Noise-Resilient Group Testing: Limitations and Constructions

Mahdi Cheraghchi*

School of Computer and Communication Sciences
EPFL, 1015 Lausanne, Switzerland
mahdi.cheraghchi@epfl.ch

Abstract. We study combinatorial group testing schemes for learning d-sparse boolean vectors using highly unreliable disjunctive measurements. We consider an adversarial noise model that only limits the number of false observations, and show that any noise-resilient scheme in this model can only approximately reconstruct the sparse vector. On the positive side, we give a general framework for construction of highly noise-resilient group testing schemes using randomness condensers. Simple randomized instantiations of this construction give non-adaptive measurement schemes, with $m = O(d \log n)$ measurements, that allow efficient reconstruction of d-sparse vectors up to $O(d)$ false positives even in the presence of δm false positives and $\Omega(m/d)$ false negatives within the measurement outcomes, for *any* constant $\delta < 1$. None of these parameters can be substantially improved without dramatically affecting the others. Furthermore, we obtain several explicit (and incomparable) constructions, in particular one matching the randomized trade-off but using $m = O(d^{1+o(1)} \log n)$ measurements. We also obtain explicit constructions that allow fast reconstruction in time $\mathsf{poly}(m)$, which would be sublinear in n for sufficiently sparse vectors.

1 Introduction

Group testing is an area in applied combinatorics that deals with the following problem: Suppose that in a large population of individuals, it is suspected that a small number possess a condition or property that can only be certified by carrying out a particular test. Moreover suppose that a *pooling strategy* is permissible, namely, that it is possible to perform a test on a chosen group of individuals in parallel, in which case the outcome of the test would be positive if at least one of the individuals in the group possesses the condition. The trivial strategy would be to test each individual separately, which takes as many tests as the population size. The basic question in group testing is: how can we do better? The idea of group testing is believed to be emerged during the screening process of draftees in World War II. Since then, a vast amount of tools and

* Research supported by Swiss NSF grant 200020-115983/1.

M. Kutyłowski, M. Gębala, and W. Charatonik (Eds.): FCT 2009, LNCS 5699, pp. 62–73, 2009.

techniques have been developed in this area, and the problem has found a large number of applications apart from its original aim (from testing for defective items, e.g., defective light bulbs or resistors, as a part of industrial quality assurance to DNA sequencing and DNA library screening in molecular biology, and less obvious applications such as multiaccess communication, data compression, pattern matching, streaming algorithms, software testing, and compressed sensing, to name a few). We refer the reader to the books by Du and Hwang [1,2] for a detailed account of the major developments in this area.

More formally, the goal in group testing is to reconstruct a d-sparse[1] boolean vector[2] $x \in \mathbb{F}_2^n$, for a known integer parameter $d > 0$, from as few observations as possible. Each observation is the outcome of a measurement that outputs the bitwise OR of a prescribed subset of the coordinates in x. Hence, a measurement can be seen as a binary vector in $\mathbb{F}_2^n$ which is the characteristic vector of the subset of the coordinates being combined together. More generally, a set of m measurements can be seen as an $m \times n$ binary matrix (that we call the *measurement matrix*) whose rows define the individual measurements.

In this work we study group testing in presence of highly unreliable measurements that can produce false outcomes. We will mainly focus on situations where up to a constant fraction of the measurement outcomes can be incorrect. Moreover, we will mainly restrict our attention to *non-adaptive* measurements; the case in which the measurement matrix is fully determined before the observation outcomes are known. Nonadaptive measurements are particularly important for applications as they allow the tests to be performed independently and in parallel, which saves significant time and cost.

On the negative side, we show that when the measurements are allowed to be highly noisy, the original vector x cannot be uniquely reconstructed. Thus in this case it would be inevitable to resort to approximate reconstructions, i.e., producing a sparse vector $\hat{x}$ that is close to the original vector in Hamming distance. In particular, our result shows that if a constant fraction of the measurements can go wrong, the reconstruction might be different from the original vector in $\Omega(d)$ positions, irrespective of the number of measurements. For most applications this might be an unsatisfactory situation, as even a close estimate of the set of positives might not reveal whether any particular individual is defective or not, and in certain scenarios (such as an epidemic disease or industrial quality assurance) it is unacceptable to miss any affected individuals. This motivates us to focus on approximate reconstructions with *one-sided* error. Namely, we will require that the support of $\hat{x}$ contains the support of x and be possibly larger by up to $O(d)$ positions. It can be argued that, for most applications, such a scheme is as good as exact reconstruction, as it allows one to significantly narrow-down the set of defectives to up to $O(d)$ *candidate positives*. In particular, as observed in [3], one can use a *second stage* if necessary and individually test the resulting set of candidates to identify the exact set of positives, hence resulting in a so-called *trivial two-stage* group testing algorithm. Next, we will show that in

[1] We define a d-sparse vector as a vector with at most d nonzero coefficients.

[2] We use the notation $\mathbb{F}_q$ for a field (or at times, an alphabet) of size q.

any scheme that produces no or little false negative in the reconstruction, only up to $O(1/d)$ fraction of false negatives (i.e., observation of a 0 instead of 1) in the measurements can be tolerated, while there is no such restriction on the amount of tolerable false positives. Thus, one-sided approximate reconstruction breaks down the symmetry between false positives and false negatives in our error model.

On the positive side, we give a general construction for noise-resilient measurement matrices that guarantees approximate reconstructions up to $O(d)$ false positives. Our main result is a general reduction from the noise-resilient group testing problem to construction of well-studied combinatorial objects known as *randomness condensers* that play an important role in theoretical computer science. Different qualities of the underlying condenser correspond to different qualities of the resulting group testing scheme, as we describe later. Using the state of the art in derandomization theory, we obtain different instantiations of our framework with incomparable properties summarized in Table 1. In particular, the resulting randomized constructions (obtained from optimal lossless condensers and extractors) can be set to tolerate (with overwhelming probability) *any* constant fraction (< 1) of false positives, an $\Omega(1/d)$ fraction of false negatives, and produce an accurate reconstruction up to $O(d)$ false positives (where the positive constant behind $O(\cdot)$ can be made arbitrarily small), which is the best trade-off one can hope for, all using only $O(d \log n)$ measurements. This almost matches the information-theoretic lower bound $\Omega(d \log(n/d))$ shown by simple counting. We will also show explicit (deterministic) constructions that can approach the optimal trade-off, and finally, those that are equipped with fully efficient reconstruction algorithms with running time polynomial in the number of measurements.

Related Work. There is a large body of work in the group testing literature that is related to the present work; in this short presentation, we are only able

Table 1. A summary of constructions in this paper. The parameters $\alpha \in [0, 1)$ and $\delta \in (0, 1]$ are arbitrary constants, m is the number of measurements, e_0 (resp., e_1) the number of tolerable false positives (resp., negatives) in the measurements, and e'_0 is the number of false positives in the reconstruction. The fifth column shows whether the construction is deterministic (Det) or randomized (Rnd), and the last column shows the running time of the reconstruction algorithm.

m	e_0	e_1	e'_0	Det/ Rnd	Rec. Time
$O(d \log n)$	αm	$\Omega(m/d)$	$O(d)$	Rnd	$O(mn)$
$O(d \log n)$	$\Omega(m)$	$\Omega(m/d)$	δd	Rnd	$O(mn)$
$O(d^{1+o(1)} \log n)$	αm	$\Omega(m/d)$	$O(d)$	Det	$O(mn)$
$d \cdot \mathsf{quasipoly}(\log n)$	$\Omega(m)$	$\Omega(m/d)$	δd	Det	$O(mn)$
$d \cdot \mathsf{quasipoly}(\log n)$	αm	$\Omega(m/d)$	$O(d)$	Det	$\mathsf{poly}(m)$
$\mathsf{poly}(d)\mathsf{poly}(\log n)$	$\mathsf{poly}(d)\mathsf{poly}(\log n)$	$\Omega(e_0/d)$	δd	Det	$\mathsf{poly}(m)$

to discuss a few with the highest relevance. The exact group testing problem in the noiseless scenario is handled by what is known as *superimposed coding* (see [4,5]) or the closely related concepts of *cover-free families* or *disjunct matrices*[3]. It is known that, even for the noiseless case, exact reconstruction of d-sparse signals (when d is not too large) requires at least $\Omega(d^2 \log n / \log d)$ measurements (several proofs of this fact are known, e.g., [6,7,8]). An important class of superimposed codes is constructed from combinatorial designs, among which we mention the construction based on MDS codes given by Kautz and Singleton [9], which, in the group testing notation, achieves $O(d^2 \log^2 n)$ measurements.

Approximate reconstruction of sparse vectors up to a small number of false positives (that is one focus of this work) has been studied as a major ingredient of trivial two-stage schemes [3,10,11,12,13,14]. In particular, a generalization of superimposed codes, known as *selectors*, was introduced in [12] which, roughly speaking, allows for identification of the sparse vector up to a prescribed number of false positives. They gave a non-constructive result showing that there are such (non-adaptive) schemes that keep the number of false positives at $O(d)$ using $O(d \log(n/d))$ measurements, matching the optimal "counting bound". A probabilistic construction of asymptotically optimal selectors (resp., a related notion of *resolvable matrices*) is given in [14] (resp., [13]), and [15,16] give slightly sub-optimal "explicit" constructions based on certain expander graphs obtained from dispersers.

To give a concise comparison of the present work with those listed above, we mention some of the qualities of the group testing schemes that we will aim to attain: (1) low number of measurements; (2) arbitrarily good degree of approximation; (3) maximum possible noise tolerance; (4) efficient, deterministic construction: As typically the sparsity d is very small compared to n, a measurement matrix must be ideally *fully explicitly constructible* in the sense that each entry of the matrix should be computable in deterministic time $\mathsf{poly}(d, \log n)$; (5) fully efficient reconstruction algorithm: For a similar reason, the length of the observation vector is typically far smaller than n; thus, it is desirable to have a reconstruction algorithm that identifies the support of the sparse vector in time polynomial in the number of measurements (which might be exponentially smaller than n). While the works that we mentioned focus on few of the criteria listed above, our approach can potentially attain *all* at the same time. As we will see later, using the best known constructions of condensers we will have to settle to sub-optimal results in one or more of the aspects above. Nevertheless, the fact that any improvement in the construction of condensers would readily translate to improved group testing schemes (and also the rapid growth of derandomization theory) justifies the significance of the construction given in this work.

[3] A d-superimposed code is a collection of binary vectors with the property that from the bitwise OR of up to d words in the family one can uniquely identify the comprising vectors. A d-cover-free family is a collection of subsets of a universe, none of which is contained in any union of up to d of the other subsets.

2 Preliminaries

For non-negative integers e_0 and e_1, we say that an ordered pair of binary vectors (x, y), each in $\mathbb{F}_2^n$, are (e_0, e_1)-close (or x is (e_0, e_1)-close to y) if y can be obtained from x by flipping at most e_0 bits from 0 to 1 and at most e_1 bits from 1 to 0. Hence, such x and y will be $(e_0 + e_1)$-close in Hamming-distance. Further, (x, y) are called (e_0, e_1)-far if they are not (e_0, e_1)-close. Note that if x and y are seen as characteristic vectors of subsets X and Y of $[n]$, respectively[4], they are $(|Y \setminus X|, |X \setminus Y|)$-close. Furthermore, (x, y) are (e_0, e_1)-close iff (y, x) are (e_1, e_0)-close. A group of m non-adaptive measurements for binary vectors of length n can be seen as an $m \times n$ matrix (that we call the *measurement matrix*) whose (i, j)th entry is 1 iff the jth coordinate of the vector is present in the disjunction defining the ith measurement. For a measurement matrix A, we denote by $A[x]$ the outcome of the measurements defined by A on a binary vector x, that is, the bitwise OR of those columns of A chosen by the support of x. As motivated by our negative results, for the specific setting of the group testing problem that we are considering in this work, it is necessary to give an *asymmetric* treatment that distinguishes between inaccuracies due to false positives and false negatives. Thus, we will work with a notion of error-tolerating measurement matrices that directly and conveniently captures this requirement, as given below:

Definition 1. Let $m, n, d, e_0, e_1, e_0', e_1'$ be integers. An $m \times n$ measurement matrix A is called (e_0, e_1, e_0', e_1')-correcting for d-sparse vectors if, for every $y \in \mathbb{F}_2^m$ there exists $z \in \mathbb{F}_2^n$ (called a *valid decoding of* y) such that for every $x \in \mathbb{F}_2^n$, whenever (x, z) are (e_0', e_1')-far, $(A[x], y)$ are (e_0, e_1)-far. The matrix A is called *fully explicit* if each entry of the matrix can be computed in time $\mathsf{poly}(\log n)$.

Intuitively, the definition states that two measurements are allowed to be confused only if they are produced from close vectors. In particular, an (e_0, e_1, e_0', e_1')-correcting matrix gives a group testing scheme that reconstructs the sparse vector up to e_0' false positives and e_1' false negatives even in the presence of e_0 false positives and e_1 false negatives in the measurement outcome. Under this notation, unique (exact) decoding would be possible using an $(e_0, e_1, 0, 0)$-correcting matrix if the amount of measurement errors is bounded by at most e_0 false positives and e_1 false negatives. However, when $e_0' + e_1'$ is positive, decoding may require a bounded amount of ambiguity, namely, up to e_0' false positives and e_1' false negatives in the decoded sequence. In the combinatorics literature, the special case of $(0, 0, 0, 0)$-correcting matrices is known as *d-superimposed codes* or *d-separable matrices* and is closely related to the notions of *d-cover-free families* and *d-disjunct* matrices (cf. [1] for precise definitions). Also, $(0, 0, e_0', 0)$-correcting matrices are related to the notion of *selectors* in [12] and *resolvable matrices* in [13].

The *min-entropy* of a distribution $\mathcal{X}$ with finite support S is given by $H_\infty(\mathcal{X}) := \min_{x \in S}\{-\log \Pr_{\mathcal{X}}(x)\}$, where $\Pr_{\mathcal{X}}(x)$ is the probability that $\mathcal{X}$ assigns to x. The *statistical distance* of two distributions $\mathcal{X}$ and $\mathcal{Y}$ defined on the same finite

[4] We use the shorthand $[n]$ for the set $\{1, 2, \ldots, n\}$.

space S is given by $\frac{1}{2}\sum_{s\in S}|\Pr_{\mathcal{X}}(s)-\Pr_{\mathcal{Y}}(s)|$, which is half the ℓ_1 distance of the two distributions when regarded as vectors of probabilities over S. 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 $\mathbb{F}_2^n$, and $X\sim\mathcal{X}$ for a random variable X drawn from a distribution $\mathcal{X}$. A function $C\colon \mathbb{F}_2^n\times\mathbb{F}_2^t\to\mathbb{F}_2^\ell$ is a *strong* $k\to_\epsilon k'$ *condenser* if for every distribution $\mathcal{X}$ on $\mathbb{F}_2^n$ with min-entropy at least k, random variable $X\sim\mathcal{X}$ and a *seed* $Y\sim\mathcal{U}_t$, the distribution of $(Y,C(X,Y))$ is ϵ-close to some distribution $(\mathcal{U}_t,\mathcal{Z})$ with min-entropy at least $t+k'$. The parameters ϵ, $k-k'$, and $\ell-k'$ are called the *error*, the *entropy loss* and the *overhead* of the condenser, respectively. A condenser with zero entropy loss is called *lossless*, and a condenser with zero overhead is called a *strong* (k,ϵ)*-extractor*. A condenser is *explicit* if it is polynomial-time computable.

3 Negative Results

In coding theory, it is possible to construct codes that can tolerate up to a constant fraction of adversarially chosen errors and still guarantee unique decoding. Hence it is natural to wonder whether a similar possibility exists in group testing, namely, whether there is a measurement matrix that is robust against a constant fraction of adversarial errors and still recovers the measured vector exactly. The result below shows that this is not possible[5]:

Lemma 2. *Suppose that an $m\times n$ measurement matrix A is (e_0,e_1,e_0',e_1')-correcting for d-sparse vectors. Then $(\max\{e_0,e_1\}+1)/(e_0'+e_1'+1)\le m/d$.* □

The above lemma[6] gives a trade-off between the tolerable error in the measurements versus the reconstruction error. In particular, for unique decoding to be possible one can only guarantee resiliency against up to $O(1/d)$ fraction of errors in the measurement. On the other hand, tolerance against a constant fraction of errors would make an ambiguity of order $\Omega(d)$ in the decoding inevitable. Another trade-off is given by the following lemma:

Lemma 3. *Suppose that an $m\times n$ measurement matrix A is (e_0,e_1,e_0',e_1')-correcting for d-sparse vectors. Then for every $\epsilon>0$, either $e_1<(e_1'+1)m/(\epsilon d)$ or $e_0'\ge(1-\epsilon)(n-d+1)/(e_1'+1)^2$.* □

As mentioned in the introduction, it is an important matter for applications to bring down the amount of false negatives in the reconstruction as much as possible, and ideally to zero. The lemma above shows that if one is willing to keep the number e_1' of false negatives in the reconstruction at the zero level (or bounded by a constant), only an up to $O(1/d)$ fraction of false negatives in the measurements can be tolerated (regardless of the number of measurements),

[5] We remark that the negative results in this section hold for both adaptive and non-adaptive measurements.

[6] The omitted proofs can be found in the full version of this paper.

unless the number e'_0 of false positives in the reconstruction grows to an enormous amount (namely, $\Omega(n)$ when $n - d = \Omega(n)$) which is certainly undesirable.

As shown in [6], exact reconstruction of d-sparse vectors of length n, even in a noise-free setting, requires at least $\Omega(d^2 \log n / \log d)$ non-adaptive measurements. However, it turns out that there is no such restriction when an approximate reconstruction is sought for, except for the following bound which can be shown using simple counting and holds for adaptive noiseless schemes as well:

Lemma 4. *Let A be an $m \times n$ measurement matrix that is $(0, 0, e'_0, e'_1)$-correcting for d-sparse vectors. Then $m \geq d \log(n/d) - d - e'_0 - O(e'_1 \log((n - d - e'_0)/e'_1))$, where the last term is defined to be zero for $e'_1 = 0$.* □

This is similar in spirit to the lower bound obtained in [12] for the size of selectors. According to the lemma, even in the noiseless scenario, any reconstruction method that returns an approximation of the sparse vector up to $e'_0 = O(d)$ false positives and without false negatives will require $\Omega(d \log(n/d))$ measurements. As we will show in the next section, an upper bound of $O(d \log n)$ is in fact attainable even in a highly noisy setting using only non-adaptive measurements. This in particular implies an asymptotically optimal trivial two-stage group testing scheme.

4 A Noise-Resilient Construction

In this section we introduce our general construction and design measurement matrices for testing D-sparse vectors[7] in $\mathbb{F}_2^N$. The matrices can be seen as adjacency matrices of certain unbalanced bipartite graphs constructed from good randomness condensers or extractors. The main technique that we use to show the desired properties is the *list-decoding view* of randomness condensers, extractors, and expanders, developed over the recent years starting from the work of Ta-Shma and Zuckerman on *extractor codes* [17]. We start by introducing the terms that we will use in this construction and the analysis.

Definition 5. (mixtures, agreement, and agreement list) Let Σ be a finite set. A *mixture* over Σ^n is an n-tuple $S := (S_1, \ldots, S_n)$ such that every S_i, $i \in [n]$, is a nonempty subset of Σ. The *agreement* of $w := (w_1, \ldots w_n) \in \Sigma^n$ with S, denoted by $\mathsf{Agr}(w, S)$, is the quantity $\frac{1}{n}|\{i \in [n] \colon w_i \in S_i\}|$. Moreover, we define the quantity $\mathsf{wgt}(S) := \sum_{i \in [n]} |S_i|$ and $\rho(S) := \mathsf{wgt}(S)/(n|\Sigma|)$, where the latter is the expected agreement of a random vector with S. For a code $\mathcal{C} \subseteq \Sigma^n$ and $\alpha \in (0, 1]$, the *α-agreement list* of $\mathcal{C}$ with respect to S, denoted by $\mathsf{LIST}_{\mathcal{C}}(S, \alpha)$, is the set[8] $\mathsf{LIST}_{\mathcal{C}}(S, \alpha) := \{c \in \mathcal{C} \colon \mathsf{Agr}(c, S) > \alpha\}$.

Definition 6. (induced code) Let $f \colon \Gamma \times \Omega \to \Sigma$ be a function mapping a finite set $\Gamma \times \Omega$ to a finite set Σ. For $x \in \Gamma$, we use the shorthand $f(x)$ to denote

[7] In this section we find it more convenient to use capital letters $D, N, \ldots$ instead of $d, n, \ldots$ that we have so far used and keep the small letters for their base-2 logarithms.

[8] When $\alpha = 1$, we consider codewords with full agreement with the mixture.

the vector $y := (y_i)_{i\in\Omega}$, $y_i := f(x,i)$, whose coordinates are indexed by the elements of Ω in a fixed order. The *code induced by* f, denoted by $\mathcal{C}(f)$ is the set $\{f(x)\colon x \in \Gamma\}$. The induced code has a natural encoding function given by $x \mapsto f(x)$.

Definition 7. (codeword graph) Let $\mathcal{C} \subseteq \Sigma^n$, $|\Sigma| = q$, be a q-ary code. The *codeword graph* of $\mathcal{C}$ is a bipartite graph with left vertex set $\mathcal{C}$ and right vertex set $n \times \Sigma$, such that for every $x = (x_1, \ldots, x_n) \in \mathcal{C}$, there is an edge between x on the left and $(1, x_1), \ldots, (n, x_n)$ on the right. The *adjacency matrix* of the codeword graph is an $n|\Sigma| \times |\mathcal{C}|$ binary matrix whose (i,j)th entry is 1 iff there is an edge between the ith right vertex and the jth left vertex.

The following is a straightforward generalization of the result in [17] that is also shown in [18]:

Theorem 8. *Let $f\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ be a strong $k \to_\epsilon k'$ condenser, and $\mathcal{C} \subseteq \Sigma^{2^t}$ be its induced code, where $\Sigma := \mathbb{F}_2^\ell$. Then for any mixture S over Σ^{2^t} we have $|\mathsf{LIST}_{\mathcal{C}}(S, \rho(S)2^{\ell-k'} + \epsilon)| < 2^k$.* □

Now using the above tools, we are ready to describe our construction of error-tolerant measurement matrices. We first state a general result without specifying the parameters of the condenser, and then instantiate the construction with various choices of the condenser, resulting in matrices with different properties.

Theorem 9. *Let $f\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ be a strong $k \to_\epsilon k'$ condenser, and $\mathcal{C}$ be its induced code, and define the capital shorthands $K := 2^k$, $K' := 2^{k'}$, $L := 2^\ell$, $N := 2^n$, $T := 2^t$. Suppose that the parameters $p, \nu, \gamma > 0$ are chosen such that $(p+\gamma)L/K' + \nu/\gamma < 1 - \epsilon$, and $D := \gamma L$. Then the adjacency matrix of the codeword graph of $\mathcal{C}$ (which has $M := TL$ rows and N columns) is a $(pM, (\nu/D)M, K - D, 0)$-correcting measurement matrix for D-sparse vectors. Moreover, it allows for a reconstruction algorithm with running time $O(MN)$.*

Proof. Let $\mathcal{M}$ be the adjacency matrix of the codeword graph of $\mathcal{C}$. It immediately follows from the construction that the number of rows of $\mathcal{M}$ (denoted by M) is equal to TL. Moreover, notice that the Hamming weight of each column of $\mathcal{M}$ is exactly T. Let $x \in \mathbb{F}_2^N$ and denote by $y \in \mathbb{F}_2^M$ its encoding, i.e., $y := \mathcal{M}[x]$, and by $\hat{y} \in \mathbb{F}_2^M$ a *received word*, or a *noisy* version of y. The encoding of x can be schematically viewed as follows: The coefficients of x are assigned to the left vertices of the codeword graph and the encoded bit on each right vertex is the bitwise OR of the values of its neighbors. The coordinates of x can be seen in one-to-one correspondence with the codewords of $\mathcal{C}$. Let $X \subseteq \mathcal{C}$ be the set of codewords corresponding to the support of x. The coordinates of the noisy encoding $\hat{y}$ are indexed by the elements of $[T] \times [L]$ and thus, $\hat{y}$ naturally defines a mixture $S = (S_1, \ldots, S_T)$ over $[L]^T$, where S_i contains j iff $\hat{y}$ at position (i,j) is 1. Observe that $\rho(S)$ is the relative Hamming weight (denoted below by $\delta(\cdot)$) of $\hat{y}$; thus, we have $\rho(S) = \delta(\hat{y}) \le \delta(y) + p \le D/L + p = \gamma + p$, where the last inequality comes from the fact that the relative weight of each

column of $\mathcal{M}$ is exactly $1/L$ and that x is D-sparse. Furthermore, from the assumption we know that the number of false negatives in the measurement is at most $\nu TL/D = \nu T/\gamma$. Therefore, any codeword in X must have agreement at least $1 - \nu/\gamma$ with S. This is because S is indeed constructed from a mixture of the elements in X, modulo false positives (that do not decrease the agreement) and at most $\nu T/\gamma$ false negatives each of which can reduce the agreement by at most $1/T$. Accordingly, we consider a decoder which simply outputs a binary vector $\hat{x}$ supported on the coordinates corresponding to those codewords of $\mathcal{C}$ that have agreement larger than $1 - \nu/\gamma$ with S. Clearly, the running time of the decoder is linear in the size of the measurement matrix. By the discussion above, $\hat{x}$ must include the support of x. Moreover, Theorem 8 applies for our choice of parameters, implying that $\hat{x}$ must have weight less than K. □

Instantiations

Now we instantiate the general result given by Theorem 9 with various choices of the underlying condenser and compare the obtained parameters.

Applying Optimal Extractors. Radhakrishan and Ta-Shma showed that non-constructively, for every k, n, ϵ, there is a strong (k, ϵ)-extractor with seed length $t = \log(n - k) + 2\log(1/\epsilon) + O(1)$ and output length $\ell = k - 2\log(1/\epsilon) - O(1)$, which is the best one can hope for [19]. In particular, they show that a random function achieves these parameters with probability $1 - o(1)$. Plugging this result in Theorem 9, we obtain a non-explicit measurement matrix from a simple, randomized construction that achieves the desired trade-off with high probability:

Corollary 10. *For every choice of constants $p \in [0, 1)$ and $\nu \in [0, \nu_0)$, $\nu_0 := (\sqrt{5 - 4p} - 1)^3/8$, and positive integers D and $N \geq D$, there is an $M \times N$ measurement matrix, where $M = O(D \log N)$, that is $(pM, (\nu/D)M, O(D), 0)$-correcting for D-sparse vectors of length N and allows for a reconstruction algorithm with running time $O(MN)$.* □

This instantiation, in particular, reproduces a result on randomized construction of approximate group testing schemes with optimal number of measurements in [14], but with stringent conditions on the noise tolerance of the scheme.

Applying Optimal Lossless Condensers. The probabilistic construction of Radhakrishan and Ta-Shma can be extended to the case of lossless condensers and one can show that a random function is with high probability a strong $k \to_\epsilon k$ condenser with seed length $t = \log n + \log(1/\epsilon) + O(1)$ and output length $\ell = k + \log(1/\epsilon) + O(1)$ [20]. This combined with Theorem 9 gives the following corollary:

Corollary 11. *For positive integers $N \geq D$ and every constant $\delta > 0$ there is an $M \times N$ measurement matrix, where $M = O(D \log N)$, that is $(\Omega(M), \Omega(1/D)M, \delta D, 0)$-correcting for D-sparse vectors of length N and allows for a reconstruction algorithm with running time $O(MN)$.* □

Both results obtained in Corollaries 10 and 11 almost match the lower bound of Lemma 4 for the number of measurements. However, we note the following distinction between the two results: Instantiating the general construction of Theorem 9 with an extractor gives us a sharp control over the fraction of tolerable errors, and in particular, we can obtain a measurement matrix that is robust against *any* constant fraction (bounded from 1) of false positives. However, the number of false positives in the reconstruction will be bounded by some constant fraction of the sparsity of the vector that cannot be made arbitrarily close to zero. On the other hand, a lossless condenser enables us to bring down the number of false positives in the reconstruction to an arbitrarily small fraction of D (which is, in light of Lemma 2, the best we can hope for), but on the other hand, does not give as good a control on the fraction of tolerable errors as in the extractor case, though we still obtain resilience against the same order of errors.

Applying the Guruswami-Umans-Vadhan's Extractor. While Corollaries 10 and 11 give probabilistic constructions of noise-resilient measurement matrices, certain applications require a fully explicit matrix that is guaranteed to work. To that end, we need to instantiate Theorem 9 with an explicit condenser. First, we use a nearly-optimal explicit extractor due to Guruswami, Umans and Vadhan, summarized in the following theorem:

Theorem 12. *[18] For all positive integers $n \geq k$ and all $\epsilon > 0$, there is an explicit strong (k, ϵ)-extractor $\mathsf{Ext}\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ with $\ell = k - 2\log(1/\epsilon) - O(1)$ and $t = \log n + O(\log k \cdot \log(k/\epsilon))$.* □

Applying this result in Theorem 9 we obtain a similar trade-off as in Corollary 10, except for a higher number of measurements which would be bounded by $O(2^{O(\log^2 \log D)} D \log N) = O(D^{1+o(1)} \log N)$.

Applying the Zig-Zag Lossless Condenser. In [20] an explicit lossless condenser with optimal output length is constructed. In particular they show the following:

Theorem 13. *[20] For every $k \leq n \in \mathbb{N}$, $\epsilon > 0$ there is an explicit $k \to_\epsilon k$ condenser*[9] *with seed length $O(\log^3(n/\epsilon))$ and output length $k + \log(1/\epsilon) + O(1)$.*

Combined with Theorem 9, we obtain a similar result as in Corollary 11, except for a higher number of measurements, namely, $M = D2^{\log^3(\log N)} = D \cdot \mathsf{quasipoly}(\log N)$.

Measurements Allowing Sublinear Time Reconstruction. The naive reconstruction algorithm of Theorem 9 works efficiently in linear time in the size of the measurement matrix. However, as mentioned in the introduction, for very sparse vectors (i.e., $D \ll N$) it might be of practical importance to have a reconstruction algorithm that runs in *sublinear* time in N, the length of the

[9] Though not explicitly mentioned in [20], these condensers can be considered to be strong.

vector, and ideally, polynomial in the number of measurements, which is merely $\mathsf{poly}(\log N, D)$ if the number of measurements is optimal.

Observe that the main computational task done by the reconstruction algorithm in Theorem 9 is in fact computation of a suitable agreement list for the induced code of the underlying condenser. Several explicit constructions of condensers are equipped with efficient algorithms for computation of agreement lists that substantially outperform exhaustive search. Namely, for such constructions the set $\mathsf{LIST}_{\mathcal{C}}(S, \rho(S)+\epsilon)$ can be computed in time $\mathsf{poly}(2^t, 2^\ell, 2^k, 1/\epsilon)$, which can be much smaller than 2^n. Here we consider two such constructions that achieve the most favorable parameters for our application: Trevisan's extractor[10] [21] and a lossless condenser due to Guruswami *et al.* [18]. We use the following improvement of Trevisan's extractor due to Raz *et al.*:

Theorem 14. *[22] For every $n, k, \ell \in \mathbb{N}$, $(\ell \leq k \leq n)$ and $\epsilon > 0$, there is an explicit strong (k, ϵ)-extractor $\mathsf{Tre}\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ with $t = O(\log^2(n/\epsilon) \cdot \log(1/\alpha))$, where $\alpha := k/(\ell-1) - 1$ must be less than $1/2$.* □

Using this result in Theorem 9, we obtain a measurement matrix for which the reconstruction is possible in polynomial time in the number of measurements. Specifically, we obtain the same parameters as in Corollary 10 using Trevisan's extractor except for the number of measurements, $M = O(D2^{\log^3 \log N}) = D \cdot \mathsf{quasipoly}(\log N)$.

In the world of lossless condensers, Guruswami *et al.* [18] show the following:

Theorem 15. *[18] For all constants $\alpha \in (0, 1)$ and every $k \leq n \in \mathbb{N}$, $\epsilon > 0$ there is an explicit strong $k \to_\epsilon k$ condenser with seed length $t = (1 + 1/\alpha)\log(nk/\epsilon) + O(1)$ and output length $\ell = d + (1+\alpha)k$. Moreover, the condenser has efficient list recovery.* □

As before, we use this construction in Theorem 9 and obtain the following:

Corollary 16. *For positive integers $N \geq D$ and any constants $\delta, \alpha > 0$ there is an $M \times N$ measurement matrix, where $M = O(D^{3+\alpha+2/\alpha}(\log N)^{2+2/\alpha})$, that is $(\Omega(e), \Omega(e/D), \delta D, 0)$-correcting for D-sparse vectors of length N, where $e := (\log N)^{1+1/\alpha} D^{2+1/\alpha}$. Moreover, the matrix allows for a reconstruction algorithm with running time $\mathsf{poly}(M)$.* □

Acknowledgment

The author is thankful to Amin Shokrollahi for introducing him to the group testing problem and his comments on an earlier draft of this paper, and to Venkatesan Guruswami for several illuminating discussions that led to considerable improvement of the results presented in this work.

[10] Trevisan's extractor belongs to a class of extractors obtained from *black-box pseudorandom generators*. Ta-Shma and Zuckerman [17] show that for any such construction the agreement list is efficiently computable.

References

1. Du, D.Z., Hwang, F.: Combinatorial Group Testing and its Applications, 2nd edn. World Scientific, Singapore (2000)
2. Du, D.-Z., Hwang, F.K.: Pooling Designs and Nonadaptive Group Testing. World Scientific, Singapore (2006)
3. Knill, E.: Lower bounds for identifying subset members with subset queries. In: Proceedings of SODA, pp. 369–377 (1995)
4. Dyachkov, A., Rykov, V.: A survey of superimposed code theory. Problems of Control and Information Theory 12(4), 229–242 (1983)
5. Knill, E., Bruno, W.J., Torney, D.C.: Non-adaptive group testing in the presence of errors. Discrete Appl. Math. 88(1–3), 261–290 (1998)
6. D'yachkov, A.G., Rykov, V.: Bounds of the length of disjunct codes. Problems of Control and Information Theory 11, 7–13 (1982)
7. Ruszinkó: On the upper bound of the size of the r-cover-free families. J. Comb. Theory., Series A 66, 302–310 (1994)
8. Füredi, Z.: On r-cover-free families. J. Comb. Theory, Series A 73, 172–173 (1996)
9. Kautz, W., Singleton, R.: Nonrandom binary superimposed codes. IEEE Transactions on Information Theory 10, 363–377 (1964)
10. Macula, A.: Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. J. Comb. Optim. 2, 385–397 (1999)
11. Berger, T., Mandell, J., Subrahmanya, P.: Maximally efficient two-stage group testing. Biometrics 56, 833–840 (2000)
12. De Bonis, A., Gasieniec, L., Vaccaro, U.: Optimal two-stage algorithms for group testing problems. SIAM Journal on Computing 34(5), 1253–1270 (2005)
13. Eppstein, D., Goodrich, M., Hirschberg, D.: Improved combinatorial group testing algorithms for real-world problem sizes. SIAM J. Comp. 36(5), 1360–1375 (2007)
14. Cheng, Y., Du, D.Z.: New constructions of one- and two-stage pooling designs. Journal of Computational Biology 15(2), 195–205 (2008)
15. Indyk, P.: Explicit constructions of selectors with applications. In: Proceedings of SODA (2002)
16. Chlebus, B., Kowalski, D.: Almost optimal explicit selectors. In: Liśkiewicz, M., Reischuk, R. (eds.) FCT 2005. LNCS, vol. 3623, pp. 270–280. Springer, Heidelberg (2005)
17. Ta-Shma, A., Zuckerman, D.: Extractor codes. IEEE Transactions on Information Theory 50(12), 3015–3025 (2004)
18. Guruswami, V., Umans, C., Vadhan, S.: Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. In: Proc. of the 22nd IEEE CCC (2007)
19. Radhakrishan, J., Ta-Shma, A.: Tight bounds for depth-two superconcentrators. In: Proceedings of the 38th FOCS, pp. 585–594 (1997)
20. Capalbo, M., Reingold, O., Vadhan, S., Wigderson, A.: Randomness conductors and constant-degree expansion beyond the degree/2 barrier. In: Proceedings of the 34th STOC, pp. 659–668 (2002)
21. Trevisan, L.: Extractors and pseudorandom generators. Journal of the ACM 48(4), 860–879 (2001)
22. Raz, R., Reingold, O., Vadhan, S.: Extracting all the randomness and reducing the error in Trevisan's extractor. JCSS 65(1), 97–128 (2002)