HOW LITTLE DOES NON-EXACT RECOVERY HELP IN GROUP TESTING?

Jonathan Scarlett and Volkan Cevher

LIONS, EPFL, Lausanne, Switzerland
{jonathan.scarlett,volkan.cevher}@epfl.ch

ABSTRACT

We consider the group testing problem, in which one seeks to identify a subset of defective items within a larger set of items based on a number of tests. We characterize the information-theoretic performance limits in the presence of list decoding, in which the decoder may output a list containing more elements than the number of defectives, and the only requirement is that the true defective set is a subset of the list, or more generally, that their overlap exceeds a given threshold. We show that even under this highly relaxed criterion, in several scaling regimes the asymptotic number of tests is no smaller than the exact recovery setting. However, we also provide examples where a reduction is provably attained. We support our theoretical findings with numerical experiments.

Keywords: Group testing, information-theoretic limits, partial recovery, list decoding, strong converse

1. INTRODUCTION

The group testing problem consists of determining a small subset of "defective" items within a larger set of items. This problem has a history in areas such as medical testing and fault detection, and has regained significant attention following new applications in areas such as communication protocols [1], pattern matching [2], and database systems [3], and new connections with compressive sensing [4, 5].

Let the items be labeled as $\{1, \dots, p\}$, and let S be the subset of defective items. Each observation is generated according to

$$Y = \mathbb{1}\Big\{ \bigcup_{i \in S} X_i = 1 \Big\}, \tag{1}$$

where the measurement vector $X = (X_1, \dots, X_p) \in \{0,1\}^p$ indicates which items are included in the test. The goal is to recover S based on a number n of tests, with the i-th measurement vector being $X^{(i)}$ and the i-th observation being $Y^{(i)}$. In the *non-adaptive* setting, all tests must be designed in advance, whereas in the *adaptive* setting, one may design the next test based on the outcomes of all previous tests. We consider a fixed number k of defective items, and assume that the support set S is uniform over the subsets of $\{1, \dots, p\}$ with cardinality k.

This work was supported by the European Commission (ERC Future Proof), SNF (200021-146750 and CRSII2-147633), and 'EPFL Fellows' program (Horizon2020 665667).

The focus on this paper is on relaxed recovery criteria compared to the usual exact recovery criterion. Specifically, we assume that the decoder outputs a list $\mathcal{L} \subseteq \{1, \dots, p\}$ of cardinality $L \geq k$ (possibly much larger than k), and all we require is that it contains at least a proportion $1 - \alpha^*$ of S; hence, the error probability is given by

$$P_{\mathrm{e}}(L, \alpha^*) := \mathbb{P}\big[|\mathcal{L} \cap S| < (1-\alpha^*)k\big]. \tag{2}$$

Note that this is distinct from the decoders considered in previous works that directly output an estimate $\hat{S}$ of S. List decoding can be of direct interest, for instance, in fault detection, if one is content with discarding some number of non-defective items as long as most of the defective items are also discarded. Another use of list decoding is as the first step in a two-step procedure, where the second step simply tests the items in the list one-by-one.

In the special case $\alpha^* = 0$, corresponding to full recovery with $S \subseteq \mathcal{L}$, we abbreviate $P_{\mathrm{e}}(L, \alpha^*)$ as $P_{\mathrm{e}}(L)$. In this case, our setup can be viewed as an analog of list decoding for channel coding, which has received considerable attention in the information theory literature (e.g., see [6, 7] and the references therein).

In summary, the main notations used throughout the rest of the paper are the number of items p, number of defectives k, number of tests n, list size L, defective set S, allowed error fraction α^*, and error probability P_{e}.

1.1. Previous Work

The information-theoretic limits of group testing have been studied for decades (e.g., see [8, 9]), and have recently become increasingly well-understood [10–15]. In particular, when the number of defectives scales as $k = O(p^{1/3})$, it is known that the number of tests n^* required to exactly identify the defective set satisfies [14]

$$n^* = \left(k \log_2 \frac{p}{k} \right)(1 + o(1)) \tag{3}$$

in the non-adaptive setting, while in the adaptive setting the same is true with $k = O(p^\theta)$ for any $\theta \in (0,1)$ [16].

Various partial recovery and list decoding results have also appeared previously. It was shown in [14] that if $L = k$ in the above setup (i.e., regular decoding instead of list decoding), then having $\alpha^* > 0$ amounts to reducing (3) by at most a multiplicative factor $1 - \alpha^*$. On the plus side, the bound

then becomes valid when $k = O(p^\theta)$ for *any* $\theta \in (0,1)$, as opposed to only $\theta \le \frac{1}{3}$.

List decoding has been used previously in group testing for various purposes [4, 17–19], including as an intermediate step for standard group testing [18], and for combating adversarial noise [17]. The work most relevant to ours is [20], which shows that when k and L behave as $O(1)$, the required number of tests remains of the form (3); see also [19] for a more stringent performance criterion related to the COMP algorithm [21].

1.2. Contributions

The study of list decoding performance limits with general scalings $k = o(p)$ is non-trivial compared to $k = O(1)$, and is the focus of this paper.

Our main findings reveal that the gain is limited in several settings, as exemplified in the following example: *If $k = O(p^{1/3})$ and we set $L = Ck$ for arbitrarily large C, $\alpha^* \in (0,1)$, and only require an arbitrarily small success rate of ϵ (i.e., $P_{\mathrm{e}}(L,\alpha^*) \le 1-\epsilon$), then the required number of tests still only improves on* (3) *by at most a multiplicative factor of $1-\alpha^*$ asymptotically*.

On the other hand, we give examples where both partial recovery and list decoding can strictly reduce the asymptotic number of tests; in the preceding example, list decoding gives an improvement whenever $L = \Theta(k^{1+\delta})$ for some $\delta > 0$, with the difference vanishing in the limit as $\delta \to 0$.

2. MAIN RESULTS

Here we present and discuss our main results; the proofs are deferred to Section 3.

2.1. Negative Result

As outlined above, the following theorem reveals a broad range of scenarios where partial recovery and list decoding do not help significantly.

Theorem 1. *Fix $\alpha^* \in (0,1)$, and suppose that $k \le L$ and $L = o(p)$. Then in order to obtain $P_{\mathrm{e}}(L,\alpha^*) \not\to 1$ as $p \to \infty$, it is necessary that*

$$n \ge (1-\alpha^*)\left(k \log_2 \frac{p}{L}\right)(1-o(1)) \tag{4}$$

even in the case of adaptive tests.

By a comparison with (3), which corresponds to the standard group testing setting with $k = O(p^{1/3})$, Theorem 1 can be viewed negatively for three reasons. First, it shows that allowing a fraction $\alpha^* k$ of the defectives to be missed gives at most an improvement by $1-\alpha^*$. Second, list decoding is of limited help; in particular, if $L = O(k)$, even with a large implied constant, then there is no gain asymptotically. Finally, even if one only seeks success a small fraction of the time, say 0.01, the result remains unchanged. The last of these is commonly known as the *strong converse*.

On the other hand, if L is significantly larger than k, e.g., $k = O(p^\theta)$ and $L = O(p^{\theta'})$ with $\theta' > \theta$, then (4) indicates that a constant-factor reduction in n may be possible due to list decoding; we will shortly see that this is provably true.

Similar observations on partial recovery were made in [14] in the absence of list decoding, i.e., with $L = k$. Moreover, setting $\alpha^* = 0$ corresponds to the problem studied in [22, 23], namely, finding a subset of non-defective items of a given size. The scaling regimes considered in [22, 23] correspond to a large list size $L = \Theta(p)$, and in this case, the potential gains are much more significant. Our results reveal the limitations of using smaller list sizes $L = o(p)$.

2.2. Positive Results

Since Theorem 1 gives a necessary condition, we cannot use it to conclude that partial recovery and list decoding can reduce the number of tests. We proceed by showing that both of these can help in some cases.

The fact that partial recovery can help (even when $L = k$) follows directly from the results of [14] and [15]. The former showed that with Bernoulli tests, the limit (3) can be achieved for any $\theta \in (0,1)$ whenever $\alpha^* > 0$. In contrast, [15] shows that the best possible coefficient to $k \log_2 \frac{p}{k}$ with Bernoulli testing tends to ∞ as $\theta \to 1$. Hence, we conclude that partial recovery provides an arbitrarily large gain.

We are not aware of any existing results revealing list decoding to help when $L = o(p)$, and in fact [20] showed that the answer is negative when k and L are $O(1)$. The following theorem reveals cases where the answer is affirmative, and moreover, where Theorem 1 is tight. We focus primarily on the non-adaptive setting with $\alpha^* = 0$, but also comment on allowing adaptivity or $\alpha^* > 0$.

Theorem 2. *Suppose that k and L satisfy $k \le L$, $L = o(p)$, and $k = O\big(\sqrt{\frac{p}{L}}\big)$. Then there exists a non-adaptive group testing procedure such that $P_{\mathrm{e}}(L) \to 0$ as $p \to \infty$ with*

$$n \le \left(k \log_2 \frac{p}{L}\right)(1+o(1)). \tag{5}$$

Moreover, if adaptivity is allowed, or if we only require $P_{\mathrm{e}}(L,\alpha^*) \to 0$ for some $\alpha^* \in (0,1)$, then this remains true even without the condition $k = O\big(\sqrt{\frac{p}{L}}\big)$.

Observe that if $k = O(p^\theta)$ and $L = O(p^{\theta'})$ with $\theta' > \theta$, then (3) and (5) have the same scaling laws, but the latter has a strictly smaller implied constant. A key implication of this observation is that *the coefficient to $k \log_2 \frac{p}{k}$ with list decoding can be strictly less than one*, even without partial recovery. This is perhaps unsurprising when $L = \Theta(p)$ (e.g., $L = p$ trivially gives zero error probability without needing any tests), but it is non-trivial in the sublinear regime.

We note that in order to satisfy the above condition $k = O\big(\sqrt{\frac{p}{L}}\big)$, it suffices that both k and L behave as $O(p^{1/3})$.

2.3. Noisy Tests

We have focused on the noiseless case for clarity, but analogous results follow in noisy scenarios using analogous tech-

niques to [13, 14, 24]. In particular, in the case of symmetric Bernoulli noise, where each test outcome is independently flipped with probability ρ, Theorems 1 and 2 remain true when the right-hand sides of (4)–(5) are divided by $1-H_2(\rho)$, where H_2 is the binary entropy function in bits.

3. PROOFS

3.1. Proof of Theorem 1

Regardless of adaptivity, the list $\mathcal{L}$ at the output of the decoder can be viewed as a function of the measurement vector $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(n)})$, and we make this explicit by writing $\mathcal{L}(\mathbf{y})$. We also let $\mathcal{S}$ denote the set of subsets of $\{1, \dots, p\}$ of cardinality k, and let $P_{\mathbf{Y}|s}$ denote the distribution of $\mathbf{Y}$ given $S = s$ for some $s \in \mathcal{S}$, which is deterministic if the tests are designed deterministically.

It follows that the probability of correct recovery, denoted by $P_{\mathrm{c}}(L, \alpha^*) := 1 - P_{\mathrm{e}}(L, \alpha^*)$, can be written as

$$P_{\mathrm{c}}(L,\alpha^*) = \sum_{s\in\mathcal{S}} \frac{1}{\binom{p}{k}} \sum_{\mathbf{y}} P_{\mathbf{Y}|s}(\mathbf{y}|s) \times \mathbb{1}\{|\mathcal{L}(\mathbf{y}) \cap s| \geq (1-\alpha^*)k\} \quad (6)$$

$$\leq \frac{1}{\binom{p}{k}} \sum_{\mathbf{y}} \sum_{s\in\mathcal{S}} \mathbb{1}\{|\mathcal{L}(\mathbf{y}) \cap s| \geq (1-\alpha^*)k\}. \quad (7)$$

We proceed by bounding the quantity

$$N(\mathbf{y}) := \sum_{s\in\mathcal{S}} \mathbb{1}\{|\mathcal{L}(\mathbf{y}) \cap s| \geq (1-\alpha^*)k\} \quad (8)$$

for a fixed value of $\mathbf{y}$, which is simply the number of defective sets of cardinality k that overlap with $\mathcal{L}(\mathbf{y})$ in at least $(1-\alpha^*)k$ positions. Since $|\mathcal{L}(\mathbf{y})| = L$, a simple counting argument gives

$$N(\mathbf{y}) \leq \sum_{d=0}^{\alpha^* k} \binom{p-L}{d}\binom{L}{k-d}, \quad (9)$$

where d represents the number of items in the defective set that are not included in $\mathcal{L}(\mathbf{y})$. We can upper bound (9) as follows:

$$N(\mathbf{y}) \leq k \max_{d\in\{0,\dots,\alpha^* k\}} \binom{p-L}{d}\binom{L}{k-d} \quad (10)$$

$$\leq k \max_{\alpha\in[0,\alpha^*]} \binom{p-L}{\alpha k}\binom{L}{(1-\alpha)k}. \quad (11)$$

Applying $\log\binom{A}{B} \leq B\log\frac{Ae}{B}$, we obtain

$$\log N(\mathbf{y}) \leq \log k + \max_{\alpha\in[0,\alpha^*]} \alpha k \log\frac{(p-L)e}{\alpha k} + (1-\alpha)k\log\frac{Le}{(1-\alpha)k} \quad (12)$$

$$= \log k + \Bigg(\max_{\alpha\in[0,\alpha^*]} \alpha k \log\frac{p-L}{k} + (1-\alpha)k\log\frac{L}{k}\Bigg)(1+o(1)) \quad (13)$$

$$= \log k + \Bigg(\alpha^* k\log\frac{p-L}{k} + (1-\alpha^*)k\log\frac{L}{k}\Bigg)(1+o(1)) \quad (14)$$

$$\leq \Bigg(\alpha^* k\log\frac{p}{k} + (1-\alpha^*)k\log\frac{L}{k}\Bigg)(1+o(1)), \quad (15)$$

where (13) follows since the $\log\frac{p-L}{k}$ term dominates $\log\frac{e}{\alpha}$ and $\log\frac{e}{1-\alpha}$ due to $L = o(p)$,[1] (14) again uses $L = o(p)$, and (15) follows since $\log k$ is dominated by $k\log\frac{p-L}{k}$.

Substituting (15) into (7) and noting that the number of $\mathbf{y}$ sequences is 2^n, we find that the following condition is necessary for $P_{\mathrm{c}}(L,\alpha^*) \not\to 0$ as $p\to\infty$:

$$n \geq \log_2\binom{p}{k} - \Bigg(\alpha^* k\log_2\frac{p}{k} + (1-\alpha^*)k\log_2\frac{L}{k}\Bigg)(1+o(1)) \quad (16)$$

$$= \Bigg((1-\alpha^*)k\log_2\frac{p}{k} - (1-\alpha^*)k\log_2\frac{L}{k}\Bigg)(1+o(1)) \quad (17)$$

$$= \Bigg((1-\alpha^*)k\log_2\frac{p}{L}\Bigg)(1+o(1)), \quad (18)$$

where follows since $\log_2\binom{p}{k} = \big(k\log_2\frac{p}{k}\big)(1+o(1))$ when $k = o(p)$. This concludes the proof.

3.2. Proof of Theorem 2

We arbitrarily partition the items $\{1,\dots,p\}$ into $\frac{pk}{L}$ groups of size $\frac{L}{k}$,denoted by $G_1,\dots,G_{pk/L}$. We then perform a "two-level" group testing procedure, where items in a common group are always tested together, and another group testing design for $p' = \frac{pk}{L}$ items is used to design a procedure that treats each group as a single item.

Since the original problem has at most k defective items, there are at most k groups containing one or more defective items; we refer to these as the *defective groups*. Using the results of [14, 15], we know that the smallest set (of groups) consistent with the observations will equal the true set of defective groups with probability approaching one, provided that (i) a Bernoulli testing design is used with $\big(k\log_2\frac{p'}{k}\big)(1+o(1))$ tests, and (ii) we have $k = O((p')^{1/3})$. Substituting the definition of p' into (i) yields the condition in (5), and rearranging (ii) gives $k = O\big(\sqrt{\frac{p}{L}}\big)$, as desired.

It remains to transfer the preceding guarantee to the list decoding guarantee for our setup. To do this, we simply use an algorithm that first identifies the defective groups by choosing the smallest set consistent with the observations, and then lets the list decoder output be the union of those groups. Since there are at most k defective groups, and their

[1] One also needs to consider the possibility of $\alpha\to 0$, but since $L = o(p)$ it is easily seen that the maximizing α is bounded away from zero.

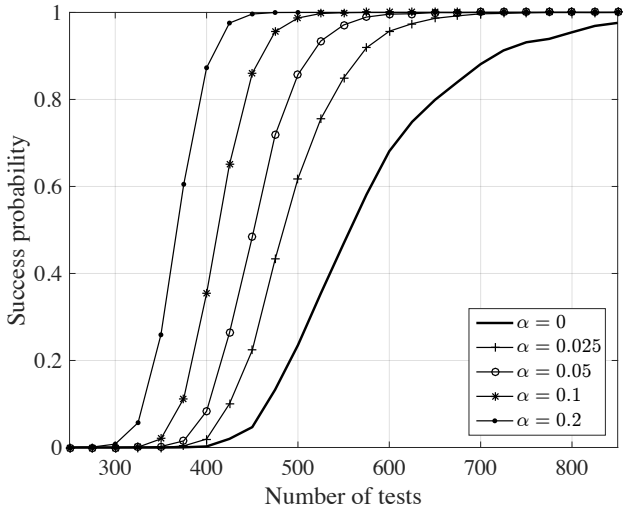

Fig. 1: Empirical performance of the SCOMP algorithm [26] with various partial recovery levels and non-list decoding ($L = k$).

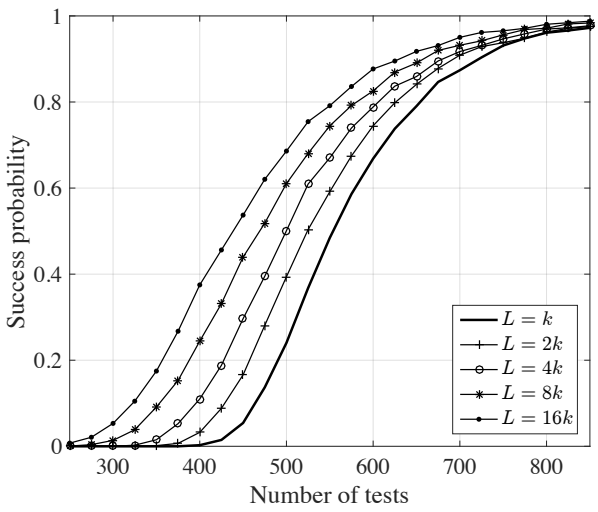

Fig. 2: Empirical performance of the SCOMP algorithm [26] combined with grouping, with various list-decoder list sizes and a full recovery requirement ($\alpha^* = 0$).

size is $\frac{L}{k}$, the overall list size is at most L. We can increase the list size to exactly L by arbitrarily adding additional items, which cannot reduce the error probability.

In the adaptive case, the claim follows by forming the groups of size $\frac{L}{k}$ in the same way, but using Hwang's adaptive algorithm [25] in the second step, which also succeeds with probability approaching one when $n = \big(k \log_2 \frac{p'}{k}\big)(1+o(1))$. In the case that we allow for partial recovery with some $\alpha^* \in (0, 1)$, the claim follows again analogously to the above, noting from [14] that the above condition $k = O((p')^{1/3})$ is not needed regardless of how small α^* is.

Note that the above grouping approach can be used to convert guarantees for any group testing algorithm into guarantees for the list decoding setting.

4. EXPERIMENTS

In this section, we provide some numerical experiments indicating the extent to which partial recovery and list decoding help in practice, and supporting the results given in Section 2. We focus on the sequential combinatorial orthogonal matching pursuit (SCOMP) algorithm described in [26], since it was seen to give state-of-the-art performance while being computationally efficient. We set the termination condition of the algorithm to be that the estimated set has cardinality k.

In each of the experiments below, we set $p = 4000$ and $k = 40$, and the empirical probability of correct decoding is computed by averaging over 5000 trials.

4.1. Partial Recovery

Figure 1 plots the empirical performance of the SCOMP algorithm with the exact recovery criterion, as well as four partial recovery criteria corresponding to different values of α^*. The behavior is qualitatively as predicted by Theorem 1: Increasing α^* shifts the performance curve to the left, but otherwise leaves it relatively unchanged.

On the other hand, the gap between $\alpha^* = 0$ and a small positive α^* appears to be slightly larger than expected. This suggests that although the asymptotics are similar for these two cases, allowing a small number of errors can improve the convergence rate and the non-asymptotic performance.

4.2. List Decoding

Figure 2 plots the empirical performance of the SCOMP algorithm with the exact recovery criterion, as well as four list decoding settings with $\alpha^* = 0$. To obtain the latter, we combine SCOMP with the grouping procedure described in Section 3.2. Once again, the behavior is as predicted, with larger list sizes giving a slightly improved performance but the same general behavior. The gain from increasing the list size appears to be particularly insignificant when the target success probability is close to one.

5. CONCLUSION

We have provided novel group testing performance bounds with list decoding and partial recovery. While our results provide a broad range of settings where these phenomena do not improve the asymptotic number of tests compared to the standard setting, we have also identified cases where they do. The grouping method used in the proof of Theorem 2 can be used to transfer guarantees from the standard setting to the list decoding setting, for any adaptive or non-adaptive group testing algorithm.

An interesting direction for future research is to further investigate noisy settings, where one might expect less stringent recovery criteria to be particularly important. The study of convergence rates to the asymptotic limit would also be of interest, since the first example in Section 4 suggests that it may be faster for partial recovery.

6. REFERENCES

[1] A. Fernández Anta, M. A. Mosteiro, and J. Ramón Muñoz, "Unbounded contention resolution in multiple-access channels," in *Distributed Computing*. Springer Berlin Heidelberg, 2011, vol. 6950, pp. 225–236.

[2] R. Clifford, K. Efremenko, E. Porat, and A. Rothschild, "Pattern matching with don't cares and few errors," *J. Comp. Sys. Sci.*, vol. 76, no. 2, pp. 115–124, 2010.

[3] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically," *ACM Trans. Database Sys.*, vol. 30, no. 1, pp. 249–278, March 2005.

[4] A. Gilbert, M. Iwen, and M. Strauss, "Group testing and sparse signal recovery," in *Asilomar Conf. Sig., Sys. and Comp.*, Oct. 2008, pp. 1059–1063.

[5] A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin, "One sketch for all: Fast algorithms for compressed sensing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, New York, 2007, pp. 237–246.

[6] G. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, 1968.

[7] N. Merhav, "List decoding: Random coding exponents and expurgated exponents," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6749–6759, Nov. 2014.

[8] M. Malyutov, "The separating property of random matrices," *Math. notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.

[9] M. B. Malyutov, "Search for sparse active inputs: A review," in *Inf. Theory, Comb. and Search Theory*, 2013, pp. 609–647.

[10] G. Atia and V. Saligrama, "A mutual information characterization for sparse signal processing," in *Int. Colloq. Aut., Lang. and Prog. (ICALP)*, Zürich, 2011.

[11] V. Tan and G. Atia, "Strong impossibility results for sparse signal processing," *IEEE Sig. Proc. Letters*, vol. 21, no. 3, pp. 260–264, March 2014.

[12] T. Laarhoven, "Asymptotics of fingerprinting and group testing: Tight bounds from channel capacities," *IEEE Trans. Inf. Forens. Sec.*, vol. 10, no. 9, pp. 1967–1980, 2015.

[13] J. Scarlett and V. Cevher, "Limits on support recovery with probabilistic models: An information-theoretic framework," 2016, accepted to *IEEE Trans. Inf. Theory*.

[14] ——, "Phase transitions in group testing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[15] M. Aldridge, "The capacity of Bernoulli nonadaptive group testing," 2015, http://arxiv.org/abs/1511.05201.

[16] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.

[17] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," in *Int. Symp. Found. Comp. Theory*, 2009, pp. 62–73.

[18] P. Indyk, H. Q. Ngo, and A. Rudra, "Efficiently decodable non-adaptive group testing," in *ACM-SIAM Symp. Disc. Alg. (SODA)*, 2010.

[19] A. G. D'yachkov, I. V. Vorobev, N. Polyansky, and V. Y. Shchukin, "Almost disjunctive list-decoding codes," *Prob. Inf. Transm.*, vol. 51, no. 2, pp. 110–131, 2015.

[20] A. G. D'yachkov, "Error probability bounds for the symmetrical model of the design of screening experiments," *Prob. Inf. Transm.*, 1982.

[21] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *Allerton Conf. Comm., Ctrl., Comp.*, Sep. 2011, pp. 1832–1839.

[22] A. Sharma and C. R. Murthy, "On finding a subset of healthy individuals from a large population," 2013, http://arxiv.org/abs/1307.8240.

[23] ——, "Computationally tractable algorithms for finding a subset of non-defective items from a large population," 2015, http://arxiv.org/abs/1502.04169.

[24] J. Scarlett and V. Cevher, "Converse bounds for noisy group testing with arbitrary measurement matrices," in *IEEE Int. Symp. Inf. Theory*, Barcelona, 2016.

[25] F. Hwang, "A method for detecting all defective members in a population by group testing," *J. Amer. Stats. Assoc.*, vol. 67, no. 339, pp. 605–608, 1972.

[26] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.