Abstract—We consider a new group testing model wherein each item is a binary random variable defined by an \textit{a priori} probability of being defective. We assume each that probability is small and that items are independent, but not necessarily identically distributed. The goal of group testing algorithms is to identify with high probability the subset of defectives via non-linear (disjunctive) binary measurements. Our main contributions are two classes of algorithms; in the first class we take adaptive measurements based either on a maximum entropy principle, or a Shannon-Fano/Huffman codes; in the second class we use non-adaptive measurements. Under loose assumptions on prior statistics, our algorithms, with high probability, only need a number of measurements that is close to the information-theoretic entropy lower bound, up to an explicit universal constant factor. We also provide simulations to support our results.

I. INTRODUCTION

The group testing model was first suggested by Dorfman [6] over sixty years ago, and has since spawned a vast affiliated literature on theory and applications (see the book [7] for a survey). The classical version of the group testing problem is that of Combinatorial Group Testing. In that version it is known that there are \( d \) defective items in a population of size \( n \) (usually it is assumed that \( d = o(n) \)). Non-linear binary disjunctive group tests are allowed, in which a subset of items is tested, and the test outcome is 1 if at least one item being tested is defective, and 0 otherwise. In that setting, if an probability of at most \( P_e \) is allowed, an information theoretic lower bound of \( (1 - P_e) \log_2 \binom{n}{d} = (1 - P_e) d \log_2 \left( \frac{d}{n} \right) + \mathcal{O}(d) \) tests are known to be necessary for both adaptive and non-adaptive algorithms (see for instance [3, 8, 13]). Adaptive group testing schemes essentially meeting this bound are known [10]; non-adaptive setting algorithms that meet this bound up to small multiplicative factors are also known [3].

In this work we focus on a model where the statistics on the likelihood of any given item to be defective are available prior to the design of the testing procedure. This model is motivated by real-world examples; for instance, in testing a large population for a given disease (Dorfman’s original motivation in [6]), historical data on the prevalence of the disease in specific sub-populations parametrized by age, gender, weight, etc are often available. Specifically, in a population of size \( n \), we denote the status of whether the \( i \)th item is defective or not by whether a corresponding binary variable \( x_i \) is 1 or 0. The length-\( n \) binary vector \( \mathbf{x} \in \{0, 1\}^n \) is the \textit{population vector}, whose recovery is the objective of the group testing algorithm.

Our working hypothesis on the prior statistics is that items might have different \textit{a priori} probabilities of being defective (non-identical) and are independent.\footnote{It is true that even this model is still fairly restrictive – it is quite natural to also consider probabilistic models with structure, such as correlation between “neighbouring” variables, or graph constraints, to model the effect of geography/social structures. These and other models are the subject of ongoing investigation.}

The knowledge of prior statistics can reduce significantly the number of required test in some scenarios. Consider the following – given the probability vector \((p_1, \ldots, p_n)\) one can compute the \textit{expected number of defective items} \( \mu \), defined as the sum \( \sum p_i \) of the individual probabilities, and in fact by standard statistical arguments [11] this quantity can even be “concentrated” (for large enough \( n \) it can be shown that with high probability the actual number of defective items is
“relatively close” to its expectation). One might then naïvely try to use existing Probabilistic Group Testing algorithms, under the assumption that an upper bound for \(d\), the number of defectives, is given by \((1 + \delta)\mu\), for some “small” \(\delta\). An immediate issue of most Probabilistic Group Testing algorithms is they assume that the prior statistics are, in one form or another, uniform – each item is equally likely to be defective.\(^2\) It is therefore by no means clear why those algorithms would have the same performance in our scenario (a naïve translation of results would indicate high probability of recovery with \(c\mu \log_2(n)\) tests for some universal constant \(c\)). Indeed, proving that such results do indeed translate, at least for one specific algorithm for the “usual” Probabilistic Group Testing model is an important module of our proof.

Another issue is performance-related. In general, \(c\mu \log_2(n)\) tests are not necessarily within a universal constant factor of the lower bounds on the number of tests required for high probability recovery. Indeed, a direct extension of known information-theoretic arguments ([3, 4, 13]) show that a natural lower bound corresponds to the entropy, \(H(x) = \sum_i h(p_i)\), of the vector of binary random variables \(x_i\) (for the sake of completeness Theorem 1 below reproduces these arguments in our probabilistic model.) It is not hard to construct distributions for the vector of random variables \(x_i\) such that the ratio between \(H(x)\) and \(\mu \log_2(n)\) is arbitrarily large.\(^3\) Indeed, there are some extremal instances of distribution on the population vector \(x\) where \(\mu\) is constant but the entropy \(H(x)\) is arbitrarily small. Therefore, existing Probabilistic Group Testing algorithms might not be optimal and have therefore no performance guarantee whatsoever under “standard” input assumptions.

We provide algorithms, both for adaptive and non-adaptive testing models, with guaranteed reconstruction of the set of defectives with high probability, that require a number of tests that are at most a small (and explicitly computed) universal factor (independent of prior statistics) away from the information-theoretic lower bound. The reason we provide both adaptive and non-adaptive algorithms, even though in principle the latter subsumes the former, is that the constant factors in the adaptive algorithms are better than in the non-adaptive case. We discuss some interesting connections between source codes and our proposed adaptive algorithms.

Previous work has attempted to analyse models involving prior statistic information. Although the model is different, from an information-theoretic point of view, interesting optimal or sub-optimal choice of parameters are shown in [16]. They deal with a particular type of prior information, where the universal set is partitioned, and within a part, all subsets of a fixed given size are uniformly distributed. This is a slight generalization of traditional group-testing and they show existence of a non-adaptive algorithm whose average performance is information-theoretically optimal, up to a small constant factor.

In this paper, we go further:

1. we provide explicit, non-adaptive and adaptive, algorithms;
2. we prove, in expectation, that they are information-theoretically sub-optimal up to some factors;
3. we show “concentration”. our algorithms are indeed optimal with some small error probability;
4. though our model is not directly comparable with [16], we believe it is broader even if an additional assumption on the total entropy \(H(x)\) is needed for the sub-optimal result of non-adaptive algorithm and concentration result of adaptive algorithm;
5. We show a direct connection between source coding and our model.

II. BACKGROUND

A. Preliminaries

1) Model and Notations: A summary of the notations used in this paper is given in Table I.

Let \(\mathcal{N} = \{x_1, x_2, \ldots, x_n\}\) denote the universal set, the set of \(n\) items being tested where each \(x_i\) is an independent binary random variable. Let \(x = (x_1, x_2, \ldots, x_n)\) be the population vector, the initial vector for our group testing. Furthermore we assume each testing item \(x_i\) being defective with the a priori probability \(p_i\) which means the independent binary random variables \(x_i\) takes value 1 with probability \(p_i\). Denote the corresponding coded vector by \(b = (b_1, b_2, \ldots, b_n)\in\{0,1\}^T\), namely the result vector. After the decoding process, an output vector is defined by \(y = (y_1, y_2, \ldots, y_n)\in\{0,1\}^n\) which is called recovery vector.

In our probabilistic model, we choose to translate the sparsity requirement of Combinatorial Group Testing into the following natural sparse property: the expected number of defective items \(\mu\) satisfies \(\mu = \sum_i p_i \ll n\). Our goal is to minimize the number of tests \(T\) used within the error probability \(P_e \geq 0\).

The algorithm should perform a sequence of measurements and guarantee that \(y\) matches \(x\) with high probability. A test\(^4\) is a subset \(S \subseteq \mathcal{N}\) to be measured, the testing procedure is the collection of all tests (possibly in order). The objective is to minimize the number of tests \(T\) given the choice of \(P_e\), which allows reconstruction of the output vector.

Notice that in Probabilistic Group Testing model, a group test (that is not an individual test) introduces an error probability \(P_e = \text{Pr}\{x \neq y\}\).

The difficulty lies in the optimization of a testing procedure, that is choosing which subsets are to be tested (and in which order for the adaptive case). The optimization should minimize the number of tests required to produce \(y\) subject to the upper bound on the error probability. Of course, the error probability depends on the particular choice of parameters of the algorithm, the trade-off of which will be discussed.

\(^2\)In the Probabilistic Group Testing literature this is indeed the assumption. In the Combinatorial Group Testing literature, the assumption is that each set of size \(d\) is equally likely to be the set of defective items.

\(^3\)For instance, consider the setting wherein the probability of the \(i\)th item being defective is \(2^{-i}\). In this case, the entropy equals \(\sum_{i=1}^{n} (2^{-i} + \frac{2^{-i}}{3} \log_2(\frac{2^i}{2^{i-1}}))\), which converges to a constant, but \(\mu \log_2(n)\) grows without bound.

\(^4\)If even one of the subset of items being tested in a group has value one (defective), then the test outcome is positive, else it is negative.
2) Pre-partitioned Model: Next introduce a way to construct a set of pre-partitioned subsets satisfying some loose assumption. Under these conditions, we will be able to show that our algorithms perform well.

**Definition 1.** The universal set \( \mathcal{N} \) is said to be a skewed set if the entropy of the population vector \( H(x) \) can be bounded by

\[
H(x) > \max(2\mu, \Gamma^2)
\]

where the measure factor \( \Gamma = \left[ \log_2 \left( \log_2 \left( \frac{2n}{P_e} \right) \right) \right] \) is defined as a function of the error probability \( P_e \) and the total number of items \( n \). Otherwise \( \mathcal{N} \) is said to be non-skewed.

**Definition 2.** For any non-empty subset \( \mathcal{U} \) of the universal set \( \mathcal{N} \), we say \( \mathcal{U} \) is a well-balanced subset if the corresponding a priori probabilities \( p_i \) of items \( x_i \) in \( \mathcal{U} \) satisfy the following constraint:

\[
p_i^2 \leq p_j \quad \forall x_i, x_j \in \mathcal{U}.
\]

Furthermore, if the following constraint also satisfied,

\[
P_e < \frac{1}{2} \quad \forall x_i \in \mathcal{U},
\]

we say \( \mathcal{U} \) is a well-balanced bounded subset. And \( \mathcal{U} \) is bounded below by \( \frac{P_e}{2n} \) if \( p_i > \frac{P_e}{2n} \); \( \mathcal{U} \) is bounded above by \( \frac{1}{2} \) if \( p_i < \frac{1}{2} \). Otherwise \( \mathcal{U} \) is said to be unbounded.

If we assume our universal set \( \mathcal{N} \) is non-skewed, then a pre-partition into subsets \( \mathcal{U}_s \) can be found using the following steps.

First we sort all the a priori probabilities \( p_i \). Then we trim the universal set \( \mathcal{N} \) into \( L \) disjoint subsets \( \mathcal{U}_s \) with index \( s \) according to the distribution of \( p_i \) such that \( \mathcal{N} \) is consists of 2 unbounded subsets and \( L - 2 \) well-balanced bounded subsets. Note that \( L \) is determined by our chosen error probability \( P_e \), each a priori probability \( p_i \) and the total number of items \( n \). The schematic diagram Figure 1 explains this procedure.

After creating the partition we classify the \( L \) subsets using the definition below.

**Definition 3.** For any non-empty well-balanced subset \( \mathcal{U} \) of the universal set \( \mathcal{N} \), we say \( \mathcal{U} \) is an ample subset if the cardinality of \( \mathcal{U} \) satisfies \( |\mathcal{U}| \geq \Gamma \). Otherwise we say \( \mathcal{U} \) is not ample.

Now for all pre-partitioned subsets \( \mathcal{U}_s \), if \( \mathcal{U}_s \) is ample, we regard it as a feasible subset for the group testing and define our group-testing algorithms on each such subset without disturbing others. On the other hand, for those subsets \( \mathcal{U}_s \) which are not ample and the last subset \( \mathcal{U}_L \) which is not bounded above by \( \frac{1}{2} \), we combine them together and test all the items in the combined set individually; for the first subset \( \mathcal{U}_1 \) which is not bounded below by \( \frac{P_e}{2n} \), we just regard all items in \( \mathcal{U}_1 \) as non-defective items without doing any test. By the methodology above for our testing procedure, one can attest acceptable concentration outcomes and upper bounds for both adaptive and non-adaptive algorithms. The results are given in section III-A with proofs in Appendix.

Fig. 1. This diagram depicts a partitioning procedure. For instance, two possible empty subsets \( \mathcal{U}_1 \) and \( \mathcal{U}_2 \) without bounds, while the remaining \( L - 2 \) trimmed subsets are well-balanced bounded. The reason why \( \mathcal{U}_1 \) are assigned 0 directly is that the probability of \( \mathcal{U}_1 \) containing defective items is quite small, thus we can still get a proper upper bound with a small error probability, details are given in appendix VI-B,VI-D. And the value between each subset implies the theoretical boundary for the partition.

3) Discussion on the Lower Bound: Before describing our algorithms, our first result states a universal information-theoretic lower bound for our model.

**Theorem 1.** Any Probabilistic Group Testing algorithm with noiseless measurements whose error probability is at most \( P_e \) requires at least \( (1 - P_e) \cdot H(x) \) tests.

As an immediate corollary, if probabilities \( p_i \) are larger than \( \frac{1}{2} \), the most efficient way to proceed is to test each element individually. Considering the disjunctive nature of measurements, it is therefore natural to test the items in the tail set \( \mathcal{U}_L \) individually.

We believe that this theorem is a witness of a relationship between compression codes and group testing. It is a counterpart of the well-known data compression lower bound. Indeed, given a probability distribution, the expected length of any code is also lower bounded by the entropy \( H(x) \) of the distribution \( x \). Further, sub-optimal/optimal codes such as Shannon-Fano/Huffman codes [5, 12, 15] meet this bound up to small additive factor. Some of our algorithms also employ such codes, in a different way, and meet the Probabilistic Group Testing lower bound up to a multiplicative factor.

This theorem will also be used in Section IV as a benchmark for our algorithms simulations.

**B. Adaptive Algorithms**

In adaptive algorithms, the order of the tests matters since we can design later tests according to the result of previous

---

3Similar techniques were used in works in the Russian literature (see for instance [9, 14] to give information-theoretic lower bounds on the required number of tests when the probabilities of items being defective are homogeneous.
The length-$L$ laminar family of all tested subsets $S_{kl}$ in the adaptive algorithms

The laminar family of all tested subsets $S_{klr}$ in the adaptive algorithms

The universal set of all items being tested $\mathcal{N}$

The disjoint pre-partitioned subsets of the universal set $\mathcal{U}$

The total number of items, $n = |\mathcal{N}|$

The number of tests used by the group-testing algorithm $T$

The number of pre-partitioned subsets $\mathcal{U}_r$

The number of pre-partitioned subsets $\mathcal{U}_s$

The length-$n$ initial population vector $(x_1, \ldots, x_n)$ population vector where $x_i$ are independent binary variables $b$

The length-$T$ binary coded result vector $(b_1, \ldots, b_n)$ where $b_i$ is the outcome of the corresponding group test $y$

The length-$n$ output recovery vector $(y_1, \ldots, y_n)$ decoded from the result vector $b$

The length-$n$ real-valued probability vector $(p_1,\ldots,p_n)$ where $p_i$ is the \textit{a priori} probability of $x_i$ to be defective $\hat{p}$

The modified probability vector got from $p$ by letting $\hat{p}_i = \frac{1-p_i}{1-p_i}$

The error probability: $P_e = \Pr[x \neq y]$

The number of subsets in the same step of tests in the laminar family $\Omega$ and $k$ is the index used for the depth of the binary tree. For example, the total number of subsets $S_{klr}$ in the $k$th stage is $m_k$

The group testing sampling parameter for the testing matrix $M$ in the non-adaptive algorithm $g$

By design, our testing procedure will always satisfy the following property: if a subset $S$ tested positive at stage $k$, then $S$ will be split into two (children) subsets to be tested at stage $k+1$. In this way, the whole testing procedure can be depicted as a tree where stages correspond to depth. Tests in child nodes correspond to subsets of items being tested in their parent node. Leaves are individual tests, thus a path in this tree identifies a single defective item. For our purposes, the first stage will usually be dealt separately.

We now describe two different ways to construct the tree. Both of them use a \textit{laminar family of subset} $\Omega$ which contains subsets $S_{klr}$. In this way, there will be no cross-testing between different trees. Each subset $S_{klr}$ in the laminar family $\Omega$ forms a node in our set of testing trees as the Figure 2. Thus our adaptive algorithm will be called \textit{Laminar Algorithm}.

1) \textbf{Laminar Algorithm Based on Maximum Entropy:} Suppose we know the first $k-1$ outcomes $(b_1', b_2', \ldots, b_{k-1}')$ where $b_i'$ is the binary result of test $i$.

We define the next test by choosing a subset such that $\Pr[b_k = 0 | b_{k-1} = b_{k-1}' = \ldots = b_2 = b_2', b_1 = b_1']$ is closest to $\frac{1}{2}$ (thus locally maximizing the information learned at each stage about the input vector).

In general getting a probability of exactly $1/2$ is not possible due to the fact that the probability vector has arbitrary entries. Quantifying the impact of these “quantization errors”, both in terms of the probability of error, and the number of tests required, is one of the major tasks required in our proof techniques. The full details are given in the third part III-B1 and the proof in the appendix (see VI-B).

In practice, we choose the subsets being tested such that the probability they contain a defective item is close to one-half, given that the previous tests were positive. This imply, by Bayes rule, that the product of $(1-p_i)$ has to be close to a certain value that we will explicit.

2) \textbf{Laminar Algorithm With Source Codes:} The second type of adaptive algorithms is based on Shannon-Fano/Huffman source codes. Instead of constructing the tree one step at a time, we do the partition at the beginning to make sure that the source coding process can be done correctly within each subset. Then set the weight $w_i$ we need as $p_i$ and the testing tree can be constructed before we doing the group testing. An similar upper bound is achieved using Shannon-Fano/Huffman coding trees with lengths of codewords $l_i$ bounded by $\lceil \log_2 \frac{1}{p_i} \rceil$. A more detailed discussion is provided in theorem 2 and section VI-B.

Figure 2 partially depicts the testing tree of the \textit{Laminar Algorithm} describing also typical structure describing the criteria to form our laminar family $\Omega$.

3) \textbf{Concentration:} It remains to show that the actual number of steps in these algorithms is close to the expected value, that is: concentrate the number of tests $T$ required. Our original model of prior is too general, preventing us from having a general concentration result with the technique we employ. To solve this, we add an extra assumption at the very first stage and assume $\mathcal{N}$ is non-skewed according to definition \ref{def:non-skewed}. Then a pre-partitioned testing pool which is comprised of $L$ different subsets can be constructed.

Then based on the pre-partitioned model, we’re going to test the ample subsets which are large enough. Moreover, another problem is that the summation of the \textit{a priori} probabilities may not be large enough within first several ample subsets. To deal with this, we combine the first several subsets until the summation of \textit{a priori} probabilities is more than half. The total number of subsets we need to combine turns to be bounded above by a relatively small value with details of proof in VI-B.

C. Non-adaptive Algorithm

Non-adaptive algorithms require the testing procedure to be fixed in advance. Therefore they might be less powerful than adaptive algorithms, but now the tests can be done in parallel, which is convenient for hardware design.
adapted from the traditional one in [4]. The group-testing procedure can be partitioned into $L$ sets, where $L$ is the number of subsets containing the items set to be tested. In this way, we have the following two methods: $ME$ and $S/H$.

For the design of our non-adaptive algorithms, we represent the testing procedure as a $T \times n$ Boolean matrix $M$. Each row of $M$ corresponds to a measurement, and each column corresponds to a single item to be tested. In this way, we have the matrix representation $M$ for the whole testing procedure can be partitioned into $L - 1$ sub-matrices $M_s$ which are all bounded below by $\frac{P_e}{2^m}$, and such that

$$M = \bigoplus_{s=2}^{L} M_s,$$

where $\bigoplus$ is the direct sum.

In the Block-Coco Algorithm, we will assume the existence of a suitable pre-partition and use, as a sub-algorithm, the Coupon Collector Algorithm for each subset $U_s$, for which we control the testing complexity and error probability.

We depict the matrix $M$ of the Block-Coco Algorithm in Figure 3.

### III. MAIN RESULTS

#### A. Upper Bound of Our Algorithms

**Theorem 2.** Laminar Algorithm needs at most $2H(x) + 2\mu$ tests in expectation. Furthermore if the universal set $\mathcal{N}$ is non-skewed, then with error probability $P_e = (2^\Gamma - \frac{1}{2}^{\frac{1}{\delta}+1})$ and $\delta \geq 2e - 1$, it requires no more than $\left(1 + \delta \right) \left(2^{\Gamma+log_2 3} + 2 \right) H(x)$ tests.

**Theorem 3.** If the universal set $\mathcal{N}$ is bounded above by $\frac{1}{2}$, then the Coupon Collector Algorithm with error probability at most $P_e = n^{-5}$ requires no more than $4e (1 + \delta) \mu \ln n$ tests.

---

**Fig. 2.** Graphical description of the Laminar Algorithm without error estimation, first we start with the initial set $x$ containing $x_1$ up to $x_n$. Then do the partition we have several subsets from $S_{11}$, up to $S_{11}$, satisfying $\prod_{i=1}^{n} (1 - p_i) = \frac{1}{2}$ and so on in this particular example. The construction of testing tree follows the two method: $ME$ and $S/H$.

**Fig. 3.** Graphical description of Block-Coco Algorithm, where the testing matrix are defined as this particular example. The population vector is $x$ which is encoded to $b$ and decoded to $y$ by doing the pre-partition. The decoding process contains three different cases: For each subset $U_s$, group testing procedure is done separately; for $U_1$, we assign it with a zero matrix which means there is no tests for it since all items are set to be 0 directly; moreover, for items in those subsets which are not ample or in $U_2$, we test them individually.
Theorem 4. If the universal set $\mathcal{N}$ is non-skewed, then with error probability at most $P_e = 2^{\Gamma - \delta + 1}$, Block-Coco Algorithm requires no more than $(12e + 2) \left( 1 + \delta \right) H(x)$ tests.

B. Algorithms

First we will implement an adaptive algorithm following the discussion in II-B. Then we describe in details the non-adaptive block design in II-C.

As we discussed, our tree-based algorithms consist of many stages, the first one is treated separately. Indeed, the very first stage is based on an initial partition. Thus the “tree” is not binary at the root but is binary afterwards (alternatively we can see it as a forest of binary trees). Indeed, each positive test at stage $k$ induces two more (child) tests at stage $k + 1$.

1) Laminar Algorithm: Maximum Entropy:

a) First Stage: In the first stage, we check $\mu = \sum_{i=1}^n p_i$ is smaller than $P_e$, if so we just return $y = 0$. Otherwise, we partition the universal set $\mathcal{N}$ into contiguous subsets $S_{11r}$, indexed by $r$, greedily: The partition is chosen such that $\Pr \exists x_j \in S_{11r} \text{ s.t. } x_j = 1$ is the closest to $\frac{1}{2}$, i.e.:

$$\min_{x_j \in S_{11r}} \prod_{x_j \in S_{11r}} (1 - p_j) - \frac{1}{2} \quad \text{subject to } p_j < 1,$$

where $j = 1, 2, \ldots, n$.

b) Second Stage: In the second stage, negative tests ensure no item is defective, we can thus forget about the corresponding subsets. If the test is positive we divide the corresponding subset $S_{11r}$ into two smaller subsets $S_{21r}$, $S_{22r}$, such that $\Pr \exists x_j \in S_{21r} \text{ s.t. } x_j = 1 | b_1 = 1$ is the closest to $\frac{1}{2}$, i.e.,

$$\min_{x_j \in S_{21r}} \prod_{x_j \in S_{21r}} (1 - p_j) - \frac{1}{2} \quad \text{subject to } p_j < 1,$$

where $j = 1, 2, \ldots, n$.

c) Later Stages: Similarly, in the $k$th stage, we forget about subsets that tested negative in the previous stage, and split each remaining subset into two parts in a similar way:

$$\min_{x_j \in S_{k1r}} \prod_{x_j \in S_{k1r}} (1 - p_j) - \frac{1}{2} \quad \text{subject to } p_j < 1,$$

where $j = 1, 2, \ldots, n$, $\forall l = 1, 3, \ldots, 2^{k-1} - 1$.

Notice that, (a) we use contiguous partitions since the probability vector is sorted; (b) all tests in a given stage involve disjoint subsets and can be thus made in parallel; and (c) this procedure terminates and the leaves of the tree correspond to individual tests.

**2) Shannon-Fano/Huffman Coding-based Algorithm (S/H):**

The idea is to notice that instead of doing using the particular binary trees for each subset in the previous partition, one can instead use a source code. This is possible, using Kraft’s inequality, assuming the probability of each subset is less than one.

Our construction of $S/H$ Algorithm goes as follows:

a) First Stage: The first stage is similar to the previous one except that we require the product of $(1 - p_i)$ in each subset being strictly larger than half. The partition is defined by:

$$\min_{x_j \in S_{11r}} \prod_{x_j \in S_{11r}} (1 - p_j) - \frac{1}{2} \quad \text{subject to } p_j < 1, \prod_{x_j \in S_{11r}} (1 - p_j) \geq \frac{1}{2}$$

where $j = 1, 2, \ldots, n$.

b) Later Stages: Next, within each subset $S_{11r}$, we have $\prod_{x_j \in S_{11r}} (1 - p_j) \geq \frac{1}{2}$. This implies that $\sum_{j \in S_{11r}} p_j \leq 1$ with details in VI-B. For each subset $S_{11r}$, we set the weights $w_i$ as the corresponding $p_i$ and apply Shannon-Fano coding, Huffman coding or any source codes to construct the testing tree.

3) Block-Coupon Collector Algorithm (Block-Coco): Suppose the universal set $\mathcal{N}$ is non-skewed and thus has a pre-partition. Each ample subset of the partition will be considered separately. We use the following steps to specify the corresponding testing sub-matrix $M_s$:

First, according to the given a priori probability vector $p$, compute the corresponding $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$ where $\hat{p}_i = \frac{1 - p_i}{n - \mu}$. Then compute the group testing sampling parameter $g$ by

$$g^* = \frac{-1}{\ln (\sum_{i=1}^n \hat{p}_i (1 - p_i))},$$

which is the optimal parameter for our purposes, as shown in VI-C.

Then for each testing sub-matrix $M_s$, in each row we choose the items with replacement $g^*$ times according to the probability distribution vector $\hat{p}$ and form the testing matrix $M$ by

$$M = \bigoplus_{s=2}^L M_s$$

as discussed in II-C.

IV. EXPERIMENTAL RESULTS

We provide experimental simulation results for both Laminar Algorithm and Coupon Collector Algorithm. We considered three different extremal kinds of probability vectors $p$: uniform, linear, and exponential.

For LA, both ME and Huffman constructions were simulated. We used 200 different points of entropy $H(x)$. As a result we can see the $E(T)$ computed from 200 trials at each entropy point looks linear in $H(x)$ where the coefficient is close to 2.
as shown in figs. 4 to 9. Moreover, the experiment is based on standard LA without using pre-partitioned model to ensure the concentration results.

For Coupon Collector, based on the different types of distribution of \( p \), we tested three different values of \( \mu \) and computed the corresponding error probability in 200 trials. We compared this error probability with the theoretic error probability \( P_e \) as shown in figs. 10 to 12.

**V. CONCLUSION AND DISCUSSION**

We studied a novel model of Probabilistic Group Testing and provided two types of algorithms. Experimental results tend to show concentration of our theorems on the expected number of tests.

Other more general models involving joint probabilities between each item \( x_i \) seems to be more attractive and useful, yet more complex to analyse, this is the subject of on-going work.
Our new model encapsulate the previous combinatorial model, and our algorithms provide similar encoding complexity. This makes us believe our new model is a natural generalisation of the traditional model.

It is also worth noting that for compressive sensing there are already a series analyses considering sparse signal with prior information and some weighted $l_1$ minimization recovery algorithm was demonstrated to have advantages over the conventional one, see [1].

REFERENCES

[1] M Amin Khajehnejad, Weiyu Xu, A Salman Avestimehr, and Babak Hassibi. Weighted $l_1$ minimization for sparse recovery with prior information. In Information Theory, 2009. ISIT 2009. IEEE International Symposium on, pages 483–487. IEEE, 2009. 8

[2] Arnon Boneh and Micha Hofri. The coupon-collector problem revisited. 1989. 10

[3] Chun Lam Chan, Pak Hou Che, Sidharth Jaggi, and Venkatesh Saligrama. Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on, pages 1832–1839. IEEE, 2011. 1, 2, 10

[4] Chun Lam Chan, Sidharth Jaggi, Venkatesh Saligrama, and Samar Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. In Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, pages 1837–1841. IEEE, 2012. 2, 5

[5] Thomas M Cover and Joy A Thomas. Elements of information theory. John Wiley & Sons, 2012. 3

[6] Robert Dorfman. The detection of defective members of large populations. The Annals of Mathematical Statistics, 14(4):436–440, 1943. 1

[7] Ding Zhu Du and Frank Hwang. Combinatorial group testing and its applications. World Scientific, 1993. 1
Then we have
\[ H(x|y) = H(e, x|y) = H(e|y) + \Pr [e = 0] H(x|y, e = 0) + \Pr [e = 1] H(x|y, e = 1) \leq H(P_e) + P_e H(x) . \] (7)

Also we have \( I(x; b) \leq H(b) \leq \log_2 |b| = T \) by the data-processing inequality. Hence we obtain that
\[ T \geq (1 - P_e) H(x) . \]
\[ \blacksquare \]

\section*{A. Proof of Theorem 1}

\textbf{Proof:} The input vector, noiseless result vector and estimated input vector form a Markov chain \( x \rightarrow b \rightarrow y \). Then we have
\[ H(x) = H(x|y) + I(x; y) . \] (4)

Define an error random variable \( e \) such that
\[ e = \begin{cases} 1, & \text{if } y \neq x \\ 0, & \text{if } y = x. \end{cases} \]

By Fano's inequality, we can bound the conditional entropy as
\[ H(e, x|y) \leq \log_2 \left( \frac{1}{P_e} \right) . \]

The length of branch \( l_i \) for each \( a \) priori probability \( p_i \) is bounded by
\[ l_i \leq \left\lceil \log_2 \frac{1}{p_i} \right\rceil . \] (8)

Note that (8) also holds for Shannon-Fano coding [15]. To justify Shannon-Fano coding is well-defined, we first introduce Lemma 1.

\textbf{Lemma 1.} Let \( k \) be a positive integer. If \( 0 < p_i < 1 \) for \( 1 \leq i \leq k \) and \( \prod_{i=1}^{k} (1 - p_i) \geq 1/2, \) then \( \sum_{i=1}^{k} p_i \leq 1. \)

\textbf{Proof:} Given \( \prod_{i=1}^{k} (1 - p_i) \geq 1/2, \) or equivalently,
\[ \prod_{i=1}^{k} \frac{1}{1 - p_i} \leq 2 . \] (9)

(9) can then be expanded by geometric sum as
\[ 2 \leq \prod_{i=1}^{k} \left( \sum_{j=0}^{\infty} p_i^j \right) \leq 1 + \sum_{i=1}^{k} p_i , \]
which yields the result \( \sum_{i=1}^{k} p_i \leq 1. \)
\[ \blacksquare \]

One thing remains to be shown is that under the partition for each subset \( S_{11} \), such that
\[ \prod_{x_j \in S_{11}} (1 - p_j) \geq \frac{1}{2} . \] (10)

then the S/H is well-defined.
Note (10) is the construction requirement of the first stage. By Lemma 1, we have the Shannon-Fano/Huffman coding procedure is well-defined since the summation of a priori probabilities within each subset is smaller or equal to 1.

Therefore, we can bound the expected number of tests $E[T]$ as

$$E[T] \leq \sum_{i=1}^{n} 2p_i l_i \quad (11)$$

$$\leq \sum_{i=1}^{n} 2p_i \left( \log_2 \frac{1}{p_i} + 1 \right) \quad (12)$$

$$\leq 2H(x) + 2\mu, \quad (13)$$

where $\mu$ is the summation of all a priori probabilities and (11) comes from our testing procedure such that a positive testing outcome, implies two more tests for both its children.

For the second part of the result, we must show that $T$ concentrates properly whenever the universal set $N$ is non-skewed and hence we can construct the Pre-partitioned Model. Then in order to make our subset large enough, we need to combine two small subset together such that in the combined subset $U$ we have $\sum_{i \in U} p_i \geq \frac{1}{2}$ such that we can have a acceptable concentration result.

The maximum number $\Lambda$ of the subsets we need to combine is bounded by

$$\Gamma \left( \frac{p_e}{2n} \right) \left( \frac{1}{2} \right) ^{\Lambda} \leq \frac{1}{2}. \quad (14)$$

Since we always have $\Gamma > 1$, we can deduce the following from (14) such that

$$\Lambda < \Gamma. \quad (15)$$

Within each ample subset $U_s$ denote the corresponding population vector by $x_s$ and the summation of the a priori probabilities by $\mu_s$, since the subsets being tested are ample and well-balanced bounded, let $p_{max} = \max_{i \in U_s} p_i$, then we have

$$l_i \leq 2l_{max} \quad \forall i \in U_s. \quad (16)$$

Thus by (15) and (16) for the combined subsets $U_2 \bigcup U_3 \bigcup \ldots \bigcup U_r$ we have

$$l_i \leq 2^r l_{max} \quad \forall i \in U_2 \bigcup U_3 \bigcup \ldots \bigcup U_r. \quad (17)$$

Hence from (17) we assume $\delta \geq 2\epsilon - 1$ and the Chernoff Bound [11] gives

$$\Pr \left[T \geq (1 + \delta) \left( 2^{\Gamma + \log_2 3} + 2 \right) \left( H(x_2) + H(x_3) + \ldots + H(x_r) \right) \right] \leq 2^{-(1+\delta)3H(x_s)}. \quad (18)$$

And since well-balanced subsets satisfy $p_i^2 \leq p_j \ \forall x_i, x_j \in U_s$ we can bound the entropy $H(x_s)$ by the following:

$$H(x_s) > \sum_{i \in U_s} p_i \log_2 \frac{1}{p_i} \geq \frac{1}{2} \mu_s \log_2 \frac{n_s}{\mu_s}. \quad (19)$$

Then combine (18) and (19) we have

$$\Pr \left[T \geq (1 + \delta) \left( 2^{\Gamma + \log_2 3} + 2 \right) \left( H(x_2) + H(x_3) + \ldots + H(x_r) \right) \right] < 2^{-(1+\delta)\Gamma+1}. \quad (20)$$

and in total we have

$$\Pr \left[T \geq (1 + \delta) \left( 2^{\Gamma + \log_2 3} + 2 \right) \right] < 2^{-(1+\delta)\Gamma+1}. \quad (21)$$

where (20) to (21) is because the subset $U_s$ is ample, namely $|U_s| > \Gamma$.

Then consider the total error probability for the group testing procedure, for $T \geq (1 + \delta) \left( 2^{\Gamma + \log_2 3} + 2 \right) H(x)$ the error probability $P_e$ satisfies

$$P_e < (2\Gamma)^{-\frac{(1+\delta)}{2^{\Gamma}}} + 1. \quad \square$$

C. Proof of Theorem 3

Proof: The proof approach is based on that of Theorem 3 in [3]. Our goal is to efficiently identify all non-defective items in the universal set $N$. Here we map the problem to Coupon Collector’s Problem. Non-defective items are the coupons. The set of negative tests which directly reveal non-defective items can be viewed as a chain of coupon collection.

Then for each row, we assume a fixed group testing sampling parameter $g$ such that the sampling process are conducted exactly $g$ times for each row. And we use a non-uniform sampling distribution $\hat{p}$ to draw the coupons. Hence the probability of outcome 0 for each group testing process is $\left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right)^g$ and in total we have $Tg$ times of selecting the coupons, namely the non-defective items from the universal set $N$. Thus we describe a test as a length-$g$ sequence of selection and when a collector obtains a full set of coupons, the number of coupons collected should be at least the stopping time $T$. In expectation, we can summarize the following equation:

$$Tg \left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right)^g \geq E[T]. \quad (22)$$

For items being drawn with non-uniform distribution $\hat{p}$, [2] suggests that the expected stopping time $E[T]$ is given by

$$E[T] = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < i_2 < \ldots < i_r \leq n} \frac{1}{p_{i_1} + \hat{p}_{i_2} + \cdots + \hat{p}_{i_r}}. \quad (23)$$
Lemma 2. Let $n \in \mathbb{Z}^+$, we have
\[
\sum_{r=1}^{n} (-1)^{r-1} \left( \frac{n}{r} \right) \frac{1}{r} = \sum_{r=1}^{n} \frac{1}{r}.
\]

Proof:
\[
\sum_{r=1}^{n} \frac{1}{r} = \int_{0}^{1} \frac{1 - s^n}{1 - s} \, ds = \int_{0}^{1} \frac{1 - (1 - t)^n}{t} \, dt = \int_{0}^{1} \left[ \sum_{r=1}^{n} (-1)^{r-1} \left( \frac{n}{r} \right) t^{r-1} \right] \, dt = \sum_{r=1}^{n} (-1)^{r-1} \left( \frac{n}{r} \right) \frac{1}{r},
\]
where (24) follows from the expansion of geometric sum, (25) follows from substituting $s = 1 - t$, and (26) follows from binomial theorem.

Let $\mu = \sum_{i=1}^{n} p_i$. (22) can be further computed as
\[
E[T] = (n - \mu) \sum_{r=1}^{n} (-1)^{r+1} \frac{1}{r} \left( 1 - \frac{p_{i_1} + p_{i_2} + \ldots + p_{i_r}}{r} \right)^{-1}
\]
\[
= (n - \mu) \sum_{r=1}^{n} (-1)^{r+1} \frac{1}{r} \left( \sum_{j=0}^{\infty} \left( \frac{p_{i_1} + p_{i_2} + \ldots + p_{i_r}}{r} \right)^j \right)
\]
\[
\leq (n - \mu) \sum_{r=1}^{n} (-1)^{r+1} \left( \frac{n}{r} \right) \frac{1}{r} \left( 1 + \frac{\mu}{n} \sum_{s=0}^{\infty} 2^{-s} \right)
\]
\[
= (n - \mu) \ln n \left( 1 + \frac{2\mu}{n} \right)
\]
(30)
\[
< (n - \mu) \ln n
\]
(31)
where (27) follows from substituting $\hat{p}_i = (1 - p_i)/(n - \mu)$, (28) follows from the expansion of geometric sum with the fact that every $p_i$ as well as the average $r^{-1} \sum_{j=1}^{r} p_{j}$ is between 0 and 1. Since we assume the universal set $\mathcal{N}$ is bounded above by $\frac{1}{2}$ using $p_i < \frac{1}{2}$ and expanding (28) we can get (29). Moreover (30) follows from lemma 2 and $\sum_{r=1}^{n} 1/r \leq \ln(n)$.

Substituting (31) into (22) and optimizing for $g$, we obtain $g^* = -1/\ln \left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right)$. Such choice of $g^*$ and the assumption $\mu \ll n$ allow (22) to be simplified as
\[
T \geq e\mu \ln n
\]
(32)
since the ratio between the expected stopping time and the expected non-defective items in a single negative test can be computed as
\[
\frac{(n + \mu) \ln n}{g^* \left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right)^{g^*}} = \frac{(n + \mu) \ln n}{\left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right) \ln \left( \sum_{i=1}^{n} \hat{p}_i (1 - p_i) \right)} = e(n + \mu) \ln n \left( \frac{\sum_{i=1}^{n} \hat{p}_i (1 - p_i)}{n - \mu} \right) = e(n + \mu) \ln n \left( \frac{(n - \mu)^{2 - \mu}}{n - 2\mu} \right) \approx e\mu \ln n
\]
Note that (22) only accounts for the expectation. Now we take variance in consideration. By Chernoff bound, the actual number of items in the negative tests can be smaller than $1 - \alpha$ times the expected number with probability at most $\exp(-\alpha^2 T)$. In tail estimate of the coupon collector problem, with probability $n^{-2\alpha + 1}$, a collector requires more than $1 - \alpha E[T]$ coupons before he can collect a full set. Thus, using union bound over two error events, (22) and (32) shall be generalized as
\[
(1 - \alpha) T \geq e\alpha \mu \ln n,
\]
(33)
which does not hold with probability $P_e$ at most $\exp(-\alpha^2 T) + n^{-2\alpha + 1}$. Taking $\alpha = \frac{1}{2}$ in (33), we can bound the error probability as
\[
P_e \leq \exp\left(-\frac{\alpha^2 T}{4}\right) + n^{-2\alpha + 1} \leq n^{-\alpha} + n^{-2\alpha + 1} \leq 2n^{-2\alpha + 1}.
\]
If we take $2n^{-2\alpha + 1}$ to be $n^{-\delta}$, we have $\chi \rightarrow 2(\delta + 1)$ as $n \rightarrow \infty$. Hence, for large $n$, theorem 3 holds.

D. Proof of Theorem 4

Proof: For non-skewed universal set $\mathcal{N}$, we use the Pre-partitioned model with the following notations. Let $l \leq L$ be the total number of ample subsets such that an ample subset $\mathcal{U}_s$ is indexed by $s \in \{1, 2, \ldots, l\}$, $\mu_s$ be the sum of a priori probabilities of items in $\mathcal{U}_s$, $n_s$ be the number of items in $\mathcal{U}_s$, $x_s$ be the population vector for each $\mathcal{U}_s$, and $\hat{p}_s = n_s^{-1} \sum_{x \in \mathcal{U}_s} p_j$.

The total number of tests $T$ is the sum of the total number of tests for ample subsets, denoted by $T_1$, and the total number of tests for unbounded non-ample subsets, denoted by $T_2$. By Theorem 3, with error probability at most $n_s^{-\delta}$ for each ample
subset $U_s$, $T_1$ can be computed as

$$ T_1 \leq 4e (1 + \delta) \sum_{s=1}^{l} \mu_s \ln n_s $$

$$ = 4e (1 + \delta) \sum_{s=1}^{l} \left( \sum_{i : x_i \in U_s} p_i \ln \frac{1}{p_s} + \mu_s \ln \mu_s \right) $$

$$ \leq 4e (1 + \delta) \sum_{s=1}^{l} \left( 2 \sum_{i : x_i \in U_s} p_i \ln \frac{1}{p_i} + \mu_s \ln \mu_s \right) $$

(34)

$$ \leq 12e (1 + \delta) \sum_{s=1}^{l} H(x_s) $$

(35)

$$ = 12e (1 + \delta) H(x) $$

(36)

where inequality (34) follows from the fact that well-balanced subsets satisfy $p_i^2 \leq p_j \forall x_i, x_j \in U_s$ and thus $p_i^2 \leq \bar{p}_s \forall x_i \in U_s$, and inequality (35) follows from ...

As to unbounded non-ample subsets, the number of tests required is at most $L^2 + 2\mu$. Since $\mathcal{N}$ is non-skewed, we can bound $L^2$ as

$$ L^2 < \left\lceil \log_2 \left( \log_2 \left( \frac{2n}{P_e} \right) \right) \right\rceil^2 $$

$$ \leq H(x) $$

and $2\mu$ is at most $H(x)$ according to Definition 1. Thus, we have

$$ T_2 < H(x). $$

(37)

Putting (36) and (37) and together, we can conclude the total number of tests is given by

$$ T = T_1 + T_2 $$

$$ < (12e \mu + 2) \left( 1 + \delta \right) H(x) $$

with error probability at most ... with probability $P_e$

$$ P_e \leq \frac{1}{2} \sum_{s=1}^{l} n_s^{-\delta} + \frac{P_e \mu}{2} $$

which is equivalent to:

$$ P_e \leq 2\Gamma^{-\delta+1}. $$