Blind Group Testing
Wasim Huleihel  Ohad Elishco  Muriel Médard

Abstract

The main goal in group testing is to recover a small subset of defective items from a larger population, while efficiently reducing the total number of (possibly noisy) required tests/measurements. The fundamental as well as the computational limits of the group testing problem become increasingly well-understood under the assumption that the input-output statistical relationship (or, channel law) is known to the recovery algorithm. Practical considerations, however, render this assumption inapplicable, and “blind” recovery/estimation procedures, independent of the input-output statistics, are desired. In this paper, we analyze the fundamental limits of a general noisy group testing problem when this relationship is unknown. Specifically, we propose an efficient scheme, based on the idea of separate-decoding of items (where each item is recovered separately), for which we derive sufficient conditions on the number of tests required for exact recovery. The difficulty in obtaining these conditions stems from the fact that we allow the number of defective items to grow with the population size, which in turn requires delicate concentration analysis of certain probabilities. We further show that in several scenarios, our proposed scheme achieves the same performance as that of the corresponding non-blind recovery algorithm (where the input-output statistics are known), implying that the proposed blind scheme is robust/universal. Finally, we also propose an inefficient combinatorial-based scheme (or, “joint-decoding”), for which we derive similar sufficient conditions.

I. INTRODUCTION

Linear and non-linear estimation problems are one of the most basic tasks in engineering, where given a known input distribution, an observed output, and an input-output statistical relationship, the goal is to infer the input. While the fundamental limits of various linear estimation problems are well-understood, characterizing the fundamental limits of non-linear estimation problems is much more challenging. Recently, there has been a large amount of work in engineering, machine learning, and statistics, on how to exploit sparsity in high dimensional data analysis. On the high level, the idea is to exploit the underlying sparsity of data (in some basis) to extract meaningful guarantees, using a number of samples that is typically much smaller than the dimension of the
original data. One example of a classical non-linear sparse estimation setting is the group testing problem, originally proposed by Dorfman in [1], with the aim of identifying syphilis infected soldiers (“defective items”) while dramatically reducing the number of required tests. Intuitively, when the number of defective items, \( d \), is much smaller than the population size, \( p \), one can conduct a smaller number of tests, \( n \), on pooled samples, instead of examining each blood sample individually. Each test outcome is negative if it includes no defective sample, and positive if it includes at least one defective sample. The problem is thus to identify the defective set using as few tests as possible. Although originated for medical purposes, group testing has many applications in computational biology [2], compressed sensing [3, 4], spectrum enforcement in cognitive radios [5, 6], communication protocols [7], pattern matching [8], and database systems [9].

Broadly speaking, several settings of the group testing problem can be formulated in accordance to the following perspectives:

- **Sparsity regime**: As in many problems, the sparsity regime, or, the way the defective set size \( d \) depends on the population size \( p \), plays a significant role. Specifically, when \( d \) is fixed, independent of \( p \), the information-theoretic limits of recovering the defective set are well-understood [10-17]. However, while the sub-linear and linear regimes are becoming increasingly well-understood, there are still gaps between known upper and lower bounds on the sample complexity [16-20], especially for the sub-linear regime which comes with significant technical challenges, as will be pointed out in the sequel.

- **Input model**: Generally speaking, there are two approaches to model the defective set. The first is the combinatorial approach, where it is assumed that any set of \( d \) items may be defective (worst-case input). The second is the probabilistic approach, where items are assumed to be drawn independently and identically distributed (i.i.d.) defective with some probability.

- **Pool design**: Similarly as for the input, both combinatorial pool design (i.e., construction of measurement matrices) to guarantee the recovery of the items of interest using a small number of tests (e.g., [21], and many references therein), and probabilistic pool design, can be considered.

- **Performance measure**: The reconstruction algorithm recovering the defective set might need to output the correct answer (“zero-error”), or, alternatively, with probability of error vanishing to zero as \( p \) increases.

- **On/Off-line design**: Both non-adaptive and adaptive pool design can be considered. Adaptive designs (see, e.g., [22-25]) use previously collected tests to guide the design and selection of
the next test in order to optimize the gain of new information. Non-adaptive designs, on the other hand, are predetermined/fixed, and naturally allow for parallel testing.

In this paper, we focus on non-adaptive probabilistic group testing, with vanishing error probability, from an information theoretic perspective. The information-theoretic limits of this problem have been deeply analyzed for the case $d = \Theta(1)$ in [10-17, 26], both for exact and partial recovery. Less is known, however, when $d$ is allowed to scale with $p$, e.g., $d = o(p)$, which comes with significant challenges as discussed in [16-18]. For $d = \Theta(p^\theta)$, where $\theta \in (0, 1)$, general achievability and converse bounds were derived in [17]. It was shown that for the noiseless (symmetric noise) model, if $\theta < \frac{1}{3}$ ($\theta \to 0$), the bounds are in fact tight.

On the practical level, several near-optimal practical algorithms for the noiseless model have been developed. To the best of our knowledge, the best known bounds under Bernoulli testing are achieved by the definite defectives (DD) algorithm proposed in [27, 28]. The same results can be achieved also by algorithms based on linear programming relaxation techniques [29]. For the symmetric noise model, the noisy combinatorial orthogonal matching pursuit (NCOMP) [30, 31] was shown to achieve optimal $O(k \log p)$ scaling, albeit with suboptimal constants. Other algorithms, without theoretical guarantees, can be found in [31-36]. More recently, the idea of separate decoding of items (where each item is recovered separately), proposed and analyzed in [11, 37], for $d = \Theta(1)$, was revisited and analyzed for general scaling in [18]. Specifically, for the noiseless and symmetric noise models, this approach is within a factor of $\log 2$ of the information theoretic optimum (i.e., optimal joint decoding) when $d = \Theta(1)$ [18, 37]. Furthermore, among other things, it was shown in [18] that the same behavior holds when $d = \Theta(p^\theta)$ and $\theta \to 0$, and that it achieves near-optimal theoretical guarantees for any $\theta \in (0, 1)$.

Based on the above introductory, it is evident that the performance of the group testing problem is typically evaluated by the following criteria:

- **Sample Complexity**: For a given error tolerance, the algorithm should require a small number of tests/samples, ideally matching the information-theoretic limits.
- **Computational Complexity**: The algorithm should run in time polynomial in the defective or population sets sizes.

The common restrictive assumption shared in all previously mentioned works is that the input-output relationship is perfectly known to the algorithm. In practice, however, these statistics are unknown and therefore the use of the joint-decoding or practical algorithms mentioned above is precluded. This issue is related to the major concern of model misspecification in high dimensional
data analysis. At the high level, the worry is that our algorithms should be able to tolerate the case when our assumed model and the true model do not perfectly coincide. This issue calls for the study of these so-called robust estimators, i.e., estimators which work in the presence of such “noise”. In other words, the algorithm should provide error guarantees even if the input-output statistics are unknown. The main question is then: “Do the statistical gains (achievable either by efficient or computationally expensive algorithms) persist in the presence of such noise?”.

To alleviate the problem described in the previous paragraph, an algorithm based on the empirical mutual information (MMI), not depending on the noise distribution, was proposed in [37]. It was shown that for $d = \Theta(1)$, this algorithm achieves the same performance achieved by the optimal joint and separate recovery algorithms, which are provided with the true input-output relationship. To the best of our knowledge, the general scaling case (where $d$ is allowed to scale with $p$) was not considered before. Similarly as in [18], this regime comes with significant challenges, giving additional requirements on the sample complexity, arising from certain large deviation events that dominate the error probability, which do not play a role when $d = \Theta(1)$. Accordingly, in this paper, we build upon [17, 18, 37] and study the group testing problem from an information-theoretic point of view, when the input-output statistics are unknown, and $d = o(p)$. Specifically, we propose two blind recovery algorithms—the first is practical and based on the previously mentioned idea of separate decoding of items, while the second is computationally expensive and based on a combinatorial search. For both schemes we derive sufficient conditions on the sample complexity (number of tests) required to achieve exact recovery. We show that when $d = \Theta(1)$ and $d = \Theta(p^\theta)$, with $\theta$ sufficiently small, our results achieve the same performance as if the input-output relationship was known and separate or optimal decoding schemes were used, implying that our strategies are universal/robust. Furthermore, for $d = \Theta(p^\theta)$ and $0 < \theta < 1$, we show that our schemes achieve order-optimal performance.

The main technical difficulty in our analysis stems from the discrepancy of the empirical measure induced by our proposed algorithms and the true distribution. In particular, it turns out that to analyze the probability of error associated with our algorithms, an explicit expression for the deviation probability of the plug-in mutual information estimator is needed. Unfortunately, standard concentration inequalities, such as the Bernstein’s inequality which was used in [18, Section II.D] for a similar task, cannot be applied here due to the simple fact that the plug-in estimator cannot be written as a sum of independent random variables. While concentration results for the plug-in estimates of the Shannon entropy and the mutual information are well-known in the literature
(see, e.g., [38, 39, Proposition 1], they are weak for our task and in some cases give sub-optimal concentration rates. To this end, we derive new concentration inequalities which are, more adequate for our task than [38, 39, Proposition 1], and allow us to prove the optimality of the proposed algorithms in the aforementioned special cases.

The paper is organized as follows. In the next section we establish notation conventions, and in Section III, we present our model and formalize the problem. Section IV presents our main results. Proofs are given in Section V. Finally, our conclusions appear in Section VI.

II. NOTATION

Throughout this paper, \((\Omega, \mathcal{F}, \mathbb{P})\) denotes a probability space, where \(\Omega\) is the sample space, \(\mathcal{F}\) is the \(\sigma\)-algebra, and \(\mathbb{P}\) is the underlying probability measure. Scalar random variables (RVs) over \((\Omega, \mathcal{F}, \mathbb{P})\) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters, e.g. \(X, x, \) and \(\mathcal{X}\), respectively. Random vectors of dimension \(n\) will be denoted by boldface letters, e.g., \(X\) denotes the random vector \((X_1, X_2, \ldots, X_n)\), and \(x = (x_1, x_2, \ldots, x_n)\) will designate a specific sample of \(X\). Random matrices of dimension \(n \times p\) will be also denoted by boldface letters, but then we explicitly emphasize the dimensions, e.g., \(X_{n \times p}\). When it is clear from the context, subscripts will be dropped. The set of all \(n\)-vectors with components taking values in a certain finite alphabet, will be denoted by the same alphabet superscripted by \(n\), e.g., \(\mathcal{X}^n\). The probability of an event \(A \in \mathcal{F}\) is denoted by \(\mathbb{P}(A)\). The cardinality of a finite set \(A\) will be denoted by \(|A|\), its complement will be denoted by \(A^c\). Finally, the indicator function of an event \(A\) will be denoted by \(\mathbb{1}\{A\}\).

The set of all probability distributions on an alphabet, say \(\mathcal{X}\), is denoted by \(\mathcal{P}(\mathcal{X})\). Probability mass functions (PMFs) are denoted by capital letters, such as \(P\) or \(Q\), with a subscript that identifies the random variable and its possible conditioning. We shall mainly consider joint distributions of two RVs \((X, Y)\) over the Cartesian product of two finite alphabets \(\mathcal{X}\) and \(\mathcal{Y}\). For brevity, we will denote any joint distribution, e.g. \(Q_{XY}\), simply by \(Q\), the marginals will be denoted by \(Q_X\) and \(Q_Y\), and the conditional distributions will be denoted by \(Q_{X|Y}\) and \(Q_{Y|X}\). The joint distribution induced by \(Q_X\) and \(Q_{Y|X}\) will be denoted by \(Q_X \times Q_{Y|X}\), and a similar notation will be used when the roles of \(X\) and \(Y\) are switched. For a sequence of random variable \(X\), if the entries of \(X\) are drawn in an independent and identically distributed (i.i.d.) manner according to \(P_X\), then for every \(x \in \mathcal{X}^n\) we have \(P_X(x) = \prod_{i=1}^n P_X(x_i)\) and we denote \(P_X(x) = P_X^n(x)\). In a similar
fashion, if for every \((x, y) \in \mathcal{X}^n \times \mathcal{Y}^m\) it holds that \(P_{Y|X}(y|x) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)\), then we denote \(P_{Y|X}(y|x) = P^\alpha_{Y|X}(y|x)\).

The expectation and variance operators will be denoted by \(\mathbb{E}\{\cdot\}\) and \(\text{Var}\{\cdot\}\), respectively, and when we wish to make the dependence on the underlying distribution \(Q\) clear, we denote it by \(\mathbb{E}_Q\{\cdot\}\) and \(\text{Var}_Q\{\cdot\}\). Information measures induced by the generic joint distribution \(Q_{XY}\), will be subscripted by \(Q\), for example, \(I_Q(X;Y)\) will denote the corresponding mutual information, etc. The relative entropy (or, Kullback-Liebler distance) between two probability measures \(P\) and \(Q\) on the same \(\sigma\)-algebra \(\mathcal{F}\) of subsets of the sample space \(\mathcal{X}\), with \(P \ll Q\) (i.e., \(P\) is absolutely continuous with respect to \(Q\)) is

\[
D_{\text{KL}}(P||Q) \triangleq \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.
\]  

(1)

The weighted divergence between two conditional distributions, \(P_{Y|X}\) and \(Q_{Y|X}\), with weight \(P_X\), is defined as

\[
D_{\text{KL}}(P_{Y|X}||Q_{Y|X}|P_X) \triangleq \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_X(x)P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)}.
\]  

(2)

The total variation distance between two discrete distributions \(P\) and \(Q\) defined over a common probability space \(\mathcal{X}\) is defined as

\[
\text{TV}(P, Q) \triangleq \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.
\]  

(3)

The Rényi entropy of order \(\alpha\), where \(\alpha \geq 0\) and \(\alpha \neq 1\), is defined as

\[
H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log \left( \sum_{i=1}^n P_X(x)^\alpha \right).
\]  

(4)

We also use Sibson’s definition of relative divergence \([40]\)

\[
I_\alpha(X;Y) \triangleq \min_{Q_Y} D_\alpha(P_{Y|X}||Q_Y|P_X),
\]  

(5)

where

\[
D_\alpha(P||Q) \triangleq \frac{1}{\alpha - 1} \log \left( \sum_{x \in \mathcal{X}} P^\alpha(x)Q^{1-\alpha}(x) \right),
\]  

(6)

is the Rényi divergence of order \(\alpha\), and \(D_\alpha(P_{Y|X}||Q_{Y|X}|P_X) \triangleq D_\alpha(P_X \times P_{Y|X}||P_X \times Q_{Y|X})\).

The set of all empirical distributions on a vector in \(\mathcal{X}^n\) is denoted by \(\mathcal{P}_n(\mathcal{X})\). For a given vector \(x \in \mathcal{X}^n\), we let \(\hat{P}_x \in \mathcal{P}_n(\mathcal{X})\) denote its empirical distribution, that is, the vector \(\{\hat{P}_x(x), x \in \mathcal{X}\}\), where \(\hat{P}_x(x)\) is the relative frequency of the letter \(x\) in the vector \(x\). For a given type \(P \in \mathcal{P}_n(\mathcal{X})\), the set \(T_n(P)\) is the type-class, that is, the set of all sequences \(x\) with type \(\hat{P}_x = P\). Similarly,
for a pair of vectors \((x, y)\), the empirical joint distribution will be denoted by \(\hat{P}_{xy}\), or simply by \(\hat{P}\), for short. All previously defined notation rules for regular distributions will also be used for empirical distributions. We use "\(\lor\)" and "\(\oplus\)" to denote the Boolean OR operation, and modulo-2 sum, respectively. Finally, we make use of the standard asymptotic notations \(O(\cdot), o(\cdot), \Theta(\cdot),\) and \(\Omega(\cdot)\).

### III. Problem Formulation

Among a population of \(p\) items, \(d\) unknowns are of interest. These \(d\) items represent the defective set denoted by \(D\). The main task in the group testing problem is to design a collection of tests (pool), which is used to detect the defective set, with the goal of reducing the number of required tests \(n\). This pool is described by a \(n \times p\) binary-valued matrix, whose \((i, j)\) entry is "1" if item \(j\)th is included in the \(i\)th designed test, and "0" otherwise. Accordingly, the outcome of the \(i\)th test is positive if and only if the \(j\)th item is a member of the test. Upon observing the output for a number of \(n\) tests, the goal is to recover the set \(D\). In this paper, we take a probabilistic approach and require that the average probability of error will be small. Also, the design of the pool will be generated randomly. We mention here that the group testing task can be formulated as a channel coding/decoding problem [16, Section III]. Accordingly, in the literature, notions which are customary in information theory are typically used also for the group testing problem: the pool/measurement matrix is in fact a “code”, the input-output relationship is a “channel”, and the recovery algorithm is a “decoder”. In the rest of this paper, we use these notions interchangeably.

We next establish some additional notation.

We assume that \(D\) is uniform on \(D\), which contain the \(\binom{p}{d}\) possible subsets of \(\{1, 2, \ldots, p\}\) of cardinality \(d\). For convenience, we will sometimes equivalently refer to a vector \(\beta \in \{0, 1\}^p\) whose \(j\)-th entry indicates whether or not item \(j\) is defective

\[
\beta_j \triangleq 1 \{j \in D\}.
\]

For each item, we associate a codeword of length \(n\), which is a binary sequence, denoted by \(X^n_j\) for \(j \in \{1, 2, \ldots, p\}\). The \(i\)th entry of \(X^n_j\) is 1 if the \(j\)th item is included in the \(i\)th test, and 0 otherwise. The pool (or, codebook) is a collection of all the codewords, namely, it is \(n \times p\) matrix denoted by \(X_{n \times p}\). The vector of \(n\) (possibly noisy) observations is denoted \(Y \in \{0, 1\}^n\). In the sequel, the \(i\)-th entry of \(Y\) is designated by \(Y_i\) and the \(i\)-th row of \(X_{n \times p}\) by \(X_i\). Given a subset \(S \subset \{1, 2, \ldots, p\}\) with cardinality \(|S|\), the matrix \(X_{n \times |S|}\) is an \(n \times |S|\) matrix formed from the columns indexed by \(S\).
We use the same measurement model described in [18]. Specifically, it is given by
\[ Y_i | X_i; \sim P_{Y|N(D, X_i)} \]  
where \( N(D, X_i) \triangleq \sum_{j=1}^{p} \mathbb{1} \{ j \in D \cap X_{i,j} = 1 \} \) denotes the number of defective items in the test. We denote the \( j \)-th column of \( X_{n \times p} \), by \( X_{:,j} \). It will be convenient to work with random variables that are implicitly conditioned on a fixed value of \( D \), say \( s = \{1, 2, \ldots, d\} \). Accordingly, we write \( P_{Y|X_s} \) in place of \( P_{Y|X_D} \) to emphasize that \( D = s \), and we define
\[ P_{X_s,Y}(x_s, y) \triangleq P_X^d(x_s) P_{Y|X_s}(y|x_s), \]  
\[ P_{X,Y}(x_s, y) \triangleq P_{X_s}^n \times d(x_s) P_{Y|X_s}^n(y|x_s). \]  
We consider partitions of the defective set \( s \in D \) into two sets \( s_{\text{dif}} \neq \emptyset \) and \( s_{\text{eq}} \). One can think of \( s_{\text{eq}} \) as corresponding to an overlap \( s \cap \bar{s} \) between the true set \( s \) and some incorrect set \( \bar{s} \), and \( s_{\text{dif}} \) corresponding to the indices \( s \setminus \bar{s} \) in one set but not the other. There are \( 2^d - 1 \) ways of performing such a partition. For a fixed set \( s \in D \) and a corresponding pair \( (s_{\text{dif}}, s_{\text{eq}}) \), we define
\[ P_{Y|X_{s_{\text{dif}}},X_{s_{\text{eq}}}}(y|x_{s_{\text{dif}}}, x_{s_{\text{eq}}}) \triangleq P_{Y|X_s}(y|x_s). \]  
Accordingly,
\[ P_{Y|X_{s_{\text{eq}}}}(y|x_{s_{\text{eq}}}) \triangleq \sum_{x_{s_{\text{dif}}}} P_X^\ell(x_{s_{\text{dif}}}) P_{Y|X_{s_{\text{dif}}},X_{s_{\text{eq}}}}(y|x_{s_{\text{dif}}}, x_{s_{\text{eq}}}) \]  
where \( \ell \triangleq |s_{\text{dif}}| \). We next describe three special cases which are captured by (7).

1) Noiseless model: In the noiseless setting, each test takes the form
\[ Y = \bigvee_{i \in D} X_i, \]  

namely, it is the Boolean sum of the items corresponding to the defective set.

2) Symmetric noise model: Each test takes the form
\[ Y = \left( \bigvee_{i \in D} X_i \right) \oplus W, \]  

where \( W \sim \text{Bernoulli}(\mu) \), for some \( \mu > 0 \). Note that contrary to the noiseless model, here, the outcome of a test can be “1” even though none of the items are included in the test.

3) Dilution model: This model was introduced in [16]. In this case, each test is given by
\[ Y = \bigvee_{i \in D} \mathcal{Z} (X_i \cdot 1_{i \in D}), \]  

where \( \mathcal{Z} \) represents the \( Z \)-channel model, i.e., it is a channel with binary input and binary output where the crossover \( 1 \rightarrow 0 \) occurs with probability \( \mu \), whereas the crossover \( 0 \rightarrow 1 \)
never occurs. Contrary to the symmetric noise model, in this model, we might have missed but no false alarms.

As mentioned before, we take a probabilistic approach (rather than combinatorial), and generate the test matrix randomly. In this paper, we will consider the \textit{i.i.d. ensemble}, or i.i.d. Bernoulli testing, where each item is placed in a given test independently with probability $P_X \triangleq \nu/d$ for some constant $\nu > 0$, which can be optimized. For the noiseless and symmetric noise models it is common and convenient to choose [15, 18]

$$\nu = \nu_{\text{sym}} \triangleq \left\{ \nu : \left( 1 - \frac{\nu}{d} \right)^d = \frac{1}{2} \right\} = \left[ 1 + o(1) \right] \cdot \log 2. \quad (13)$$

Given $X_{n \times p}$ and $Y$, a decoder forms an estimate of $\hat{D}(X_{n \times p}, Y)$ of $D$. In this paper, we focus on the the error probability as our performance measure, which is defined by

$$P_e \triangleq \Pr(\hat{D}(X_{n \times p}, Y) \neq D)$$

where the probability is evaluated w.r.t. the randomness of $D$, $X_{n \times p}$, and $Y$. Our main goal is to derive necessary conditions on $n$ and $d$ (as function of $p$) such that $P_e$ vanishes as $p \to \infty$. Specifically, we wish to find a sequence of values $n^*$, indexed by $p$, such that for all $\eta > 0$, we have $P_e \to 0$ when $n \geq n^*(1 + \eta)$.

It is a well-known fact that the maximum-likelihood (ML) decoder minimizes the above error probability. Naturally, this decoder requires the knowledge of the codebook distribution, and more importantly, the knowledge of the channel law. Furthermore, the computational complexity of the ML decoder is exponential, which precludes its use in practice. As was mentioned in the introduction, the main goal of this paper is to devise decoders which are both computationally efficient and robust. By robustness we essentially mean that they are independent of the actual channel (and hence on the probability law of the noise) but yet perform “well” in some sense. Such decoders are called universal if they achieve the same performance as the ML decoder.

In this paper, we will propose two blind recovery algorithms, capturing two sides of the problem. The first, which we actually focus on in this paper, is sub-optimal, but computationally efficient. As shall be seen in the sequel this decoder mimics an empirical version of the separate decoding of items scheme [11, 18]. The second algorithm, on the other hand, is comprised of a combinatorial search, and therefore computationally expensive, but in some cases provably achieves the performance of the ML decoder, and thus universal. We next describe these decoders.
Separate decoding: We next propose a family of decoders, termed separate decoding of items, which have linear computational complexity. Specifically, when the channel law is known, \( \hat{\beta}_j \) is a function of \( x_{\cdot,j} \) and \( y \) only, i.e., (see, e.g., [18])

\[
\hat{\beta}_j = \varphi_j(x_{\cdot,j}, y), \quad j = 1, 2, \ldots, p,
\]

for some set of function \( \{ \varphi_j \}_j \) defined on \( \{0, 1\}^n \times \{0, 1\}^n \to \mathbb{R} \). Since we are concerned with robust group testing the above decoder is inadequate for our task. Accordingly, we define the following decoder, for some \( \gamma > 0 \), and two given sequences \( x_{\cdot,j} \) and \( y \),

\[
\varphi_j(x_{\cdot,j}, y) \triangleq \mathbb{1}\left\{ \log \frac{\hat{P}_n(x_{\cdot,j}, y)}{\hat{P}_n(x_{\cdot,j}) \hat{P}_n(y)} > \gamma \right\}, \quad j = 1, 2, \ldots, p, \tag{14}
\]

for \( j = 1, 2, \ldots, p \), where \( \hat{P}_n(A) \) designates the Binomial empirical measure, namely, for a given binary sequences \( x, y \) of length \( n \),

\[
\frac{1}{n} \log \hat{P}_n(x, y) \triangleq \sum_{x,y\in\{0,1\}} \hat{P}_{x,y}(x,y) \log \hat{P}_{x,y}(x,y) \tag{15}
\]

\[
= -H(\hat{P}_{x,y}). \tag{16}
\]

For simplicity of notation we define the empirical information density as follows

\[
i_n(x_{\cdot,j}, y) \triangleq \log \frac{\hat{P}_n(x_{\cdot,j}, y)}{\hat{P}_n(x_{\cdot,j}) \hat{P}_n(y)}.
\]

Joint decoding: Fix a set \( s \in \mathcal{D} \) and a corresponding pair \( (s_{\text{dif}}, s_{\text{eq}}) \). Let

\[
i_n(x_{s_{\text{dif}}}; y|x_{s_{\text{eq}}}) \triangleq \log \frac{\hat{P}_n(x_{s_{\text{dif}}}, s_{\text{eq}}; y)}{\hat{P}_n(x_{s_{\text{eq}}}, y) \hat{P}_n(x_{s_{\text{dif}}})}. \tag{17}
\]

Fix some constants \( \gamma_1, \gamma_2, \ldots, \gamma_d \in \mathbb{R} \). The first decoder searches for the unique set \( s \in \mathcal{D} \) such that

\[
i_n(x_{s_{\text{dif}}}; y|x_{s_{\text{eq}}}) > \gamma_{|s_{\text{dif}}|} \tag{18}
\]

for all \( 2^d - 1 \) partitions \( (s_{\text{dif}}, s_{\text{eq}}) \) of \( s \) with \( s_{\text{dif}} \neq \emptyset \). An error occurs if no such \( s \) exists, if multiple exist, or if such a set differs from the true one. It is evident that similarly to the ML decoder, the above decoder suffers also from an exponential computational complexity, as it requires a search over all \( \binom{p}{d} \) possible sets of size \( d \).

Remark 1. We mention here that the above decoder is closely related to the empirical mutual information (MMI) decoder which goes through all \( \binom{p}{d} \) possible sets of size \( d \), and choose the most “likely” one. Specifically, given \( Y = y \),

\[
\hat{D} = \max_{|S|=d} \hat{i}_n(x_{n\times|S|}; y) \tag{19}
\]
where the maximization is over all subsets \( S \) of \( \{1, 2, \ldots, p\} \) with cardinality \(|S| = d\), and \( \hat{I}_n(\cdot; \cdot) \) is the empirical mutual-information,

\[
\hat{I}_n(x_{n \times |S|}; y) \triangleq \frac{1}{n} \log \frac{\hat{P}_n(x_{n \times |S|}; y)}{P_{X_{n \times |S|}}(x_{n \times |S|}) \hat{P}_n(y)}.
\]

(20)

*It turns out that analyzing (18) is more convenient.*

We emphasize that the above decoders are indeed independent of the actual channel law, and are only based on empirical joint statistics of the observations \( Y \) and the test matrix \( X_{n \times p} \). Also, in contrast to classical studies in the topic of universal decoding (see, e.g., [41, 42]), where it is customary assumed that the decoder designer knows that the channel belongs to some family of channels, here, we do not assume any such prior knowledge. We do assume that the input and output alphabets are known.

Finally, we mention that the forthcoming analysis will apply for any given choice of the defective set \( D \), due to the symmetry of the observation model (7), and the fact that test matrix \( X_{n \times p} \) is i.i.d. Therefore, throughout the rest of this paper we focus on the specific choice \( D = \{1, 2, \ldots, d\} \). In particular, we assume that item \#1 is defective, and we accordingly define

\[
P_{Y|X_1}(y|x_1) = P_{Y|X_1,\beta_1}(y|x_1, 1).
\]

In the following \( I(X_1; Y) \) refers to the mutual information evaluated w.r.t. \( P_X \times P_{Y|X_1} \). Similarly, note that by symmetry, the mutual information \( I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \) depends only on \(|s_{\text{dif}}|\) for each \((s_{\text{dif}}, s_{\text{eq}})\). As shall be seen later on, our main results will be expressed in terms of these mutual information quantities. The closed-form expressions of these quantities are known in some special cases, and we next mention a few of them for completeness. For the classical choice of \( \nu = \log 2 \), in the noiseless setting, it can be shown that [15, 43]

\[
I(X_1; Y) = \frac{(\log 2)^2}{d}[1 + o(1)]
\]

(21)

where asymptotic notation such \( \rightarrow, o(\cdot), \) etc., are w.r.t. \( k, p \to \infty \) with \( k = o(p) \). Also, for the symmetric noise model with parameter \( \mu \in (0, 1/2) \), it can be shown [43, 18] that for the same choice of \( \nu \),

\[
I(X_1; Y) = \left[\log 2 - H_2(\mu)\right] \frac{\log 2}{d}[1 + o(1)].
\]

(22)

We mention here that at least for the symmetric noise model the choice of \( \nu = \log 2 \) might be suboptimal and \( I(X_1; Y) \) may appear differently for the optimal choice of \( \nu \). Formulas for \( I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \) can be found in [17]. In the next section we present our main result on the performance of the separate decoder in (14), and then discuss the joint decoder in (18).
IV. MAIN RESULTS

A. Separate Decoding

In this subsection, we present our main results regarding the empirical separate decoding of items scheme in (14). The proofs appear in Section V. We start with the following result which provides an upper-bound on the error probability using the proposed separate decoding procedure in (14) for the general noisy group testing model described in Section III.

**Theorem 1.** For the general group testing with Bernoulli($\nu/d$) testing and separate decoding of items (14), we have

$$P_e \leq d \cdot \mathbb{P} \left[ \hat{i}_n(X_{i,1}, Y) \leq \gamma \right] + C \cdot (p - d) \cdot e^{-\gamma}$$

where the probability term in (23) is evaluated w.r.t. $P^n_X \times P^n_{Y|X_1}$.

An analogue of Theorem 1 in the case where the channel law is known can be found in [18, Theorem 1]. While the above statement resembles that of [18, Theorem 1], the proofs are quite different. In particular, the proof of the above result turns out to be more involved due to the discrepancy of the empirical measure induced by the decoder and the true distribution. Next, using Theorem 1 we would like to get some guarantees on the sample complexity $n$ needed to drive the error probability $P_e$ to zero. The following result is a consequence of Theorem 1.

**Theorem 2.** Consider the general group testing with Bernoulli($\nu/d$) testing and separate decoding of items (14). Suppose that there exists some function $\psi_n(\delta_1)$ such that

$$\mathbb{P} \left[ \hat{i}_n(X_{i,1}, Y) \leq (1 - \delta_1) \cdot \mathbb{E} \hat{i}_n(X_{i,1}, Y) \right] \leq \psi_n(\delta_1)$$

for some $\delta_1 \in (0, 1)$. Also, suppose that

$$n \geq \frac{\log \left[ \frac{1}{\delta_2^2} (p - d) \right]}{(1 - \delta_1) \cdot I(X_1; Y)}$$

and

$$d \cdot \psi_n(\delta_1) \to 0,$$  \hspace{1cm} (25b)

for some $\delta_2 \to 0$. Then, $P_e \to 0$, using $\gamma = \log \frac{\nu - d}{\delta_2^2}$.

In order to apply Theorem 2 on specific channels, it is evident that we need to derive an explicit expression for $\psi_n$. This simply means that we need to study the concentration of the empirical mutual-information density function $\hat{i}_n(X_{i,1}, Y)$. To get an explicit expression for $\psi_n$ in the known channel case, the Bernstein’s inequality was used in [18, Section II.D]. However, since $\hat{i}_n(X_{i,1}, Y)$
do not tensorize, or, in other words, cannot be represented in terms of summation of independent random variables, the use of Bernstein’s-like concentration inequalities is precluded. Nonetheless, concentration results for the plug-in estimates of the Shannon entropy and the mutual information, are well-known in the literature (see, e.g., [38, 39, Proposition 1], and more recently [44, Eqn. 88]). These inequalities are and typically obtained using bounded-difference type concentration inequalities [45]. These concentration results are, however, universal in the sense that the upper bounds are independent of the true distribution, and thus potentially weak for our task. We next propose two explicit expressions for \( \psi_n \); the first obtained using [38, 39, Proposition 1], and the second is proved using Chernoff’s inequality.

**Lemma 1.** The following hold:

1) For any \( \delta_1 > 0 \),

\[
P \left[ \hat{i}_n(X; Y) \leq (1 - \delta_1) \cdot \bar{E}_n \hat{i}_n(X; Y) \right] \leq \exp \left( -n \cdot \frac{[\delta_1 \cdot I(X_1; Y)]^2}{18 \log^2 n} \right). \tag{26}
\]

2) For any \( \rho \in \mathbb{R}_+ \), define

\[
h_{\rho}(P_{XY}) \triangleq \min_{Q_{XY}} \left[ \rho \cdot I_Q(X; Y) + D_{KL}(Q_{XY} \| P_{XY}) \right] \tag{27}
\]

\[
\geq \min_{Q_X} \left[ D_{KL}(Q_X \| P_X) + \rho \cdot I_{1/n}^{\rho}(Q_X \times P_{Y|X}) \right] \tag{28}
\]

where \( I_\alpha(X; Y) \) is defined in (5). Then, for any \( \delta_1 > 0 \), and \( \rho \in \mathbb{R}_+ \),

\[
P \left[ \hat{i}_n(X; Y) \leq (1 - \delta_1) \cdot \bar{E}_n \hat{i}_n(X; Y) \right] \leq n^{3/2} e^{-n \left[ h_{\rho}(P_{XY}) - (1 - \delta_1) \cdot \rho \cdot I(X_1; Y) - \Theta(\frac{1}{n}) \right]}. \tag{29}
\]

As will be seen later on the advantage of (26) over (29) lies in its simplicity. Indeed, to apply (26) one only needs to evaluate the mutual information \( I(X_1; Y) \) term as in (21)-(22). To apply (29), however, a relatively involved optimization problem should be solved, which might impose some difficulties in some cases. On the other hand, (29) is provably tighter than (26). Indeed, since (26) exhibits dependency on the squared value of \( I(X_1; Y) \), when \( d \) grows with \( p \), this dependency might cause a poor scaling of \( n \) as a function of \( d \). We will demonstrate later on that (29) resolves this issue. We next use Lemma 1 to get an explicit formula for \( \psi_n(\cdot) \), which in turn will imply certain conditions on the sample complexity \( n \) through condition (25b). Combining Theorem 1, Theorem 2, and Lemma 1, we obtain the following result.
Corollary 1. For the noisy group testing problem, we can achieve $P_e \to 0$ using (14) provided that

$$n \geq \log \left[ \frac{1 - \delta_2 (p - d)}{(1 - \delta_1) \cdot I(X_1; Y)} \right] (1 + \eta),$$

(30a)

$$n \geq \log \frac{d}{\min_{\rho \geq 0} \rho (P_{XY}) - (1 - \delta_1) \cdot \rho \cdot I(X_1; Y)} (1 + \eta),$$

(30b)

for some $\delta_1, \eta > 0$, and $\delta_2 \to 0$. Alternatively, (30b) can be replaced with

$$\frac{n}{\log^2 n} \geq \frac{\log d}{\frac{1}{2} \delta_1 \cdot I(X_1; Y)^2} (1 + \eta).$$

(31)

Proof of Corollary 1: We only need to show the conditions in (30b) and (31). These follow immediately by plugging (26) and (29) in (25b) and solving for $n$.

The above results are general and can be applied for any noisy group testing model in the form of (7), and general scaling of $d = o(p)$. In the next subsection, we consider a few examples, and compare our results with [18, Th. 2], where the known channel case was considered. As shall be seen, in some special cases, our proposed decoder achieves the same performance as if the channel was known.

B. Comparison and Examples

In the following, we provide a few examples illustrating the results presented in the previous subsection. To this end, we first briefly discuss the results of [18], where general achievability (lower-bounds) and converse (upper-bounds) results for the noisy group testing problem under separate decoding of items and known channel law were proposed. Unfortunately, the results in [18] are not tight in the sense that the upper and lower bounds on $n$ do not match in general. Tight results were shown, nonetheless, for the case where $d = \Theta(p^\theta)$ with $\theta \to 0$ (after taking $p \to \infty$). Comparing Theorem 2 with [18, Th. 2], we see that conditions (25a)-(25b) are in fact the same. The main difference, though, lies in the fact that the upper bound $\psi_n$ in our case and in [18, Th. 2] on the probability of deviation are different, which in turn, results in a different constraint on $n$ through (25b).

1) Fixed Defective Set Size: We first consider the case where $d = \Theta(1)$, namely, the cardinality of the defective set is a constant independent of the population size. When the channel law is known, the information theoretic limits of group testing are well-known [10-17]. Specifically, for the symmetric noise model (11) with $\mu \in [0, 1/2]$, we have

$$n^* = \frac{d \log \frac{p}{d}}{\log 2 - H_2(\mu)} [1 + o(1)].$$

(32)
Furthermore, it was proved in [37], that when the channel law is known, separate decoding of items achieves exact recovery with vanishing error probability provided that
\[ n \geq \frac{\log p}{I(X_1; Y)} \left[ 1 + o(1) \right]. \tag{33} \]

As noted in [18], in the noiseless setting with \( \nu = \log 2 \) we have \( I(X_1; Y) = \frac{\log^2 2}{d}(1 + o(1)) \), and thus in this case (33) matches (32) up to a constant factor of \( \log 2 \approx 0.7 \). Actually, the same phenomenon is true for the symmetric noise model as well. We next show that these results are true also when the channel law is unknown, and the decoder in (14) is used.

**Corollary 2.** For the noisy group testing problem, if \( d = \Theta(1) \), then \( P_e \to 0 \) using (14), provided that
\[ n \geq \frac{\log p}{I(X_1; Y)} \left[ 1 + o(1) \right] \tag{34} \]
for any \( \nu > 0 \).

**Proof of Corollary 2:** The statement follows immediately by noticing that when \( d = \Theta(1) \), the condition in (25a) dominates. To see this, note that the r.h.s. of (31) is a constant while the r.h.s. of (30a) grows with \( p \). Hence, we get the condition in (34) by choosing \( \delta_2 \to 0 \) sufficiently slowly.

Comparing Corollary 2 to (33) and [18, Corollary 6], we can conclude that our decoder is robust in the sense that it achieves exactly the same performance as the decoder that knows the channel. Accordingly, the performance of (14) matches (32) up to a constant factor of \( \log 2 \approx 0.7 \). The above result demonstrates that when \( d = \Theta(1) \) there is no loss induced by the fact that channel law is unknown. Indeed, the fact that \( d = \Theta(1) \) makes the estimation of the mutual information easier in terms concentration rates.

2) The \( \theta \to 0 \) case: For the noiseless and symmetric noise models it was shown in [18] that it is sufficient and necessary to have
\[ n \approx \frac{\log p}{I(X_1; Y)} \left[ 1 + o(1) \right] \]
in order to get \( P_e \to 0 \) when \( d = \Theta(p^\theta) \) with \( \theta \to 0 \) (which taken after letting \( p \to \infty \)). We next show that the same result is true in our setting as well, namely, in this case, (25a) dominates and thus we achieve exactly the same performance as if the channel law was known. Contrary to Corollary 2 this result is not a priori obvious, and in fact it requires the more delicate concentration result in (29).
Corollary 3. For the noiseless and symmetric noise group testing models, if \( d = \Theta(p^{\theta}) \) with \( \theta \to 0 \), then \( P_e \to 0 \) using (14), provided that
\[
n \geq \frac{\log p}{I(X_1; Y)} \left[ 1 + o(1) \right], \tag{35}\]
for \( \nu = \nu_{\text{sym}} \).

Similarly as in the previous subsection we may conclude that when \( \theta \) is sufficiently small the proposed decoder matches the performance of the optimal ML decoder up to a log 2 factor. We emphasize the standard upper bound in (26), and accordingly, the condition in (30a), are not suffice to obtain the above result. Indeed, the condition in (30a) boils down to \( n \approx \log p/[I(X_1; Y)]^2 \), and since typically \( I(X_1; Y) \approx O(1/d) \), when \( d \) grows with \( p \), this condition is strictly worse than the one in (35).

3) General Scaling: We next consider the general scaling case where \( d = \Theta(p^{\theta}) \) for \( \theta \in (0, 1) \) independent of \( p \). Unfortunately, in this case, the information-theoretic upper and lower bounds do not coincide [46, Subsection IV.F], either due to lack of strong concentration results or perhaps tighter converse bounds. The same is true for the efficient separate decoding of item scheme. Nonetheless, order-wise (exact constants aside) the upper and lower bounds do coincide. In particular, for the noiseless and symmetric noise models, it is shown in [18, 46] that it is suffice to have \( n \geq C_\theta \cdot (d \log p) \), where \( C_\theta \) is a constant which depends on \( \theta \) and the parameter of the channel (if exists), and might be different depending on whether optimal decoding or separate decoding are being used. In light of these results, here, we focus only on the general scaling of \( n \) with \( d \) and \( p \), and show that (14) achieves order-optimal performance, albeit with worse constants.

Corollary 4. For the noiseless and symmetric noise group testing models, with \( \nu = \nu_{\text{sym}} \), if \( d = \Theta(p^{\theta}) \) with \( \theta \in (0, 1) \), then \( P_e \to 0 \) using (14), provided that
\[
n \geq \Omega(d \log p). \tag{36}\]

C. Joint Decoding

In this subsection, moving computational considerations aside, we present some initial results concerning the combinatorial based decoder in (18). Generally speaking, the main ideas and conclusions of the following results are similar to those in Subsection IV-A. We start with the following result which is in parallel to Theorem 2, and has the same flavor as [46, Theorem 1], but with the similar important differences as in Theorem 1 (see the discussion right after the statement of Theorem 1).
Theorem 3. Consider the a general group testing with Bernoulli($\nu/d$) testing and the decoder in (18), and suppose that there exist some functions $\psi_n(\delta_1,\ell)$, for $\ell = 1, 2, \ldots, d$, such that

$$\mathbb{P}\left[\hat{i}_n(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}}) \leq (1 - \delta_1,\ell) \cdot \mathbb{E}\hat{i}_n(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}})\right] \leq \psi_n(\delta_1,\ell),$$

for some constants $\{\delta_1,\ell\}_{\ell = 1}^d \in (0, 1)^d$. Also, suppose that

$$n \geq \max_{(s_{\text{diff}}, s_{\text{eq}}): s_{\text{diff}} \neq \emptyset} \frac{\log \left[ \frac{d \cdot (p - d)}{|s_{\text{diff}}|} \right]}{(1 - \delta_1,|s_{\text{diff}}|) \cdot I(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}})},$$

for some $\delta_2 \to 0$. Then, $P_e \to 0$, using $\gamma_{|s_{\text{diff}}|} = \log \left[ \frac{d \cdot (p - d)}{|s_{\text{diff}}|} \right]$ in (18).

As for Theorem 1 (and Theorem 2), we notice that the error probability associated with the joint empirical decoder in (18), can be decomposed into two parts – the first term corresponds to the true set failing the threshold test, and the second term corresponds to some incorrect set passing the threshold test. Accordingly, the conditions in (38a) and (38b) ensure that these contributions are asymptotically small. As before, to apply the above result we need to derive explicit expressions for $\{\psi_n(\delta_1,\ell)\}_{\ell = 1}^d$. This can be done in a similar fashion to Lemma 1. Specifically, for simplicity of demonstration we provide here an upper bound which is in the spirit of (26). A tighter bound, similar to (29) can be derived in the same fashion, however, since the obtained result will be at least as difficult to evaluate as (29) we opted to focus on the following bound. The proof is similar to Lemma 1 and thus relegated.

Lemma 2. For any $\{\delta_1,\ell\}_{\ell = 1}^d \subseteq (0, 1)$,

$$\mathbb{P}\left[\hat{i}_n(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}}) \leq 1 - \delta_1,\ell \cdot \mathbb{E}\hat{i}_n(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}})\right] \leq \exp \left( -n \frac{[I(X_{s_{\text{diff}}}; Y_1|X_{s_{\text{eq}}})]^2 \delta_1^2}{18 \log_2^2 n} \right).$$

As in the previous subsection, depending on the scaling of $d$ with $p$, in some cases the condition in (38a) dominates, while in other cases (38b) do. As before, when the defective set size is fixed, $d = \Theta(1)$, the condition in (38a) dominates, similarly to Corollary 2.

Corollary 5. For the noisy group testing problem, if $d = \Theta(1)$, then $P_e \to 0$ using (18), provided that

$$n \geq \max_{(s_{\text{diff}}, s_{\text{eq}}): s_{\text{diff}} \neq \emptyset} \frac{\log \left[ \frac{d \cdot (p - d)}{|s_{\text{diff}}|} \right]}{(1 - \delta_1,|s_{\text{diff}}|) \cdot I(X_{s_{\text{diff}}}; Y|X_{s_{\text{eq}}})},$$

for any $\nu > 0$. 

March 14, 2018 DRAFT
In fact, it was shown in [16] that (40) is also a necessary condition when the channel is known. This of course implies that our decoder is universal, namely, it achieves the performance of the ML decoder.

V. PROOFS OF MAIN RESULTS

A. Proof of Theorem 1

Correct decoding means that the $d$ defective items pass the threshold test in (14), and the rest $(n-d)$ non-defective items fails the same test. The union bound then gives

$$P_e \leq d \cdot \mathbb{P} \left[ \hat{t}_n(X_{-1}, Y) \leq \gamma \right] + (p - d) \cdot \mathbb{P} \left[ \hat{t}_n(X_{-1}, Y) > \gamma \right].$$  \hspace{1cm} (41)

We next upper-bound the second term at the r.h.s. of (41). To this end, we will use the following lemma whose proof is relegated to Appendix A. First, for a given $\varepsilon > 0$, define

$$A_{\varepsilon, n} \triangleq \{ Q_{XY} \in \mathcal{P}_n(X \times Y) : D_{\text{KL}}(Q_{XY} || P_{XY}) \leq \varepsilon \}$$

where the dependency of $A_{\varepsilon, n}$ on $P_{XY}$ is implicit. We have the following result.

**Lemma 3.** For any $\varepsilon > 0$,

$$|A_{\varepsilon, n}| \leq 8 \cdot \left( (2n\sqrt{2\varepsilon})^3 + 1 \right).$$

Markov inequality implies that

$$\mathbb{P} \left[ \hat{t}_n(X_{-1}, Y) > \gamma \right] \leq e^{-\gamma} \cdot \mathbb{E} \exp \left[ \hat{t}_n(X_{-1}, Y) \right]$$

$$= e^{-\gamma} \sum_{x,y} P^n_X(x) P^n_Y(y) \frac{\hat{P}_n(x,y)}{P_n(x) P_n(y)}. \hspace{1cm} (42)$$

(43)

Recall the fact that for every $x \in X^n$,

$$|\mathbb{T}(\hat{P}_x)| = 2^{nH(\hat{P}_x) - \frac{1}{2} \log n + O(1)}.$$  \hspace{1cm} (44)

Expressing the summands in (43) in terms of types, we get

$$\sum_{x,y} P^n_X(x) P^n_Y(y) \frac{\hat{P}_n(x,y)}{P_n(x) P_n(y)} = \sum_{x,y} 2^{-n \cdot [D_{\text{KL}}(\hat{P}_x || P_X) + D_{\text{KL}}(\hat{P}_y || P_Y) + H(\hat{P}_x,y)]}$$

$$= \sum_{Q_{XY} \in \mathcal{P}_n} |\mathbb{T}(Q_{XY})| 2^{-n \cdot [D_{\text{KL}}(Q_X || P_X) + D_{\text{KL}}(Q_Y || P_Y) + H(Q_{XY})]}$$

$$= C' \cdot \sum_{Q_{XY} \in \mathcal{P}_n} \frac{1}{n^{3/2}} 2^{nH(Q_{XY})} 2^{-n \cdot [D_{\text{KL}}(Q_X || P_X) + D_{\text{KL}}(Q_Y || P_Y) + H(Q_{XY})]}$$

$$= C' \cdot \sum_{Q_{XY} \in \mathcal{P}_n} 2^{-n \cdot [D_{\text{KL}}(Q_X || P_X) + D_{\text{KL}}(Q_Y || P_Y) + H(Q_{XY})]}$$
where $C' > 0$ is a constant. In the following we estimate the last summation term. For any $\varepsilon > 0$, consider the sets $\{A_{t,\varepsilon, n}\}_{t=1}^{\infty}$. We may write
\[
\sum_{Q_{X Y} \in P_n} 2^{-n[D_{KL}(Q_X \| P_X) + D_{KL}(Q_Y \| P_Y)]} \leq \sum_{t=1}^{\infty} |A_{t,\varepsilon, n}| \leq |A_{(t-1),\varepsilon, n}| 2^{-n2(t-1)\varepsilon}.
\]
This is due to the fact that over the set $|A_{t,\varepsilon, n}| - |A_{(t-1),\varepsilon, n}|$, the minimal value $D_{KL}(Q_X \| P_X)$ and $D_{KL}(Q_Y \| P_Y)$ can take is $(t-1)\varepsilon$. Clearly, we have $|A_{t,\varepsilon, n}| \geq |A_{(t-1),\varepsilon, n}|$, and thus
\[
\sum_{Q_{X Y} \in P_n} 2^{-n[D_{KL}(Q_X \| P_X) + D_{KL}(Q_Y \| P_Y)]} \leq \sum_{t=1}^{\infty} |A_{t,\varepsilon, n}| 2^{-n2(t-1)\varepsilon}.
\]
Lemma 3 now gives
\[
\sum_{x,y} P_n^m(x)P_n^m(y) \frac{\hat{P}_n(x,y)}{\hat{P}_n(x)\hat{P}_n(y)} = 64C' \sum_{t=1}^{\infty} (2nt\varepsilon)^2 2^{-2n(t-1)\varepsilon} + 8C' \sum_{t=1}^{\infty} 2^{-2n(t-1)\varepsilon}. \tag{45}
\]
Taking $\varepsilon = \frac{1}{n}$, we readily see that the r.h.s. of (45) is converging to some constant $C$, which concludes the proof.

B. Proof of Theorem 2

We first study the expectation of the empirical density. It is well-known that (see, e.g., [47]) that
\[
I(X_1; Y) - \frac{c_1}{n} \leq \mathbb{E} \hat{i}_n (X_{1,1}, Y) \leq I(X_1; Y) + \frac{c_2}{n} \tag{46}
\]
where $c_1, c_2 > 0$ are absolute constants. Setting $\gamma = \log \frac{p-d}{\delta_2}$ in Theorem 1, we readily get
\[
P_\varepsilon \leq d \cdot \mathbb{P} \left[ \hat{i}_n (X_{1,1}, Y) \leq \log \frac{p-d}{\delta_2} \right] + C \cdot \delta_2.
\]
Now, if
\[
n(1-\delta_1) \cdot \left[ I(X_1; Y) - \frac{c_1}{n} \right] \geq \log \frac{p-d}{\delta_2}, \tag{47}
\]
then by (46) we also have that
\[
(1-\delta_1) \cdot \mathbb{E} \hat{i}_n (X_{1,1}, Y) \geq \log \frac{p-d}{\delta_2}, \tag{48}
\]
so we obtain
\[
\mathbb{P} \left[ \hat{i}_n (X_{1,1}, Y) \leq \log \frac{p-d}{\delta_2} \right] \leq \mathbb{P} \left[ \hat{i}_n (X_{1,1}, Y) \leq (1-\delta_1) \cdot \mathbb{E} \hat{i}_n (X_{1,1}, Y) \right] \leq \psi_n (\delta_1).
\]
Accordingly, we get that $P_\varepsilon \leq d \cdot \psi_n (\delta_1) + C \cdot \delta_2$, and hence the result is followed by the assumption that $\delta_2 \to 0$ and (25b). Note that (47) is actually equivalent to (25a) since it can be written as
\[
n \geq \frac{\log \frac{p-d}{\delta_2}}{(1-\delta_1) \cdot I(X_1; Y)}, \tag{49}
\]
where $\delta_2' = e^{-c_1(1-\delta_1)} \cdot \delta_2$, and $\delta_2' \to 0$ as $\delta_2 \to 0$. 

March 14, 2018 DRAFT
C. Proof of Lemma 1

Since \( \hat{i}_n(X_{:,1}, Y) \) cannot factorize as a sum of independent random variables, we need another machinery to desired concentration result. In fact, concentration results for maximum-likelihood entropy and mutual information estimates are well-known in the literature (see, e.g., [38, 39, Proposition 1], and more recently [44, Eqn. 88]). The idea is to show that \( \hat{i}_n(X_{:,1}, Y) \) satisfy the so-called bounded difference property and then apply Azuma’s or McDiarmid’s inequalities. Let

\[
C_{\text{mean}} \triangleq \frac{1}{n} \mathbb{E} \hat{i}_n(X_{:,1}, Y).
\]

We use the following proposition [38, Proposition 1].

**Proposition 1.** Let \( X_1, \ldots, X_n \) be independent random variables defined on a common probability space \( \mathcal{X} \), and assume that \( \hat{F} : \mathcal{X}^n \rightarrow \mathbb{R} \) satisfies

\[
\sup_{x_1, \ldots, x_n, x'_j \in \mathcal{X}} \left| \hat{F}(x_1, \ldots, x_n) - \hat{F}(x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_n) \right| \leq c_j, \ 1 \leq j \leq n. \tag{50}
\]

Then for any \( \epsilon > 0 \)

\[
P \left\{ \hat{F}(X_1, \ldots, X_n) - \mathbb{E} \hat{F}(X_1, \ldots, X_n) > \epsilon \right\} \leq \exp \left( -2\epsilon^2 / \sum_{j=1}^{n} c_j^2 \right). \tag{51}
\]

In our case \( \hat{F} \) is the empirical mutual information \( \hat{i}_n(X_{:,1}, Y) \). Note that the empirical mutual information can be decomposed as \( \hat{i}_n(X_{:,1}, Y) = H(\hat{P}_{X_{:,1}Y}) - H(\hat{P}_{X_{:,1}}) - H(\hat{P}_Y) \). Furthermore, notice that for all integers \( 0 \leq j \leq n \),

\[
\left| \frac{j+1}{n} \log_2 \frac{j+1}{n} - \frac{j}{n} \log_2 \frac{j}{n} \right| \leq \frac{\log_2 n}{n}. \tag{52}
\]

By changing one sample point of \( Y \), there can be two values \( i' \) and \( i'' \) such that the \( \hat{P}_Y(i') \) increases by \( 1/n \) and \( \hat{P}_Y(i'') \) decreases by the same amount. The same is true of course for \( \hat{P}_{X_{:,1}} \) and \( \hat{P}_{X_{:,1}Y} \). Thus, by triangle inequality and (52) we may conclude that \( \hat{i}_n(X_{:,1}, Y) \) satisfies (50) with \( c_j = 6 \log_2 n \). Accordingly, using Proposition 1, we get

\[
P \left[ \hat{i}_n(X_{:,1}, Y) \leq (1 - \delta_1) \cdot \mathbb{E} \hat{i}_n(X_{:,1}, Y) \right] \leq \exp \left( -n \frac{C_{\text{mean}}^2 \delta_1^2}{18 \log_2 n} \right) \tag{53}
\]

\[
\leq \exp \left( -n \frac{[I(X_1;Y)]^2 \delta_1^2}{18 \log_2 n} \right) \tag{54}
\]

where in the last inequality we have used (46).

In fact, we can get better bounds for our propose. Indeed, the power of the above result lies in the fact that it is a universal bound independent of the true underlying distribution. We, however,
do not really need such a strong result. We next derive a better bound. By Chernoff’s inequality, for any \( \rho \geq 0 \), we have

\[
P \left[ \hat{i}_n(X_{i:1}, Y) \leq (1 - \delta_1) \cdot \mathbb{E}\hat{i}_n(X_{i:1}, Y) \right] \leq e^{(1 - \delta_1) \cdot \rho \cdot \mathbb{E}\hat{i}_n(X_{i:1}, Y)} e^{-\rho \hat{i}_n(X_{i:1}, Y)}. \tag{55}
\]

Now, using the method of types,

\[
\mathbb{E}e^{-\rho \hat{i}_n(X_{i:1}, Y)} = \sum_{x,y} P^n_{XY}(x,y)e^{-\rho \hat{i}_n(x,y)} \\
\leq \frac{1}{n^{3/2}} \sum_{Q_{XY}} e^{-n[I_Q(X; Y) + D_{KL}(Q_{XY} \| P_{XY})]} \\
\leq n^{3/2} \exp \left\{-n \cdot \min_{Q_{XY}} \left[ \rho \cdot I_Q(X; Y) + D_{KL}(Q_{XY} \| P_{XY}) \right] \right\} \\
\triangleq n^{3/2} \exp \left[ -n \cdot h_{\rho}(P_{XY}) \right]
\]

where

\[
h_{\rho}(P_{XY}) \triangleq \min_{Q_{XY}} \left[ \rho \cdot I_Q(X; Y) + D_{KL}(Q_{XY} \| P_{XY}) \right].
\]

Accordingly, we get that

\[
P \left[ \hat{i}_n(X_{i:1}, Y) \leq (1 - \delta_1) \cdot \mathbb{E}\hat{i}_n(X_{i:1}, Y) \right] \leq n^{3/2} \exp \left[ -n \cdot h_{\rho}(P_{XY}) \right] \\
\leq n^{3/2} \exp \left[ -n \cdot h_{\rho}(P_{XY}) - (1 - \delta_1) \rho \mathbb{E}\hat{i}_n(X_{i:1}, Y) \right]. \tag{56}
\]

and the last inequality follows from (46). Finally, note that

\[
h_{\rho}(P_{XY}) = \min_{Q_X} \left[ D_{KL}(Q_X \| P_X) + \min_{Q_{Y|X}} \left[ \rho \cdot I_Q(X; Y) + D_{KL}(Q_{Y|X} \| P_{Y|X} Q_X) \right] \right] \tag{57}
\]

\[
\geq \min_{Q_X} \left[ D_{KL}(Q_X \| P_X) + \rho \cdot I_{\frac{1}{1+\rho}}(Q_X \times P_{Y|X}) \right] \tag{58}
\]

where the last inequality follows from \([48, \text{Theorem 7}]\).

**D. Proof of Corollaries 3 and 4**

We next show that the condition in (30a) dominates (30b) which proves the desired result. To this end, we analyze the denominator in (30b) when \( \rho = \varepsilon \), for some sufficiently small \( \varepsilon > 0 \).

First, we apply Taylor’s expansion on Sibon’s divergence and get,

\[
D_\alpha(P||Q) = D_{KL}(P||Q) + (\alpha - 1) \frac{d}{d\alpha} D_\alpha(P||Q) \bigg|_{\alpha = 1} + o(\alpha - 1). \tag{59}
\]

Using L’Hôpital’s rule twice, it can be shown that

\[
\lim_{\alpha \to 1} \frac{d}{d\alpha} D_\alpha(P||Q) = \frac{1}{2} \mathbb{V}ar_P \left[ \log \frac{P(X)}{Q(X)} \right].
\]
Therefore,

\[ D_\alpha(P\|Q) = D_{KL}(P\|Q) + \frac{\alpha - 1}{2} \cdot \text{Var} P \left[ \log \frac{P(X)}{Q(X)} \right] + o(\alpha - 1). \]  

(60)

Accordingly,

\[ I_\alpha(X;Y) \geq \min_{Q_Y} \left\{ D_{KL}(Q_X P_{Y|X}\|Q_X Q_Y) + \frac{\alpha - 1}{2} \cdot \text{Var}_{Q_X P_{Y|X}} \left[ \log \frac{Q_X P_{Y|X}}{P_X Q_Y} \right] \right\} \]  

(61)

where the inequality follows from the fact that the Taylor’s expansion of Sibon’s divergence appears in the form of an alternating sum and thus the first order term which is negative provides a lower bound. Plugging the above in (28), with \( \rho = \varepsilon \), we obtain

\[ h_{\varepsilon}(P_{XY}) \geq \min_{Q_X, Q_Y} \left\{ D_{KL}(Q_X \| P_X) + \varepsilon \cdot D_{KL}(Q_X P_{Y|X}\|Q_X Q_Y) \right. \]

\[ - \frac{\varepsilon^2}{2(1+\varepsilon)} \cdot \text{Var}_{Q_X P_{Y|X}} \left[ \log \frac{Q_X P_{Y|X}}{P_X Q_Y} \right] \} \]  

(62)

Let \( Q_{X,\varepsilon}^* \) and \( Q_{Y,\varepsilon}^* \) be the minimizers of the above optimization problem. It is evident that \( Q_{X,\varepsilon}^* \rightarrow P_X \) and \( Q_{Y,\varepsilon}^* \rightarrow P_Y \), as \( \varepsilon \rightarrow 0 \). Choose \( \nu \) (the parameter controlling \( P_X \)) such that

\[ \left[ Q_{X,\varepsilon}^*(0) \right]^d = \frac{1}{2}. \]

Then, a straightforward calculation reveals that in the noiseless case

\[ D_{KL}(Q_{X,\varepsilon}^* P_{Y|X}\|Q_{X,\varepsilon}^* Q_{Y,\varepsilon}^*) = Q_{X,\varepsilon}^*(1) \cdot \log 2 \cdot (1 + o(1)). \]  

(63)

The same is true for the symmetric noise model. Then, solving the minimization problem in (62) for \( Q_X \), we get

\[ Q_{X,\varepsilon}^*(1) = P_X(1) + \varepsilon \cdot P_X(1) \cdot (1 - P_X(1)), \]

(64)

and accordingly,

\[ D_{KL}(Q_{X,\varepsilon}^* P_{Y|X}\|Q_{X,\varepsilon}^* Q_{Y,\varepsilon}^*) = I(X_1;Y) + \Omega (\varepsilon/d). \]

Therefore,

\[ h_{\varepsilon}(P_{XY}) \geq \varepsilon \cdot I(X_1;Y) - \Omega (\varepsilon^2/d). \]  

(65)

Similar calculations reveal the same result for the symmetric noise model. Using the last results, (30b) simplifies to

\[ n \geq \frac{\log d}{\varepsilon \cdot \delta_1 \cdot I(X_1;Y) - \Omega (\varepsilon^2/d)}. \]  

(66)
For both the noiseless and symmetric noise models it is well known that $I(X_1; Y) \approx C/d$ for some constant $C$. Accordingly, (66) simplifies to

$$n \geq \frac{d \log d}{\varepsilon \cdot \delta_1 \cdot C - \Omega(\varepsilon^2)}.$$  

(67)

Comparing (30a) and (66), it is clear that for $d = \Omega(p^\delta)$ (30a) always dominates (66) for sufficiently small $\theta$ and $\varepsilon$, which concludes the proof of Corollary 3. Corollary 4 also follows by noticing that both (30a) and (66) share the same order of $\Omega(d \log p)$.

E. Proof of Theorem 3

Since the joint distribution of $(X_s, Y_s|D = s)$ is the same for all $s$ in our setup, and the decoder that we have chosen exhibits a similar symmetry, we can condition on a fixed and arbitrary value of $D$, say $s = \{1, \ldots, d\}$. By the union bound, the error probability is upper bounded by

$$P_e \leq \mathbb{P} \left[ \bigcup_{(s_{\text{dif}}, s_{\text{eq}})} \left\{ \hat{h}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq \gamma_{|s_{\text{dif}}|} \right\} \right] + \sum_{\bar{s} \in D \setminus s} \mathbb{P} \left[ \hat{h}_n(X_{\bar{s}\setminus s}; Y|X_{\bar{s}\cap s}) > \gamma_{|\bar{s}\setminus s|} \right].$$

(68)

The first term corresponds to the true set failing the threshold test, and the second term corresponds to some incorrect set $\bar{s}$ passing the threshold test. In the summand of the second term, we have upper bounded the probability of an intersection of $2^d - 1$ events by just one such event, namely, the one corresponding to $s_{\text{dif}} = \bar{s}\setminus s$ and $s_{\text{eq}} = \bar{s}\cap s$. Let $\ell \triangleq |\bar{s}\cap s|$. Then, using the same method used in the proof of Theorem 1 (see, (43)-(45)), it can be shown that

$$\mathbb{P} \left[ \hat{h}_n(X_{\bar{s}\setminus s}; Y|X_{\bar{s}\cap s}) > \gamma_\ell \right] \leq C \cdot e^{-\gamma_\ell}$$

(69)

for some constant $C$. Therefore,

$$P_e \leq \mathbb{P} \left[ \bigcup_{(s_{\text{dif}}, s_{\text{eq}})} \left\{ \hat{h}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq \gamma_{|s_{\text{dif}}|} \right\} \right] + C \cdot \sum_{\ell=1}^d \left( \begin{array}{c} p-d \varepsilon \\ \ell \end{array} \right) \left( \begin{array}{c} d \varepsilon \\ \ell \end{array} \right) e^{-\gamma_\ell}.$$  

(70)

Taking $\gamma_\ell = \log \left[ \frac{d}{\delta_2} \left( \begin{array}{c} p-d \varepsilon \\ \ell \end{array} \right) \left( \begin{array}{c} d \varepsilon \\ \ell \end{array} \right) \right]$, we finally arrive at

$$P_e \leq \mathbb{P} \left[ \bigcup_{(s_{\text{dif}}, s_{\text{eq}})} \left\{ \hat{h}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq \log \left[ \frac{d}{\delta_2} \left( \begin{array}{c} p-d \varepsilon \\ \ell \end{array} \right) \left( \begin{array}{c} d \varepsilon \\ \ell \end{array} \right) \right] \right\} \right] + C \cdot \delta_2$$

(71)

$$\leq \sum_{\ell=1}^d \left( \begin{array}{c} d \varepsilon \\ \ell \end{array} \right) \cdot \mathbb{P} \left[ \hat{h}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq \log \left[ \frac{d}{\delta_2} \left( \begin{array}{c} p-d \varepsilon \\ \ell \end{array} \right) \left( \begin{array}{c} d \varepsilon \\ \ell \end{array} \right) \right] \right] + C \cdot \delta_2.$$  

(72)

We next study the expectation of the empirical density appears in the probability term at the r.h.s. of (72). It is evident that

$$I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) - \frac{c_1}{n} \leq \frac{\mathbb{E} \hat{h}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}} \leq I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) + \frac{c_2}{n}$$

(73)
where \( c_1, c_2 > 0 \) are absolute constants. Now, if
\[
n(1 - \delta_{1, \ell}) \cdot \left[I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) - \frac{c_1}{n} \right] \geq \log \left[ \frac{d}{\delta_2} \left( \frac{p-d}{\ell} \right) \left( \frac{d}{\ell} \right) \right],
\]
then by (73) we also have that
\[
(1 - \delta_{1, \ell}) \cdot \mathbb{E} \hat{i}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \geq \log \left[ \frac{d}{\delta_2} \left( \frac{p-d}{\ell} \right) \left( \frac{d}{\ell} \right) \right],
\]
so we obtain
\[
\mathbb{P} \left[ \hat{i}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq \log \left[ \frac{d}{\delta_2} \left( \frac{p-d}{\ell} \right) \left( \frac{d}{\ell} \right) \right] \right] \leq \mathbb{P} \left[ \hat{i}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \leq (1 - \delta_{1, \ell}) \cdot \mathbb{E} \hat{i}_n(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}}) \right] \leq \psi_n(\delta_{1, \ell}).
\]
Accordingly, we get that
\[
P_e \leq \sum_{\ell=1}^{d} \binom{d}{\ell} \cdot \psi_n(\delta_{1, \ell}) + C \cdot \delta_2,
\]
and hence the result is followed by the assumption that \( \delta_2 \to 0 \) and (38b). Note that (74) is actually equivalent to (38a) since it can be written as
\[
n \geq \frac{\log \left[ \frac{d}{\delta_2} \left( \frac{p-d}{\ell} \right) \left( \frac{d}{\ell} \right) \right]}{(1 - \delta_{1, \ell}) \cdot I(X_{s_{\text{dif}}}; Y|X_{s_{\text{eq}}})},
\]
where \( \delta_2' = e^{-c_1(1 - \delta_{1, \ell})} \cdot \delta_2 \), and \( \delta_2' \to 0 \) as \( \delta_2 \to 0 \).

VI. CONCLUSION

In this paper, we have provided an information-theoretic framework for the blind group testing problem. Specifically, we proposed two blind recovery algorithm schemes which are based on the idea of separate and joint decoding of items. Overall, the results of this paper establish that our empirical separate decoding of items scheme is both computationally efficient and achieves near-optimal theoretical guarantees. An interesting future work is to sharpen our sample complexity bounds and study the loss (if exists) in terms of constants, due to the fact that the channel is unknown. To this end, either sharper (and simpler) upper bounds on \( \psi_n \) or stronger blind recovery algorithms are needed. Furthermore, while the separate decoding of items scheme is computationally efficient, it is slower than recent sublinear-time algorithms [35, 36], and it would interesting to propose and analyze blind counterparts of these algorithms. While the focus of this paper is on non-adaptive schemes, it would be interesting to understand the blind adaptive counterpart as well. Finally, another somewhat related question is the mismatched setting where the recovery algorithm is basing his decision on an incorrect input-output statistics. It is then interesting to study the loss in performance as a function of the assumed (incorrect) and true models.
APPENDIX A

PROOF OF LEMMA 3

To prove Lemma 3 we first prove the following result.

Lemma 4. Let \( \mathcal{X} \) be a finite set and consider \( \mathcal{X}^n \). Let \( P \in \mathcal{P}(\mathcal{X}) \), fix \( \varepsilon > 0 \), and define

\[
B_\varepsilon(P) \triangleq \{ Q \in \mathcal{P}_n(\mathcal{X}) : \TV(P, Q) \leq \varepsilon \}.
\]

Then,
\[
|B_\varepsilon(P)| \leq 2^{|\mathcal{X}|-1} \left( (2n\varepsilon)^{|\mathcal{X}|-1} + 1 \right).
\]

Proof: Let \( |\mathcal{X}| = m \) and note that since we consider sequences of length \( n \), every type must be of the form \((a_0/n, \ldots, a_{m-1}/n)\) such that \( \sum_{i=0}^{m-1} a_i = n \). Moreover, since \( \mathcal{X} \) is finite, the total variation norm can be written in terms of the \( \ell_1 \) norm,

\[
\TV(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |Q(x) - P(x)|.
\]

The case for which \( |\mathcal{X}| = 1 \) is trivial. Let us consider the case in which \( |\mathcal{X}| = 2 \). For this case, the type is determined by the distribution of a symbol. The inequality \( \TV(P, Q) \leq \varepsilon \) implies that we may change the probability of a symbol \( x \in \mathcal{X} \) by at most \( \varepsilon \). Since this probability must be of the form \( a/n \) we may choose \( Q(x) \) to be one of at most \( 2n\varepsilon + 1 \) options (In case \( \min \{ P(x), 1 - P(x) \} \leq n\varepsilon \) we have less options). Therefore, we have that \( |B_\varepsilon(P)| \leq (2n\varepsilon + 1) \).

For \( |\mathcal{X}| = m \), let \( x_0, \ldots, x_{m-1} \) be the symbols of \( \mathcal{X} \). By the definition of the total variation norm we obtain that \( \max_{x \in \mathcal{X}} |Q(x) - P(x)| \leq \varepsilon \). Hence, we have at most \( 2n\varepsilon + 1 \) options to choose \( Q(x_0) \). After the choosing of \( Q(x_0) \) we remain with at most \( 2n\varepsilon + 1 \) options to choose \( Q(x_1) \). We may continue this calculation for \( m - 1 \) times where the value \( Q(x_{m-1}) \) is determined by \( \sum_{i=0}^{m-1} Q(x_i) = 1 \). We obtain at most \( (2n\varepsilon + 1)^{m-1} \) to choose \( Q \) and therefore \( |B_\varepsilon(P)| \leq (n\varepsilon + 1)^{|\mathcal{X}|-1} \). Since it is always true that \( (t + r)^m \leq 2^m (t^m + r^m) \) we obtain

\[
|B_\varepsilon(P)| \leq 2^{|\mathcal{X}|-1} \left( (2n\varepsilon)^{|\mathcal{X}|-1} + 1 \right).
\]

We next prove Lemma 3. First, recall that

\[
\TV(P_{XY}, Q_{XY}) \leq \sqrt{2D_{\text{KL}}(Q_{XY} \| P_{XY})}
\]

by Pinsker’s inequality. Hence,

\[
\mathcal{A}_{\varepsilon,n} \subseteq \left\{ Q_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}) : \TV(P_{XY}, Q_{XY}) \leq \sqrt{2\varepsilon} \right\}
\]
which means that

\[ |\mathcal{A}_{\varepsilon,n}| \leq \left| \left\{ Q_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}) : \text{TV}(P_{XY}, Q_{XY}) \leq \sqrt{2\varepsilon} \right\} \right|. \]

Applying Lemma 4 we obtain the upper bound in Lemma 3.

REFERENCES


