construction makes it optimal in this case (even up to an almost matching constant).

We next investigate the more general case where the codeword weights are bounded by $w_{\max} \leq ld+1$ for some integer $l > 1$.

2) The General Case $w = ld+1$: We note that by choosing $k_q = l+1$ and $t_q = ld+1$ in Lemma 2, we get a d-disjunct matrix with $t = (ld+1)n^{\frac{1}{l+1}}$ tests and $w = ld+1$ column weights using the Kautz-Singleton construction when $q \geq t_q$ is satisfied. In this case we can show that this construction is nearly optimal.

Theorem 5. *For all integers $d, n, l \geq 2$ such that $ld + 1 \leq n^{\frac{1}{l+1}}$ and $n^{\frac{1}{l+1}}$ is a prime power, the Kautz-Singleton construction provides a $t \times n$ matrix that is d-disjunct with constant column weights $w = ld+1$ and*

$$t = (ld+1)n^{\frac{1}{l+1}}.$$

On the other hand, for all integers $d, n, l \geq 2$, a $t \times n$ matrix that is d-disjunct with maximum column weight $w_{\max} \leq ld+1$ must satisfy

$$t \geq \left(\frac{(l-1)^{l+1}(d-1)^{l+1}}{2\ e^l(l-1)(d-1)^{l-1}+1}\right)^{\frac{1}{l+1}} n^{\frac{1}{l+1}}.$$

Proof. We begin with the lower bound. Let M be a $t \times n$ d-disjunct matrix with $w_{\max} \leq ld+1$. We similarly separate the columns of M into $l+1$ groups and construct the sets $\mathcal{N}_i$ for $i = 1, \ldots, l+1$ such that $\mathcal{N}_1 \cup \ldots \cup \mathcal{N}_{l+1} = [n]$ and $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$ for any $i, j \in [l+1]$ with $i \neq j$. We construct the sets $\mathcal{N}_1, \ldots, \mathcal{N}_{l+1}$ as follows. We keep the first set $\mathcal{N}_1$ as the columns whose weight is less than or equal to d, as well as the ones that have weight equal to $d+1$ such that they have at least one private row. For $i = 2, \ldots, l$, the set $\mathcal{N}_i$ consists of two types of columns. The first one is the columns that have weight between $(i-2)d+2$ and $(i-1)d+1$ and they have no private set of size $i-1$. The second one is the columns that have weight between $(i-1)d+2$ and $id+1$ and they have at least one private set of size i. Finally, the last set $\mathcal{N}_{l+1}$ consists of the columns that have weight between $(l-1)d+2$ and $ld+1$ and they have no private set of size l. More formally,

$$\begin{aligned}
\mathcal{N}_1 \coloneqq \{&j \in [n] \mid w_j \leq d \text{ or} \\
&w_j = d+1 \text{ and } M_j \text{ has at least one private row}\}, \\
\text{for } i = 2, \ldots, l,& \\
\mathcal{N}_i \coloneqq \{&j \in [n] \mid (i-2)d+2 \leq w_j \leq (i-1)d+1 \text{ and } M_j \\
&\text{has no private set of size } i-1 \text{ or} \\
&(i-1)d+2 \leq w_j \leq id+1 \text{ and } M_j \text{ has at least one} \\
&\text{private set of size } i\}, \\
\mathcal{N}_{l+1} \coloneqq \{&j \in [n] \mid (l-1)d+2 \leq w_j \leq ld+1 \text{ and } M_j \text{ has} \\
&\text{no private set of size } l\}.
\end{aligned}$$

Note that by construction, $\mathcal{N}_1 \cup \ldots \cup \mathcal{N}_{l+1} = [n]$ and $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$ for any $i, j \in [l+1]$ such that $i \neq j$, hence $n = |\mathcal{N}_1| + \ldots + |\mathcal{N}_{l+1}|$. In the following, we will bound the size of these sets.

Since there could be at most t private rows, we have $|\mathcal{N}_1| \leq t$. Consider the sets $\mathcal{N}_i$ for $i = 2, \ldots, l$. For any column $j \in \mathcal{N}_i$, if we have $(i-1)d+2 \leq w_j \leq id+1$, then by construction M_j has at least one private set of size i. For the case $(i-2)d+2 \leq w_j \leq (i-1)d+1$, we claim that all the subsets of its support with size i must be private for the column M_j. We similarly show this by contradiction. Assume that M_j has a subset of size i in its support that is not private. Then we can find a column that can cover these positions. By the construction of set $\mathcal{N}_i$, the column M_j has no private set of size $i-1$, therefore, one can find at most $((i-1)d+1-i)/(i-1) = d-1$ columns that will cover the rest of the positions in the support of M_j. Hence, we have at most d columns covering the column M_j, which contradicts the d-disjunctiveness. Therefore, we obtain that all the columns in the set $\mathcal{N}_i$ must have at least one private set of size i. Since the private sets cannot be shared among the columns and we have at most $\binom{t}{i}$ private sets of size i, it yields $|\mathcal{N}_i| \leq \binom{t}{i}$.

For the last set $\mathcal{N}_{l+1}$, similar arguments apply and for each column, it should be the case that all the subsets of its support with size $l+1$ must be private. Since $w_j \geq (l-1)d+2$ for $j \in \mathcal{N}_{l+1}$, we have $|\mathcal{N}_{l+1}|\binom{(l-1)d+2}{l+1} \leq \binom{t}{l+1}$. Therefore,

$$\begin{aligned} n &= |\mathcal{N}_1| + \ldots + |\mathcal{N}_{l+1}| \\ &\leq \sum_{i=1}^{l} \binom{t}{i} + \frac{\binom{t}{l+1}}{\binom{(l-1)d+2}{l+1}} \\ &\overset{(i)}{\leq} \left(\frac{et}{l}\right)^l + \frac{t \ldots (t-l)}{((l-1)d+2)\ldots((l-1)(d-1)+1)} \\ &\overset{(ii)}{\leq} \frac{e^l t^l}{l^l} + \frac{t^{l+1}}{((l-1)(d-1))^{l+1}} \\ &\overset{(iii)}{\leq} \frac{e^l t^l}{(l-1)^l} \frac{t}{(d-1)^2/2} + \frac{t^{l+1}}{((l-1)(d-1))^{l+1}} \\ &= t^{l+1}\left(\frac{2\ e^l}{(l-1)^l(d-1)^2} + \frac{1}{(l-1)^{l+1}(d-1)^{l+1}}\right) \\ &= t^{l+1}\left(\frac{2\ e^l(l-1)(d-1)^{l-1}+1}{(l-1)^{l+1}(d-1)^{l+1}}\right), \end{aligned}$$

where (i) is due to the inequality $\sum_{i=0}^{l}\binom{t}{i} \leq \left(\frac{et}{l}\right)^l$ for $t \geq l \geq 1$, (ii) is bounding all the terms in the numerator by t and denominator by $(l-1)(d-1)$ and in (iii) we use (1) and $\binom{d+2}{2} \geq \frac{(d-1)^2}{2}$. This completes the proof for the lower bound.

For the achievability, we use the Kautz-Singleton construction. We choose $w = t_q = ld+1$ and $k_q = l+1$ in Lemma 2 to get a d-disjunct matrix. Since $n = q^{k_q}$, we obtain $q = n^{\frac{1}{l+1}}$ and therefore $t = (ld+1)n^{\frac{1}{l+1}}$. In order to satisfy the requirement $q \geq t_q$ where q is any prime power, we must ensure that $ld+1 \leq n^{\frac{1}{l+1}}$ and $q = n^{\frac{1}{l+1}}$ must be a prime power. This completes the proof for the achievability. □

Note that as we increase the weights as a multiple of d (i.e., $w_{\max} = ld+1$), the minimum number of required tests decreases exponentially in l. As we see from Theorem 5, for a fixed l the lower bound we get is $\Theta\left(d^{\frac{2}{l+1}}n^{\frac{1}{l+1}}\right)$ whereas the upper bound is $\Theta(dn^{\frac{1}{l+1}})$. While we have a matching lower and upper bound in terms of the scaling with respect to n, there is an increasing gap of $d^{\frac{l-1}{l+1}}$ between them, which approaches d for large l. In the extreme case where we can choose l such that $ld+1 = n^{\frac{1}{l+1}}$, we get $l = O\left(\frac{\log n}{\log\log n + \log d}\right)$ and $t = O(d^2\log_d^2\ n)$, which captures the original Kautz-Singleton construction.

Before presenting our results in the noisy setting, we discuss the assumptions we have in our above results. Obviously the assumption that $n^{\frac{1}{l+1}}$ is a prime power may not hold for all positive integers n. However, one can always pad the items with "dummy" non-defectives until we get a prime power, which can only change the constants in our results and does not affect the scaling. On the other hand, the assumption of $ld+1 \leq n^{\frac{1}{l+1}}$ may seem somewhat more restrictive. We point out that the results for the classical group testing as well as our results are most significant when $d = O(\text{poly}(\log n))$. In this regime, the assumption $ld+1 \leq n^{\frac{1}{l+1}}$ is asymptotically valid for all integers l and all of our bounds match within logarithmic factors. For the regime where $d = \Theta(n^\alpha)$ for some constant $\alpha \in (0, 1/2)$, there are cases where $d > n^{\frac{1}{l+1}}$, violating our assumption. However, note from Theorem 2 that, the original Kautz-Singleton construction already achieves the order-optimal number of tests with $w = O(d)$ column weight in this case. We know from Proposition 2 that $w_{\max} \leq d$ requires $t = n$, therefore, in this case we can trivially achieve the order-optimal number of test with minimal column weight.

We continue our discussion with the noisy case.

3) The Noisy Setting: As we have seen that it is impossible to achieve $t < n$ when $w_{\max} \leq d$ in the noiseless case, a similar result can be observed in the noisy case as well. Our next result extends this to the noisy setting with an arbitrary noise parameter ν.

Proposition 4. *If $w_{\max} \leq d+\nu$, then $t(d,\nu,n) = (\nu+1)n$.*

The proof of the above proposition can be found in Appendix A. Proposition 4 similarly shows that one cannot do better than individual testing corresponding to the more general noisy setting if the codeword weights are bounded by $d+\nu$.

We note that it is sufficient to have $w_{\min} \geq d\lambda_{\max} + \nu + 1$ to obtain a (d,ν)-disjunct matrix. We can employ the Kautz-Singleton construction and fix $k_q = 2$ and $t_q = d+\nu+1$ to get a (d,ν)-disjunct matrix with $t = (d+\nu+1)\sqrt{n}$ tests and $w = d+\nu+1$ column weights when $q \geq t_q$ is satisfied. The following theorem shows that this is order-optimal when $w_{\max} \leq d+\nu+1$.

Theorem 6. *For all integers $d, n \geq 2$ and $\nu \geq 0$, a $t \times n$ matrix that is (d,ν)-disjunct with maximum column weight $w_{\max} \leq d+\nu+1$ must satisfy*

$$t \geq \min\left\{(\nu+1)n, \sqrt{(d+\nu)(d+\nu+1)n}\right\}.$$

The proof of the above theorem can be found in Appendix B. It is interesting to observe that by increasing $w_{\max}$ from $d+\nu$ to $d+\nu+1$, we are able to reduce to $t = \Theta((d+\nu)\sqrt{n})$ from $t = (\nu+1)n$. Going further, we can generalize this to the case where the codeword weights are bounded by $w_{\max} \leq ld+\nu+1$ for some integer $l > 1$. Fixing $k_q = l+1$ and $t_q = ld+\nu+1$ in the Kautz-Singleton construction provides us with a (d,ν)-disjunct matrix that has $t = (ld+\nu+1)n^{\frac{1}{l+1}}$ tests and $w = ld+\nu+1$ column weights. The next theorem shows that this construction is nearly optimal.

Theorem 7. *For all integers $d, n, l \geq 2$ and $\nu \geq 0$, a $t \times n$ matrix that is (d,ν)-disjunct with maximum column weight $w_{\max} \leq ld+\nu+1$ must satisfy*

$$t \geq n^{\frac{1}{l+1}}\left(\frac{2\ e^l}{(d+\nu)^2(l-1)^l} + \frac{1}{((l-1)(d-1)+\nu)^{l+1}}\right)^{-\frac{1}{l+1}}$$

The proof of the above theorem can be found in Appendix C. Similar to the noiseless case, we have a matching lower and upper bounds in terms of the scaling with respect to n and order-wise the lower bound in Theorem 7 is $\Theta\left((d+\nu)^{\frac{2}{l+1}}n^{\frac{1}{l+1}}\right)$ whereas the Kautz-Singleton construction provides $\Theta\left((d+\nu)n^{\frac{1}{l+1}}\right)$ tests.

B. Sparse Test

In the sparse test setting, we focus on a model where each test can include a limited number of items. In other words, we restrict the row weights of M, and derive lower and upper bounds on the minimum number of tests so that M is a (d,ν)-disjunct matrix in the more general noisy case (including the noiseless setting as a special case under $\nu=0$).

Our first theorem provides a fundamental lower bound on the minimum required number of tests under a row weight constraint and an upper bound, which is again based on the Kautz-Singleton construction.

Theorem 8. *For all integers $d,n\geq 2$ and $\nu,l\geq 0$ such that $ld+\nu+1\leq n^{\frac{1}{l+1}}$ and $n^{\frac{1}{l+1}}$ is a prime power, the Kautz-Singleton construction provides a $t\times n$ matrix that is (d,ν)-disjunct with constant row weights $\rho=n^{\frac{l}{l+1}}$ and*

$$t=(ld+\nu+1)n^{\frac{1}{l+1}}.$$

On the other hand, for all integers $d,n\geq 2$ and $\nu\geq 0$, a $t\times n$ matrix that is (d,ν)-disjunct with maximum row weight $\rho_{\max}$ must satisfy

$$t\geq\begin{cases}\dfrac{(d+\nu+1)n}{\rho_{\max}} & \text{if } \rho_{\max}>\dfrac{d+\nu+1}{\nu+1},\\ (\nu+1)n & \text{if } \rho_{\max}\leq\dfrac{d+\nu+1}{\nu+1}.\end{cases}$$

Proof. We begin with the lower bound. Suppose we have a $t\times n$ matrix that is (d,ν)-disjunct and all the row weights are bounded with $\rho_{\max}$. We consider the columns that have weight less than or equal to $d+\nu$. All these columns must have at least $\nu+1$ private rows as we have the condition $|\text{supp}(M_i)\setminus\cup_{j\in S}\text{supp}(M_j)|>\nu$ for every set S of columns with $|S|\leq d$ and every $i\in[n]\setminus S$ in Definition 2. Let us remove these columns along with the corresponding $\nu+1$ private rows for each column. Note that if a column has more than $\nu+1$ private rows, we can arbitrarily choose and remove any $\nu+1$ of them. We will be removing $\nu+1$ private rows per such column and this is possible since they all have at least $\nu+1$ private rows and a private row cannot be shared by two distinct columns.

Let us denote t_1 by the number of columns whose weight is less than or equal to $d+\nu$. From the definition of private row, it follows that $0\leq t_1\leq t/(\nu+1)$. After the removal operation, the dimension of the resulting matrix is $(t-t_1(\nu+1))\times(n-t_1)$ and it is still (d,ν)-disjunct since we are only removing the zero-entries of the rest of the columns, therefore, the resulting matrix must still satisfy the (d,ν)-disjunctiveness. We also note that the weight of the rows are still bounded with $\rho_{\max}$.

We observe that in the resulting matrix, the weight of all columns will be at least $d+\nu+1$. Therefore, the total number of ones in the resulting matrix can be lower bounded by $(d+\nu+1)(n-t_1)$ and upper bounded by $\rho_{\max}(t-t_1(\nu+1))$. Hence,

$$\begin{aligned}&\rho_{\max}(t-t_1(\nu+1))\geq(d+\nu+1)(n-t_1),\\ &\rho_{\max}t\geq(d+\nu+1)n+t_1\left(\rho_{\max}(\nu+1)-(d+\nu+1)\right).\end{aligned}\tag{6}$$

If $\rho_{\max}\leq\frac{d+\nu+1}{\nu+1}$, then from $t_1\leq t/(\nu+1)$ and (6) we have

$$\rho_{\max}t\geq(d+\nu+1)n+\frac{t}{\nu+1}\left(\rho_{\max}(\nu+1)-(d+\nu+1)\right)$$

It follows that $t\geq(\nu+1)n$. On the other hand, if $\rho_{\max}>\frac{d+\nu+1}{\nu+1}$, then from $t_1\geq 0$ and (6) we have

$$\rho_{\max}t\geq(d+\nu+1)n.$$

This yields that $t\geq\frac{(d+\nu+1)n}{\rho_{\max}}$.

For the achievability, we use the Kautz-Singleton construction. We choose $t_q=ld+\nu+1$ and $k_q=l+1$ to get a (d,ν)-disjunct matrix. Since $n=q^{k_q}$, we obtain $q=n^{\frac{1}{l+1}}$ hence $t=(ld+\nu+1)n^{\frac{1}{l+1}}$ and $\rho=n^{\frac{l}{l+1}}$ from Lemma 2. In order to satisfy the requirement $q\geq t_q$ where q is any prime power, we must ensure that $ld+\nu+1\leq n^{\frac{1}{l+1}}$ and $q=n^{\frac{1}{l+1}}$ must be a prime power. This completes the proof for the achievability. □

Observe that for any fixed integer $l\geq 1$ that satisfies the conditions stated in Theorem 8, the number of tests we get using the Kautz-Singleton construction is $\Theta\left((d+\nu)n^{\frac{1}{l+1}}\right)$ with constant row weight $\rho=n^{\frac{l}{l+1}}$. Substituting this in the lower bound of Theorem 8, the required number of tests is also $\Theta\left((d+\nu)n^{\frac{1}{l+1}}\right)$. It is interesting to note that the Kautz-Singleton construction is order-optimal in this setting.

From Proposition 2, we know that if the weights of the columns are bounded by d, one cannot do better than the individual testing; i.e., $t=n$. Theorem 8 states an analogous result for the case with row weight constraint: if the weights of the rows are bounded by $\frac{d+\nu+1}{\nu+1}$, we have $t=(\nu+1)n$, which means that we cannot do better than the individual testing. Another very interesting result that can be obtained from Theorem 8 is that for the special case where $l=1$, the Kautz-Singleton construction is optimal with matching constants.

Corollary 1. *For all integers $d,n\geq 2$ and $\nu\geq 0$, the Kautz-Singleton construction provides an optimal (with matching constants) (d,ν)-disjunct matrix under the maximum row weight constraint $\rho_{\max}\leq\sqrt{n}$.*

We emphasize that the Kautz-Singleton construction in Theorem 8 provides us with codes that have constant row weight of $n^{\frac{l}{l+1}}$; i.e., when ρ is a fractional power of n in the form $\frac{l}{l+1}$ in the interval $[1/2,1)$. It is natural to ask whether there exist group testing codes with $\rho=n^{\alpha}$ for an arbitrary $\alpha\in(0,1)$ that achieves the lower bound in Theorem 8. The following theorem shows the existence of such codes when $d=O(\text{poly}(\log n))$ by using a random construction.

Theorem 9. *For any $\alpha\in(0,1)$, there exists a randomized design that is (d,ν)-disjunct with $t=O\left((d+\nu)n^{1-\alpha}\right)$ number of tests and $\rho_{\max}=\Theta(n^{\alpha})$ maximum row weight in the regime where $d=O(\text{poly}(\log n))$.*

Proof. We consider the following randomized design. For a fixed $\alpha\in(0,1)$, we take $t=c(d+\nu)n^{1-\alpha}$ for some constant $c>0$ that we will fix later. We create a matrix M with size $t\times n$ by choosing the columns of this matrix uniformly at

random among the codewords of size t with weight w where we set $w = c(d+\nu)$.

We next calculate the probability of not having a (d,ν)-disjunct matrix. Let us fix $d+1$ columns of the matrix M and denote them as $M_1,\ldots,M_{d+1}$. Let us further fix a single column among them, say M_{d+1}. The probability of violating the condition $|\text{supp}(M_{d+1})\backslash \cup_{j\in S}\text{supp}(M_j)| > \nu$ where $S = \{1,2\ldots,d\}$ can be bounded as

$$\begin{aligned}\mathbb{P}\left(|\text{supp}(M_{d+1})\backslash \cup_{j\in S}\text{supp}(M_j)| \le \nu\right) &\overset{(i)}{\le} \frac{\binom{|\cup_{j\in S}\text{supp}(M_j)|}{w-\nu}\binom{t-w+\nu}{\nu}}{\binom{t}{w}} \\ &\overset{(ii)}{\le} \frac{\binom{dw}{w-\nu}\binom{t-w+\nu}{\nu}}{\binom{t}{w}},\end{aligned}$$

where in (i), we select $w-\nu$ positions from $\cup_{j\in S}\text{supp}(M_j)$ and then we select ν positions from the remaining $t-(w-\nu)$ positions in the numerator. The denominator is for all possible choices of weight w. This ensures that $\text{supp}(M_{d+1})$ intersects in at least $w-\nu$ positions with $\cup_{j\in S}\text{supp}(M_j)$. Note that certain combinations are counted more than once but that's not a problem since we are computing an upper bound. On the other hand, (ii) is because $M_1,\ldots,M_d$ can have at most dw non-intersecting number of ones.

Using the union bound over the choice of $d+1$ columns and fixing one among them, the probability that the matrix M does not satisfy the (d,ν)-disjunctiveness property can be bounded as

$$\mathbb{P}\left(M \text{ is not } (d,\nu)\text{-disjunct}\right) \le (d+1)\binom{n}{d+1}\frac{\binom{dw}{w-\nu}\binom{t-w+\nu}{\nu}}{\binom{t}{w}}.$$

We can further bound this as

$$\begin{aligned}&\mathbb{P}\left(M \text{ is not } d\text{-disjunct}\right) \qquad (7)\\ &\overset{(i)}{\le} (d+1)\left(\frac{ne}{d+1}\right)^{d+1}\frac{\left(\frac{dwe}{w-\nu}\right)^{w-\nu}\left(\frac{(t-w+\nu)e}{\nu}\right)^{\nu}}{\left(\frac{t}{w}\right)^{w}}\\ &= (d+1)\left(\frac{ne}{d+1}\right)^{d+1}\frac{\left(\frac{dw}{w-\nu}\right)^{w-\nu}\left(\frac{t-w+\nu}{\nu}\right)^{\nu}e^{w}}{n^{w(1-\alpha)}}\\ &\overset{(ii)}{\le} \frac{n^{d+1}e^{d+w+1}}{d^{d}}\cdot\frac{(2d)^{w-\nu}(t/\nu)^{\nu}}{n^{w(1-\alpha)}}, \qquad (8)\end{aligned}$$

where (i) is due to the inequality $\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{ne}{k}\right)^k$ and (ii) is due to $w-\nu \ge \nu/2$ for $c \ge 2$ and $t-w+\nu \le t$ and $d+1 \ge d$. Taking the logarithm of the last term in (8) gives us

$$\begin{aligned}&(d+1)\log n + d + w + 1 + (w-\nu)\log 2 + \nu\log(t/\nu)\\ &\quad + (w-\nu-d)\log d - w(1-\alpha)\log n \le -\tilde{c}(d+\nu)\log n,\end{aligned}$$

which holds with a constant $\tilde{c} > 0$ for sufficiently large n and appropriately chosen constant $c > 4/(1-\alpha)$ when $d = O(\text{poly}(\log n))$. Hence,

$$\mathbb{P}\left(M \text{ is not } (d,\nu)\text{-disjunct}\right) \le n^{-\tilde{c}(d+\nu)}. \qquad (9)$$

We next investigate the weights of the rows of the matrix M. We consider the first row. Note that by our random construction, it follows that each entry in the first row is independent and identically distributed with Bernoulli distribution where the probability of having one is $\frac{\binom{t-1}{w-1}}{\binom{t}{w}} = \frac{w}{t}$. Denoting ρ_1 as the weight of the first row, we have $\mathbb{E}[\rho_1] = \frac{w}{t}n = n^{\alpha}$. Using Hoeffding's inequality along with the union bound, we achieve the following upper bound on the probability that there exists a row with its weight deviating from n^{α}

$$\begin{aligned}\mathbb{P}(\exists\ i\in[t] \text{ s.t. } |\rho_i - n^{\alpha}| \ge \delta n^{\alpha}) &\le t2e^{-2n^{\alpha}\delta^2}\\ &= 2c(d+\nu)n^{1-\alpha}e^{-2n^{\alpha}\delta^2},\end{aligned} \qquad (10)$$

for some fixed constant $0 < \delta < 1$. For sufficiently large n, the right-hand side of (10) can be bounded as $e^{-\bar{c}n^{\alpha}}$ for some constant $\bar{c} > 0$.

Using the union bound over (9) and (10) to bound the probability of either not having a (d,ν)-disjunct matrix or not satisfying the $\rho_{\max} = \Theta(n^{\alpha})$ condition, we conclude that with probability approaching to 1 the matrix is (d,ν)-disjunct with row weight $\rho = \Theta(n^{\alpha})$ and $t = \Theta((d+\nu)n^{1-\alpha})$. □

In the regime where $d = O(\text{poly}(\log n))$, the lower bound in Theorem 8 suggests that the minimum number of tests is $\Omega((d+\nu)n^{1-\alpha})$ when $\rho = \Theta(n^{\alpha})$ for some $\alpha \in (0,1)$. The randomized construction in Theorem 9 proves that there exist codes that achieve $t = \Theta((d+\nu)n^{1-\alpha})$. This matches the lower bound in Theorem 8.

IV. Encoding & Decoding

We have so far focused on investigating the fundamental trade-offs between t and (d,ν,n) under constraints on either the number of items that can participate in a test (sparse test) or the number of tests an item can participate in (sparse codeword) without considering the encoding or decoding complexities. However, due to the emerging applications involving massive datasets there is a recent research effort towards low-complexity decoding schemes [25]–[29]. The computational complexity of encoding and decoding might be just as critical, therefore, it is desirable not to sacrifice on encoding or decoding complexity to achieve the optimal trade-off between t and (d,ν,n). In this section, we discuss the encoding and decoding complexity of the explicit constructions we presented earlier in this paper.

In the classical combinatorial group testing framework, the focus has been on designing testing strategies that can be decoded in poly(t)-time while achieving the best known upper bound $t = O(d^2\log n)$. Guruswami et al. present an efficiently decodable ($O(t)$ time decoding) d-disjunct matrix in [30] and their constructions require $O(d^4\log n)$ tests. The first result that achieves efficient decoding time while matching the $O(d^2\log n)$ bound on the number of tests was recently presented in [26]. Furthermore, the construction in [26] can be derandomized in the regime $d = O(\log N/\log\log N)$. Later in [27] the constraint on d is removed and an explicit construction is provided that can be decoded in time poly(t). The main idea considered in [26] was using *list-disjunct*

matrices and a similar idea was considered in [25] to obtain explicit constructions of non-adaptive group testing schemes that handle noisy tests and return a small superset of the defective items.

We now show that our explicit constructions can be decoded in $(\text{poly}(d) + O(t))$-time and each entry in any codeword can be computed in $\text{poly}(\log n)$ memory space by following a similar approach to [26]. This shows that these constructions not only (nearly) achieve the fundamental lower bound in the energy constrained setting, but also do that with a favorable encoding and decoding complexity. We begin with the following result, which is based on the noiseless setting.

Theorem 10. *The construction and guarantees of Theorem 5 and Theorem 8 ($\nu = 0$) can be achieved with decoding time* $\text{poly}(d) + O(t)$ *and* $\text{poly}(\log n)$ *memory space for the computation of each entry.*

Proof. We describe the decoding procedure as follows. For an output vector $Y \in \{0,1\}^t$, we can consider it as $Y = (Y_1, \ldots, Y_{t_q})$, a vector in $(\{0,1\}^q)^{t_q}$ where $t_q = ld+1$ is the block length of the outer Reed-Solomon code. Note that since we use the identity code as the inner code, for each $i \in [t_q]$, Y_i will have at most d ones and the support of Y_i will correspond to the symbols of defective items in the outer code. We now apply the following procedure. For each $i \in [t_q]$, we create the sets $S_i \subseteq [q]$ such that S_i is the support of Y_i. It follows that $|S_i| \leq d$ for every $i \in [t_q]$. We further have the following property. For any defective item, the corresponding codeword $(c_1, \ldots, c_{t_q})$ in the outer code must satisfy $c_i \in S_i$ for all $i \in [t_q]$ and for any non-defective item, the corresponding codeword $(c_1, \ldots, c_{t_q})$ in the outer code will include a symbol c_i such that $c_i \notin S_i$. Note that this step can be done in $O(t)$ time.

The second step is to output all codewords $(c_1, \ldots, c_{t_q})$ in the outer code such that $c_i \in S_i$ for all $i \in [t_q]$ given $S_i \subseteq [q]$ with $|S_i| \leq d$ for every $1 \leq i \leq t_q$. This problem is an instance of the error-free list recovery problem [31]–[33]. When each set S_i has at most s elements, it is referred to as list recovering with input lists of size s. It has been shown that the corresponding error-free list recovery problem can be solved in polynomial time for a $[t_q, k_q, t_q - k_q + 1]_q$ Reed-Solomon code as long as the parameter s satisfies $s < \lceil \frac{t_q}{k_q - 1} \rceil$ [31], [34]. We note that in our case, we have $s = d$, $t_q = ld + 1$, and $k_q = l + 1$, therefore $s < \lceil \frac{t_q}{k_q - 1} \rceil$. It follows that the second step can be done in time $\text{poly}(t_q)$. In particular, we can use the algorithm in [35] that runs in time $\text{poly}(d) \cdot t_q \log^2 t_q \log\log t_q$, which is $\text{poly}(d)$ with our choice of t_q (see Appendix D for a more detailed discussion about list recovery problem and its connection to [35]). Combining the two steps, we conclude that the decoding can be done in time $\text{poly}(d) + O(t)$.

The space complexity of an algorithm is the memory space required to solve an instance of the computational problem as a function of the size of the input. Any position of a Reed-Solomon codeword is an evaluation of a k_q-degree polynomial at a q-ary evaluation point. Therefore, any position in a Reed-Solomon codeword can be computed in $\text{poly}(k_q, \log q)$ memory space. Additionally, due to the identity mapping, any bit value of the identity inner code can be computed in $O(\log q)$ memory space. This provides the claimed space complexity for the reconstruction of the matrix. □

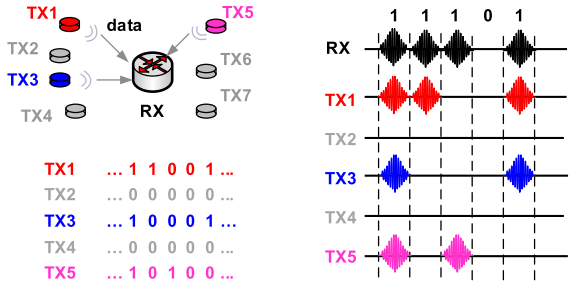

Fig. 1. Massive random access with on-off keying at the transmitters and energy detection at the receiver.

A similar approach can be taken in the case of noisy measurements by slightly modifying the parameters of the construction (only changing the constant in front of ν) to be appropriate for the noisy case. Our next result whose proof we give in Appendix E shows the validity of efficient decoding in the case of arbitrary errors.

Theorem 11. *The construction and guarantees of Theorem 7 and Theorem 8 under the modification $t_q = ld + (l+2)\nu + 1$ can be achieved with decoding time* $\text{poly}(d, \nu) + O(t)$ *and* $\text{poly}(\log n)$ *memory space for the computation of each entry.*

V. Low-Energy Massive Random Access

In this section, we discuss an application of our framework to wireless random access. Consider n devices (or sensors) that are associated with a single access point and assume that at most d of them can be active at any given time, where $n \gg d \gg 1$. We adopt the following modulation and detection technique at the transmitters and the receiver respectively: each device uses on-off signaling; i.e., it transmits a binary sequence of 0's and 1's, which corresponds to either transmitting a pulse or no pulse in every time slot. The access point simply detects whether or not there is energy in the channel in every time slot. This leads to a (potentially noisy) Boolean OR-channel from the devices to the access point. This simple modulation and detection technique is often used in low-rate applications in practice due to its simplicity. Energy detection does not require any channel state information at the receiver and thus it eliminates the need for channel training and estimation. This setting is depicted in Figure 1. To simplify the discussion, we focus on the device discovery problem, though as we argue in [36] the same group testing framework can be used to develop solutions for jointly discovering active devices and transmitting data, and for transmitting data without communicating device identities. The device discovery problem can be formulated as follows. Given n devices, design a length-t binary signature for each device (i.e., $M_i \in \{0,1\}^t$ for $i = 1, \ldots, n$) such that for any set $S \subset \{1, \ldots, n\}$ of active devices with size $|S| \leq d$, we can exactly identify the set S (the active devices) from

$$Y = \left(\bigvee_{i \in S} M_i \right) + \xi,$$

where $\bigvee$ denotes entry-wise Boolen-OR operation and '$+$' denotes entry-wise modulo-2 addition, and ξ is a length-t binary vector representing occasional errors in the energy detection at the receiver (flips of the output) due to noise. We assume that ξ has at most ν ones.

It can be readily observed that the problem statement above corresponds to a (non-adaptive) combinatorial group testing problem and the binary signature vectors M_i can be taken to be the columns of a $t \times n$ (d, ν)-disjunct group testing matrix. In particular, (d, ν)-disjunct matrices with small t will lead to short binary signatures for the devices. In practice, transmitters are subject to energy constraints, which are especially limiting in IoT applications where devices are required to operate on small batteries for many years or harvest their energy from their environment [37], [38]. In such applications, while it may still be desirable to minimize the length of the signature codewords to increase spectral efficiency, it may be also desirable to limit the energy needed to transmit each codeword. The energy spent for transmitting a codeword is proportional to the number of pulses, i.e. the number of "1"s, in the codeword (ignoring the standby energy for keeping the device active). Using our previous notation, this corresponds to imposing a constraint on the total number of "1"s in each column of M. Note that if energy efficiency were the only metric of interest, we could have resorted to the trivial solution that tests every item individually. This leads to a single "1" in each column of M, but the length of each column, and therefore that of our signature codewords, becomes equal to n. Our sparse group testing framework provides a way to optimally trade energy and spectral efficiency in this framework.

VI. Conclusion and Discussion

In this paper, we studied the combinatorial group testing problem under constraints on the number of items that can participate in each test (sparse test) or the number of tests each item can participate in (sparse codeword). We developed explicit group testing codes that minimize the number of tests under such constraints and proved that they are order optimal or nearly order optimal. Our results show that the minimal number of tests exhibits a particularly favorable behavior in the sparse codeword case, since the number of tests decreases drastically when the number of tests each item participates in increases beyond a bare minimum.

There are a few remaining gaps in our results that would be interesting to consider in future work. Firstly, as the number of tests per item increases linearly with d (i.e., $w = ld + 1$), the gap between our lower and upper bounds on t increases as a function of d. It would be interesting to see if this gap can be closed with sharper lower bounds or improved constructions that yield better performance. Secondly, the Kautz-Singleton construction provides d-disjunct matrices with row weight $\rho = n^{\frac{l}{l+1}}$. Therefore, the Kautz-Singleton construction cannot achieve a row weight of n^α for $\alpha < 1/2$. Nevertheless, as proven in Theorem 9, d-disjunct matrices with $\rho = n^\alpha$ for any real number $\alpha \in (0, 1)$ do exist. It would be interesting to know if there are optimal explicit constructions that can achieve $\rho = n^\alpha$ for $\alpha < 1/2$. Finally, while we have exclusively focused on the combinatorial group testing framework in the current paper, where the defective set is to be exactly recovered, we show in [39] that the Kautz-Singleton construction we consider in this work is also relevant in the probabilistic setting, where the defective set is to be recovered with a small probability of error. In [39], we build on the Kautz-Singleton construction to develop the first-order optimal strongly explicit construction for probabilistic group testing.

Appendix

A. Proof of Proposition 4

The achievability can trivially be obtained by individual testing, i.e., testing each item alone $\nu+1$ times. Note that this satisfies (d, ν)-disjunctiveness, therefore, $t(d, \nu, n) \leq (\nu+1)n$.

We can show that for a $t \times n$ binary matrix M that is (d, ν)-disjunct with the condition $w_{\max} \leq d+\nu$, all columns need at least $\nu + 1$ private rows, hence $t(d, \nu, n) \geq (\nu + 1)n$. Assume there exists a column $i \in [n]$ with at most ν private rows. It follows that this column has at least $w_i - \nu$ non-private rows. Fix any $w_i - \nu$ non-private rows. Since $w_{\max} \leq d + \nu$, it follows that $w_i - \nu \leq d$ and we can find at most d other columns covering these rows. Therefore, there exists a set S of columns with $|S| \leq d$ and $i \notin S$ such that $|\text{supp}(M_i) \backslash \cup_{j \in S} \text{supp}(M_j)| \leq \nu$, which contradicts with (d, ν)-disjunctiveness of M.

B. Proof of Theorem 6

Let M be a $t \times n$ (d, ν)-disjunct matrix with $w_{\max} \leq d + \nu + 1$. We will separate the columns of M into disjoint groups whose union is $[n]$. We define

$$\mathcal{N}_1 \coloneqq \{j \in [n] \mid w_j \leq d + \nu \text{ or } w_j = d + \nu + 1 \text{ and } M_j \text{ has at least } \nu + 1 \text{ private rows}\},$$
$$\mathcal{N}_{2,k} \coloneqq \{j \in [n] \mid w_j = d + \nu + 1 \text{ and } M_j \text{ has } k \text{ private rows}\} \text{ for } 0 \leq k \leq \nu.$$

Note that by construction, $\mathcal{N}_1 \cup (\cup_{0 \leq k \leq \nu} \mathcal{N}_{2,k}) = [n]$ and $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$, hence $n = n_1 + \sum_{0 \leq k \leq \nu} n_{2,k}$ where we denote $n_1 \coloneqq |\mathcal{N}_1|$ and $n_{2,k} \coloneqq |\mathcal{N}_{2,k}|$ for $0 \leq k \leq \nu$ respectively. In the following, we will bound the size of these sets.

We note that each column in the set $\mathcal{N}_1$ has at least $\nu + 1$ private rows and each column in the set $\mathcal{N}_{2,k}$ has k private rows for $0 \leq k \leq \nu$ respectively. Therefore, we have at least $\alpha \coloneqq (\nu + 1)n_1 + \sum_{k=0}^{\nu} k n_{2,k}$ private rows. Since a private row cannot be shared by two distinct columns and there could be at most t private rows, we have $0 \leq \alpha \leq t$.

For any $0 \leq k \leq \nu$, consider the set $\mathcal{N}_{2,k}$. Take any column $M_i \in \mathcal{N}_{2,k}$ if $\mathcal{N}_{2,k} \neq \emptyset$. Note that M_i has k private rows and $w_i = d + \nu + 1$. Considering the rest of $d + \nu + 1 - k$ non-private rows, we claim that all the pair of positions here must be private. We prove this by contradiction. Assume there exists another column covering any pair of positions among $d + \nu + 1 - k$ non-private rows. It follows that excluding this pair, among the rest of $d + \nu - 1 - k$ non-private rows, one can find at most $d - 1$ other columns covering $d - 1$ rows. Therefore, there exists a set S of columns with $|S| \leq d$ and

$i \notin S$ such that their union covers $d+1$ positions in the support of M_i. Since $w_i \leq d+\nu+1$, it follows that $|\text{supp}(M_i)\backslash \cup_{j\in S} \text{supp}(M_j)| \leq \nu$, which contradicts with (d,ν)-disjunctiveness. Therefore, there are $\binom{d+\nu+1-k}{2}$ private pairs for any $M_i \in \mathcal{N}_{2,k}$. By definition of a private set it cannot be shared by two distinct columns and we excluded the private rows in our calculation for the number of private pairs, hence we have $\sum_{0\leq k\leq \nu} n_{2,k}\binom{d+\nu+1-k}{2}$ private pairs whereas there could be at most $\binom{t-\alpha}{2}$ private pairs. Therefore, we have

$$\sum_{0\leq k\leq \nu} n_{2,k}\binom{d+\nu+1-k}{2} \leq \binom{t-\alpha}{2} \leq \frac{(t-\alpha)^2}{2}.$$

Hence, this gives

$$\begin{aligned} t &\geq \sqrt{\sum_{0\leq k\leq \nu} n_{2,k}(d+\nu+1-k)(d+\nu-k)} + \alpha \\ &= \sqrt{\sum_{0\leq k\leq \nu} n_{2,k}(d+\nu+1-k)(d+\nu-k)} + (\nu+1)n_1 \\ &\qquad + \sum_{k=0}^{\nu} k n_{2,k}, \end{aligned}$$

with the condition $n = n_1 + \sum_{0\leq k\leq e} n_{2,k}$. We can also write this as

$$\begin{aligned} t \geq \sqrt{\sum_{0\leq k\leq \nu} n_{2,k}(d+\nu+1-k)(d+\nu-k)} \\ + (\nu+1)\left(n - \sum_{0\leq k\leq \nu} n_{2,k}\right) + \sum_{k=0}^{\nu} k n_{2,k}, \end{aligned}$$

with the condition $0 \leq \sum_{0\leq k\leq \nu} n_{2,k} \leq n$.

Since this is a concave function over $\{n_{2,k}\}_{k=0}^{\nu}$ and $0 \leq \sum_{0\leq k\leq \nu} n_{2,k} \leq n$ is a convex set, minimum is attained over one of the extreme points. Therefore, we obtain

$$t \geq \min\Bigg\{(\nu+1)n, \tag{11}$$

$$\min_{0\leq k\leq \nu} \sqrt{(d+\nu+1-k)(d+\nu-k)n} + nk\Bigg\} \tag{12}$$

$$= \min\left\{(\nu+1)n, \sqrt{(d+\nu)(d+\nu+1)n}\right\},$$

where in the last step, we observe that the term that depends on k in (12) is concave over $0 \leq k \leq \nu$ and attains its minimum at either $k = 0$ or $k = \nu$ and $\sqrt{(d+\nu)(d+\nu+1)n} < \sqrt{d(d+1)n} + n\nu$.

C. Proof of Theorem 7

Let M be a $t \times n$ (d,ν)-disjunct matrix $w_{\max} \leq ld+\nu+1$. We define

$\mathcal{N}_1 \coloneqq \{j \in [n] \mid w_j \leq d+\nu$
or $w_j = d+\nu+1$ and M_j has at least one private row$\}$,
for $i = 2, \ldots, l$,
$\mathcal{N}_i \coloneqq \{j \in [n] \mid (i-2)d+\nu+2 \leq w_j \leq (i-1)d+\nu+1$
and M_j has no private set of size $i-1$ or
$(i-1)d+\nu+2 \leq w_j \leq id+\nu+1$ and M_j has at least one private set of size $i\}$,
$\mathcal{N}_{l+1} \coloneqq \{j \in [n] \mid (l-1)d+\nu+2 \leq w_j \leq ld+\nu+1$
and M_j has no private set of size $l\}$.

Note that by construction, $\mathcal{N}_1 \cup \ldots \cup \mathcal{N}_{l+1} = [n]$ and $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$ for any $i, j \in [l+1]$ such that $i \neq j$, hence $n = |\mathcal{N}_1| + \ldots + |\mathcal{N}_{l+1}|$. In the following, we will bound the size of these sets.

Note that $|\mathcal{N}_1| \leq t$. Consider the sets $\mathcal{N}_i$ for $i = 2, \ldots, l$. For any column $j \in \mathcal{N}_i$, if we have $(i-1)d+\nu+2 \leq w_j \leq id+\nu+1$, then by construction M_j has at least one private set of size i. For the case $(i-2)d+\nu+2 \leq w_j \leq (i-1)d+\nu+1$, using similar arguments as in the proof of Theorem 5, one can show that all the subsets of the support of M_j with size i must be private. Hence, all the columns in the set $\mathcal{N}_i$ must have at least one private set of size i. Since the private sets cannot be shared among columns and we have at most $\binom{t}{i}$ private sets of size i, it yields that $|\mathcal{N}_i| \leq \binom{t}{i}$. For the last set $\mathcal{N}_{l+1}$, similar arguments apply and for each column all the subsets of its support with size $l+1$ must be private. Since $w_j \geq (l-1)d+\nu+2$ for $j \in \mathcal{N}_{l+1}$, we have $|\mathcal{N}_{l+1}|\binom{(l-1)d+\nu+2}{l+1} \leq \binom{t}{l+1}$. Therefore,

$$\begin{aligned} n &= |\mathcal{N}_1| + \ldots + |\mathcal{N}_{l+1}| \\ &\leq \sum_{i=1}^{l}\binom{t}{i} + \frac{\binom{t}{l+1}}{\binom{(l-1)d+\nu+2}{l+1}} \\ &\overset{(i)}{\leq} \left(\frac{et}{l}\right)^l \\ &\qquad + \frac{t\ldots(t-l)}{((l-1)d+\nu+2)\ldots((l-1)(d-1)+\nu+1)} \\ &\overset{(ii)}{\leq} \frac{e^l t^l}{l^l} + \frac{t^{l+1}}{((l-1)(d-1)+\nu)^{l+1}} \\ &\overset{(iii)}{\leq} \frac{e^l t^l}{(l-1)^l}\frac{2t}{(d+\nu)^2} + \frac{t^{l+1}}{((l-1)(d-1)+\nu)^{l+1}} \\ &= t^{l+1}\left(\frac{2\ e^l}{(d+\nu)^2(l-1)^l} + \frac{1}{((l-1)(d-1)+\nu)^{l+1}}\right) \end{aligned}$$

where (i) is due to the inequality $\sum_{i=0}^{l}\binom{t}{i} \leq \left(\frac{et}{l}\right)^l$ for $t \geq l \geq 1$, (ii) is bounding all the terms in the numerator by t and denominator by $(l-1)(d-1)+\nu$ and in (iii) we use $t \geq \binom{d+\nu+2}{2} \geq \frac{(d+\nu)^2}{2}$. This completes the proof of the lower bound.

D. List Recovery

In this section, we discuss the (error-free) list recovery problem that comes into play in the decoding procedure of our construction in Theorem 10. The error-free list recovery problem is a special case of a more general problem known as soft decoding, which is defined as follows. The decoder is given a set of non-negative weights corresponding to each row and each symbol ($w_{i,\alpha}$, $i \in [t_q]$, $\alpha \in [q]$) and a threshold

$W \geq 0$. The decoder needs to output all codewords $(c_1, \ldots, c_{t_q})$ in q-ary code of block length t_q that satisfy

$$\sum_{i=1}^{t_q} w_{i,c_i} \geq W.$$

Note that the error-free list recovery is a special case of soft decoding under the parameters $W = t_q$ and $w_{i,\alpha} = 1$ for $\alpha \in S_i$ and $w_{i,\alpha} = 0$ otherwise. The soft decoding is related to weighted polynomial reconstruction problem, which is defined as follows. Given N points $(x_1, y_1), \ldots, (x_N, y_N)$, N non-negative integer weights $w(x_1, y_1), \ldots, w(x_N, y_N)$, and a parameter k, find all polynomials p of degree at most k such that $\sum_{i:p(x_i)=y_i} w(x_i, y_i) \geq W$. The algorithm presented in [35] solves this problem and runs in time poly(d) translated to our case.

E. Proof of Theorem 11

The decoding procedure follows what is described in the proof of Theorem 10. For each $i \in [t_q]$, we create the sets $S_i \subseteq [q]$ such that S_i is the support of Y_i. Due to the noise, we can only guarantee that $|S_i| \leq d + \nu$ for every $i \in [t_q]$ in this case. Similarly, for any defective item, the corresponding codeword $(c_1, \ldots, c_{t_q})$ in the outer code must satisfy $|\{i : c_i \in S_i\}| \geq t_q - \nu$ and for any non-defective item, the corresponding codeword $(c_1, \ldots, c_{t_q})$ in the outer code will include at least $(l+1)\nu \geq 2\nu$ symbols c_i such that $c_i \notin S_i$. Note that this step can be done in $O(t)$ time.

The second step is to output all codewords $(c_1, \ldots, c_{t_q})$ in the outer code such that $|\{i : c_i \in S_i\}| \geq t_q - \nu$ given $S_i \subseteq [q]$ with $|S_i| \leq d + \nu$ for every $1 \leq i \leq t_q$. This problem is an instance of the list recovery problem [31]–[33] and it can be solved in polynomial time for a $[t_q, k_q, t_q - k_q + 1]_q$ Reed-Solomon code, as long as $t_q - \nu > \sqrt{(k_q - 1)(d + \nu)t_q}$. We note that in our case, we have $t_q = ld + (l+2)\nu + 1$, and $k_q = l + 1$, which satisfies the requirement.

ACKNOWLEDGMENT

The authors thank the Associate Editor and the anonymous reviewers for helpful comments and suggestions.

REFERENCES

[1] D.-Z. Du and F. K. Hwang, *Combinatorial Group Testing and Its Applications*, vol. 12. Singapore: World Scientific, 2000.

[2] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Statist.*, vol. 14, no. 4, pp. 436–440, Dec. 1943.

[3] H.-B. Chen and F. K. Hwang, "A survey on nonadaptive group testing algorithms through the angle of decoding," *J. Combinat. Optim.*, vol. 15, no. 1, pp. 49–59, 2008.

[4] A. Ganesan, S. Jaggi, and V. Saligrama, "Learning immune-defectives graph through group tests," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3010–3028, May 2017.

[5] D. Malioutov and K. Varshney, "Exact rule learning via Boolean compressed sensing," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 2013, vol. 28, no. 3, pp. 765–773.

[6] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, "Group testing and sparse signal recovery," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, Oct. 2008, pp. 1059–1063.

[7] M. T. Goodrich, M. J. Atallah, and R. Tamassia, "Indexing information for data forensics," in *Applied Cryptography and Network Security*. Berlin, Germany: Springer, 2005, pp. 206–221.

[8] A. Emad and O. Milenkovic, "Poisson group testing: A probabilistic model for nonadaptive streaming Boolean compressed sensing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3335–3339.

[9] W. Kautz and R. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inf. Theory*, vol. IT-10, no. 4, pp. 363–377, Oct. 1964.

[10] T. Berger, N. Mehravari, D. Towsley, and J. Wolf, "Random multiple-access communication and group testing," *IEEE Trans. Commun.*, vol. COM-32, no. 7, pp. 769–779, Jul. 1984.

[11] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 2, pp. 185–191, Mar. 1985.

[12] A. K. Fletcher, S. Rangan, and V. K. Goyal, "A sparsity detection framework for on-off random access channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2009, pp. 169–173.

[13] J. Luo and D. Guo, "Neighbor discovery in wireless ad hoc networks based on group testing," in *Proc. 46th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2008, pp. 791–797.

[14] A. G. D'yachkov and V. V. Rykov, "Bounds on the length of disjunctive codes," *Probl. Peredachi Inf.*, vol. 18, no. 3, pp. 7–13, 1982.

[15] Z. Füredi, "On r-cover-free families," *J. Combinat. Theory A*, vol. 73, no. 1, pp. 172–173, 1996.

[16] E. Porat and A. Rothschild, "Explicit non-adaptive combinatorial group testing schemes," in *Automata, Languages and Programming*. Berlin, Germany: Springer, 2008, pp. 748–759.

[17] V. Gandikota, E. Grigorescu, S. Jaggi, and S. Zhou, "Nearly optimal sparse group testing," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2760–2773, May 2019.

[18] A. J. Macula, "A simple construction of d-disjunct matrices with certain constant weights," *Discrete Math.*, vol. 162, no. 1, pp. 311–312, Dec. 1996.

[19] N. Alon, D. Moshkovitz, and S. Safra, "Algorithmic construction of sets for k-restrictions," *ACM Trans. Algorithms*, vol. 2, no. 2, pp. 153–177, Apr. 2006.

[20] W. W. Peterson, *Error-Correcting Codes*. Cambridge, MA, USA: MIT Press, 1961.

[21] R. Singleton, "Maximum distance q-nary codes," *IEEE Trans. Inf. Theory*, vol. IT-10, no. 2, pp. 116–118, Apr. 1964.

[22] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 2, pp. 300–304, Jun. 1960.

[23] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

[24] R. M. Kainkaryam and P. J. Woolf, "Pooling in high-throughput drug screening," *Current Opinion Drug Discovery Develop.*, vol. 12, no. 3, pp. 339–350, 2009.

[25] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," in *Fundamentals of Computation Theory*. Berlin, Germany: Springer, 2009, pp. 62–73.

[26] P. Indyk, H. Q. Ngo, and A. Rudra, "Efficiently decodable non-adaptive group testing," in *Proc. 21st Annu. ACM-SIAM Symp. Discrete Algorithms*, 2010, pp. 1126–1142. [Online]. Available: http://dl.acm.org.stanford.idm.oclc.org/citation.cfm?id=1873601.1873692

[27] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications," in *Automata, Languages and Programming*. Berlin, Germany: Springer, 2011, pp. 557–568.

[28] K. Lee, R. Pedarsani, and K. Ramchandran, "SAFFRON: A fast, efficient, and robust framework for group testing based on sparse-graph codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2016, pp. 2873–2877.

[29] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Efficient algorithms for noisy group testing," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2113–2136, Apr. 2017.

[30] V. Guruswami and P. Indyk, "Linear-time list decoding in error-free settings," in *Proc. 31st Int. Colloq. Autom., Lang. Program. (ICALP)*. Berlin, Germany: Springer, 2004, pp. 695–707.

[31] V. Guruswami and A. Rudra, "Limits to list decoding Reed–Solomon codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3642–3649, Aug. 2006.

[32] V. Guruswami and P. Indyk, "Expander-based constructions of efficiently decodable codes," in *Proc. 42nd IEEE Symp. Found. Comput. Sci.*, Oct. 2001, pp. 658–667.

[33] A. Ta-Shma and D. Zuckerman, "Extractor codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3015–3025, Dec. 2004.

[34] V. Guruswami and M. Sudan, "Improved decoding of Reed–Solomon and algebraic-geometry codes," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1757–1767, Sep. 1999.

[35] M. Alekhnovich, "Linear diophantine equations over polynomials and soft decoding of Reed–Solomon codes," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2257–2265, Jul. 2005.
[36] A. H. Inan, P. Kairouz, and A. Ozgur, "Sparse group testing codes for low-energy massive random access," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput.* Oct. 2017, pp. 658–665.
[37] H. A. Inan and A. Ozgur, "Online power control for the energy harvesting multiple access channel," in *Proc. 14th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2016, pp. 1–6.
[38] H. A. Inan, D. Shaviv, and A. Özgür, "Capacity of the energy harvesting Gaussian MAC," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2347–2360, Apr. 2018.
[39] H. A. Inan, P. Kairouz, M. Wootters, and A. Özgür, "On the optimality of the Kautz–Singleton construction in probabilistic group testing," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5592–5603, Sep. 2019.

Huseyin A. Inan is a Ph.D. candidate in the department of Electrical Engineering at Stanford University. He received the B.Sc. degrees in Electrical & Electronics Engineering and Mathematics from Bogazici University, Istanbul, Turkey, in 2012 and the M.Sc. degree in Electrical & Electronics Engineering from Koc University, Istanbul, Turkey in 2014. His research interests include coding theory, wireless communication, and machine learning.

Peter Kairouz is a research scientist at Google. Before joining Google, he was a postdoctoral research fellow at Stanford University. He received his Ph.D. in ECE, M.S. in Maths, and M.S. in ECE from the University of Illinois at Urbana-Champaign (UIUC). During his Ph.D., he interned twice at Qualcomm and once at Google, where he designed privacy-aware utility-optimal unsupervised learning algorithms. He taught classes on big data and probabilities at UIUC, and was the General Chair for the 10th Annual Coordinated Science Laboratory Student Conference. His work on data privacy was recently featured on Forbes. He is the recipient of the 2012 Roberto Padovani Scholarship from Qualcomm's Research Center, the 2015 ACM SIGMETRICS Best Paper Award, the 2015 Qualcomm Innovation Fellowship Finalist Award, and the 2016 Harold L. Olesen Award for Excellence in Undergraduate Teaching from UIUC. His research interests span the areas of privacy-preserving data analysis, machine learning, and information theory.

Ayfer Özgür (SM'06) received her B.Sc. degrees in electrical engineering and physics from Middle East Technical University, Turkey, in 2001 and the M.Sc. degree in communications from the same university in 2004. From 2001 to 2004, she worked as hardware engineer for the Defense Industries Development Institute in Turkey. She received her Ph.D. degree in 2009 from the Information Processing Group at EPFL, Switzerland. In 2010 and 2011, she was a post-doctoral scholar at the same institution. She is currently an Associate Professor in the Electrical Engineering Department at Stanford University where she is a Hoover and Gabilan Fellow. Her current research interests include distributed communication and learning, wireless systems, and information theory. Dr. Özgür received the EPFL Best Ph.D. Thesis Award in 2010, the NSF CAREER award in 2013, the Okawa Foundation Research Grant and the IEEE Communication Theory Technical Committee (CTTC) Early Achievement Award in 2018, and was selected as the inaugural Goldsmith Lecturer of the IEEE ITSoc in 2020.