Lower Bounds in the Pooled Data Problem with Noise from the Multinomial Distribution

Arends, Jasper R.M.

Award date:
2022

Lower Bounds in the Pooled Data Problem with Noise from the Multinomial Distribution

J.R.M. Arends, 1337238

June 30, 2022

Supervised by N.S. Müller

Department of Mathematics and Computer Science

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Abstract

In this research, a noisy variant of the pooled data problem is investigated. Here, the labels associated to each items in a large population are recovered through a series of pooled tests that reveal the number of items for each label in that queried pool. Particularly, the noise is generated by assuming that the items are correctly identified with a certain probability, and is otherwise uniformly classified with one of the other labels. For this model, lower bounds using the framework set up by Scarlett and Cevher (2017) were established in the deterministic and non-adaptive setting, where respectively the pools were pre-determined and items were pooled with a certain probability, independent of all other items. This was first done in the binary case, where only two labels were considered, and then generalized to an arbitrary number of labels. The results indicated that the minimum number of tests required to recover the labels of all items with a certain confidence is at least linear in the population's size when the probability of correct identification is taken as a constant. This holds for the binary case in both the deterministic and non-adaptive setting. Yet in the general case, this is only valid for the deterministic setting, otherwise this bound is of a logarithmic order. When the probability of correct identification would converge to one, a noiseless model is approached and the minimum number of tests is of the same order as the noiseless setting.

Contents

1 Notation

We define $[n] = \{1, ..., n\}$ for all positive integers n. For a discrete random variable X, we denote its corresponding probability mass function by f_X and its support by its corresponding calligraphic letter, i.e. $\mathcal{X}$. In numerous equations, the logarithmic function $\log_b(\cdot)$ with base $b > 0$ will be used. The parameter b will be omitted when the natural logarithm is used.

1.1 Entropy of random variables

In information-theory, concepts such as (conditional) entropy and mutual information are commonly used. They can be defined as follows.

Definition 1 ((Conditional) entropy)**.** *Let X and Y be discrete random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ with probability mass functions f_X and f_Y respectively. Then the entropy of X is defined as*

$$H(X) = -\sum_{x \in \mathcal{X}} f_X(x) \log(f_X(x)) \tag{1.1}$$

and the conditional entropy of X given Y is

$$H(X|Y) = \sum_{y \in \mathcal{Y}} f_Y(y) H(X|Y = y). \tag{1.2}$$

Given these definitions, the mutual information between two random variables can be defined as follows.

Definition 2 (Mutual information)**.** *The mutual information between two discrete random variables X and Y is defined as*

$$I(X;Y) = H(X) - H(X|Y). \tag{1.3}$$

Whereas Cover and Thomas (2006) also give definitions for the (conditional) entropy and mutual information of continuous random variables, this research is limited to the discrete case only.

1.2 Asymptotic notation

Concepts in information theory often require asymptotic notations. The most relevant notions will be considered here.
Let $f, g : \mathbb{N} \to \mathbb{R}$ be two functions. We can describe different orders of asymptotic growth using the following Bachmann-Landau notation.
To describe that f grows at least as fast as g, the Ω notion can be used. We say that $f(n) = \Omega(g(n))$ if and only if

$$\exists c > 0,\ \exists n \in \mathbb{N},\ \forall n \geq N: \quad c \cdot |g(n)| \leq |f(n)|.$$

Moreover, when g is both an asymptotic lower and upper bound, Θ can be applied. We say that $f(n) = \Theta(g(n))$ if and only if

$$\exists c_1, c_2 > 0,\ \exists n \in \mathbb{N},\ \forall n \geq N: \quad c_1 \cdot |g(n)| \leq |f(n)| \leq c_2 \cdot |g(n)|.$$

Many concepts in information theory require approximations. The two most notable notions little-o and big-O are described by Paulsen (2014) as follows. Consider again two functions $f, g : \mathbb{R} \to \mathbb{R}$. Firstly, we say that $f(x)$ is of order $g(x)$ as x approaches $a \in \mathbb{R}$ and write $f(x) = O(g(x))$ if and only if

$$\exists M, \varepsilon > 0,\ \forall x \in \mathbb{R}: \quad |x - a| \leq \varepsilon \implies |f(x)| \leq M|g(x)|.$$

Similarly, $f(x)$ is of order $g(x)$ as $x \to \infty$ when

$$\exists M, N > 0,\ \forall x > N: \quad |f(x)| \leq M|g(x)|.$$

A more stronger statement is the little o notation. We say that $f(x)$ is less than order $g(x)$ as x approaches $a \in \mathbb{R}$ if $g(x)$ has no zeros in a neighbourhood of a and write $f(x) = o(g(x))$ when

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0.$$

1.3 Random variables

Throughout this thesis, several different types of random variables will be presented. We will use the following shorter notation to indicate certain distributions.

A binomial variable with distribution $\mathrm{Bin}(n, p)$ counts the number of successes among $n \in \mathbb{N}$ independent trials, where each trial is successful with a probability $p \in [0, 1]$.

A multinomial random variable with distribution $X \sim \mathrm{Multin}(n, \boldsymbol{p})$, where $\boldsymbol{p} \in [0, 1]^d$ and $p_1 + ... + p_d = 1$, counts the outcomes of n independent trials, each with an outcome in a fixed set with cardinality $d \in \mathbb{N}$. For $i = 1, ..., d$, let X_i denote the number of trials with the i-th outcome, and p_i the probability that a trial has the i-th result.

A hypergeometric random variable $H \sim \mathrm{Hyp}(k, m, n)$ counts the number of successes when k elements are selected from a group with a total of $n \in \mathbb{N}$ elements. Here, $m \in \mathbb{N}$ is the number successful items.

2 Introduction

2.1 The pooled data problem

The pooled data problem is a mathematical problem that is applied in many different scientific fields. Here, a large population of items is considered, where each item has an associated label that are unknown at first. The goal is to recover the label of each item through a series of test, that can be applied as follows. A group of items, or a so called 'pool', is selected, and a test reveals how many items with each label can be found in that specific group, though not the individual labels. By cleverly pooling the items and combining the test results, the labels of all items can be recovered.

The pooled data problem was based on the mathematical problem introduced during World War II by Robert Dorfman. Here, draftees needed to be screened for syphilis. Rather than testing the people one by one, blood samples of a group of people were collected and tested together. When this group tested positive, each person would then be tested again, but individually. On the contrary, when the pool result was negative, each person would be diagnosed as negative. This allowed for a significant decrease in the number of tests, which was overall much more efficient and cost effective.
Nowadays, pooling data is often applied in DNA sequencing, particularly to test for the frequency of certain allele in large-scale studies (Sham, Bader, Craig, O'Donovan, & Owen, 2002). Simply put, a gene is a DNA segment that causes some (physical) property to occur, for example a certain gene eventually determines the color of a person's eyes. Allele are variances of particular gene. Within the same example, this includes the colors brown, blue etc. Investigating the allele's of a person could indicate potential health-issues. For example, there are several type of allele's that result in a higher risk of obesity. Pooling data can be used to identify who has such specific types of alleles in large populations. A test indicates how many people in the queried pool contain a specific type, and simultaneously how many people do not. By choosing the pools of people strategically, only a few tests need to be conducted, instead of testing everyone individually.

Generally, data can be pooled to reduce costs and increase efficiency. This is mainly related to the number of tests that are required to recover the labels of each item. In the pooled data problem, the goal is to optimize the recovering process in terms of minimizing the number of tests. This objective raises several questions, such as how many tests are needed to identify each item, how does this depend on the population size and how can the items be pooled strategically to achieve this result?

There are several forms of label recovery through pooling data, based on how the queried pools are determined. Often a distinction between adaptive and non-adaptive group testing is made (Wang, Zhao, & Chuah, 2018).
The first type concerns deterministic setting or non-adaptive group testing: the pools are predetermined and remain fixed. A specific example of this form concerns a random dense setting. Here, the pools are random subsets of the population, and its size is proportional to the total number of items (e.g. see El Alaoui, Ramdas, Krzakala, Zdeborova, and Jordan (2019)). Particularly, each item is considered with a certain constant probability, independently of all other items.
Alternatively, a pool can be determined based on the results of the previously queried pools. This is formally known as adaptive group testing. Though using previous test results to determine the new pools can result in fewer tests, finding an algorithm that does this optimally is more complex. Therefore, non-adaptive group testing is often applied.

2.2 Noisy model

In contrast to the classical pooled data problem, a new variant also considers noise. Here, the tests do not function perfectly and contain some random noise. There are many different types of noise that can be considered. This includes letting the test results deviate according to a certain

distribution, or assuming that an item may be incorrectly identified with a different label.

Similar noise can also occur in the DNA sequencing example that was explained before. An allele may be incorrectly identified as another type, which is referred to genotyping errors. Such noise can arise from several reasons, such as laboratory errors or misinterpreting data (Douglas, Skol, & Boehnke, 2002). Note that a primary difference between this application and the investigated model is that when an item is incorrectly identified, it is uniformly assigned one of the remaining labels, although this may not be cause when considering this application.

Adding noise to the problem makes it more difficult to correctly identify the label of each item. Besides, because there is a possibility that noise will occur in each of the test results, one will never be able to identify the items correctly with absolute certainty. Therefore, it becomes interesting to investigate how confident one can be whether the items are correctly identified using a certain number of test results, or the contrary, how many tests are required to recover the items' labels for some confidence. Besides, like in the classical model, how does this relate to the population's size?

2.3 Results in prior research

Extensive research has been conducted on the required number of tests to recover the items' labels. This particularly concerns information-theoretic bounds. These are bounds that are based on the amount of information needed to solve a given problem. Hahn-Klimroth and Müller (2021) give a more formal definition of this concept as follows. A sequence m_0 is called an information-theoretic lower bound for m when $m \geq m_0$ and the probability of making an error does not tend to one. Similarly, a sequence m_1 is an information-theoretic upper bound for m when $m \leq m_1$ and the probability that the decoder correctly recovers the labels of the items tends to one. In the pooled data problem, such bounds could not be used as an exact bound, however they indicate the order of the number of tests in terms of the population's size.

In the classical model without noise, El Alaoui, Ramdas, Krzakala, Zdeborova, and Jordan (2016) have given an information-theoretic lower bound in the random dense regime, i.e. each item is pooled with a constant probability, all independently. They have proven that when n denotes the total population size, the maximum number of tests required is of order $O(\frac{n}{\log n})$. Scarlett and Cevher (2017) used a different framework to show that an information-theoretic lower bound in the same setting is also of order $\Omega(\frac{n}{\log n})$. However, their results rely on the notion that the decoder works optimally. Nevertheless, it closes the information-theoretic gap.

Scarlett and Cevher (2017) also introduce the concept of random noise in the measurements. They provide a general framework for such models and use that to give an information-theoretic lower bound on the number of items. This bound is given as a maximization over a a number of items whose labels are already revealed before a decoder starts. This framework was then used to illustrate that the notion of noise may result in a much larger information-theoretic lower bound for a few specific examples.

2.4 Research problem

As previously mentioned, a pooled item may be incorrectly identified. This affects the test results for multiple labels. As this is a common form of noise, though it has not been investigated yet, this research will derive an information-theoretic lower bound for the number of tests required to obtain a correct estimate of the item labels under this concept. Specifically, it is investigated how these bounds grow in relation to the population's size and how these depend on the probability of correct identification. Moreover, as the selection of queried pools can be difficult, we are also interested in the differences between these bounds in a deterministic regime, i.e. when the tests are pre-determined, and a random dense setting, where the items are independently selected to be queried with a certain probability.

The noisy model can be specified more precisely as follows. An item is correctly identified in a test with some probability that is independent of the item's label. Moreover, when an item is incorrectly classified, then its assumed that its diagnosed label is uniform over all other labels. This assumption allows for some simplifications during the computations and avoids too complex cases, though it may not occur as commonly in real-life applications.

2.5 Setup

The examples of random noise researched by Scarlett and Cevher (2017) dealt with Gaussian noise in the test results. Specifically, the tests results indicated how many items of each type were present in the queried pool separately, with an additive noise from a Gaussian distribution independently for each type.
As mentioned before, an item may be incorrectly identified. Consequently, the noise that results from this notion affects the test results for multiple labels. Although this form of noise is different than the examples investigated by Scarlett and Cevher (2017), their framework can be used to derive an information-theoretic bound for the number of tests when this concept of noise is considered. Therefore, the relevant parts of their framework is first discussed in section 3.
The main result of their research can then be applied to derive an information-theoretic lower bound. This will be done first for the binary case, i.e. when the number of labels is two. This then serves as a foundation for the general case, where there is an arbitrary number of labels.
For both cases, two different settings regarding the test designs will be investigated. Firstly, it is assumed that the pooled sets are pre-determined and fixed. These computations are then revisited under the random dense regime. To avoid confusion with the concept of random noise in the test results, these cases will be further referred to as the deterministic and Bernoulli settings respectively.

3 Setting

As the computations revolve around applying the main result of Scarlett and Cevher (2017), their framework first needs to be explored.

3.1 Items and their labels

We consider a population $\{1, ..., n\}$, where each of these items has a label in $\{1, ..., d\}$. Let $\pi = (\pi_1, ..., \pi_d)$ be the vector with the proportions having a certain label, i.e. there are a total of $\pi_t n$ items with label $t \in [d]$ in the population. Let $\pi_l := \max_{t\in[d]} \pi_t$ be the largest occurring proportion. The set defined by

$$\mathcal{B}(\pi) = \left\{ \beta \in [d]^n : \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\beta_i = 1\}, ..., \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\beta_i = d\} \right) = \pi \right\} \tag{3.1}$$

is the set of all label vectors β with empirical distribution π. A vector $\beta \in \mathcal{B}(\pi)$ has at its i-th coordinate the label that correspond to the i-th item. We let the choice of β be uniform on this set $\mathcal{B}(\pi)$.
The goal of the pooling data is to find a correct estimate of β. To describe this more formally, let $\hat{\beta}$ be its estimate, then the error probability can be defined as

$$P_e = \mathbb{P}(\beta \neq \hat{\beta}). \tag{3.2}$$

3.2 Tests

We now turn to the setting of the tests. The total number of performed tests is denoted by $m \in \mathbb{N}$. For a test $i = 1, ..., m$, the random vector $X^{(i)} \in \{0, 1\}^n$ indicates which items are included, i.e. $X_j^{(i)} = 1$ when the j-th item is present and 0 otherwise. The number of items in a test $X^{(i)}$ with a certain label $t \in [d]$, as defined in β, is denoted by $N_t(\beta, X^{(i)})$. Let $|X^{(i)}|$ be the total number of items in that test, i.e.

$$|X^{(i)}| = \sum_{t=1}^{d} N_t(\beta, X^{(i)}).$$

Analogously to the measurement vectors, we define $Y^{(i)} \in \mathbb{N}^d$ to be the vector of observation or pool histograms, whose t-th entry indicates how many items in the i-th test are labelled as t.

Following the problem description in the introduction, for any $i = 1, ..., m$, the distribution of $Y^{(i)}$ can be derived as follows. Knowing that any item's type is correctly identified with a (constant) probability $p \in (0, 1)$, we can write

$$Y^{(i)} = \sum_{t=1}^{d} M_t^{(i)}, \tag{3.3}$$

where $M_t^{(i)} \sim \text{Multin}\left(N_t(\beta, X^{(i)}), \boldsymbol{p}^{(t)}\right)$ for $t = 1, ..., d$ are multinomial random variables (with different probability vectors). Here, $\boldsymbol{p}^{(t)} \in (0, 1)^d$ is the vector with probabilities, which can be defined more precisely as follows. For a label $t \in [d]$, we define $\boldsymbol{p}^{(t)}$ to be the vector with at its t-th coordinate the constant probability $p \in (0, 1)$ that a test correctly yields label t, and at its other coordinates $k \neq t$ the probability that the test incorrectly yields label k. Recall that the latter occurs uniformly, therefore

$$p_j^{(t)} = \begin{cases} p & \text{if } j = t, \\ \frac{1-p}{d-1} & \text{if } j \neq t. \end{cases} \tag{3.4}$$

Note that indeed $\sum_j p_j(t) = 1$ for all $t \in [d]$. Moreover, we assume that the labels are independently incorrectly identified, and therefore the random variables $\{M^{(i)}(t)\}_{t\in[d]}$ are assumed to be independent.

3.3 Genie argument

The main theorem of Scarlett and Cevher (2017) gives a lower bound for the number of tests that need to be applied to recover the labels of each item. This bound revolves around the maximization of a vector $\ell = (\ell_1, ..., \ell_d)$ that is defined as follows. For each $t \in [d]$, $\ell_t \in \{0, ..., \pi_t n\}$ denotes the number of items with label t that are still unknown after the labels of $\pi_t n - \ell_t$ items have been revealed. This vector therefore describes how many items of each type still need to be identified.
Accompanying this vector is a random set of items $S_\ell \subseteq \{1, ..., n\}$. The indices in this set are taken such that of each type $t \in [d]$, there exist exactly a total of ℓ_t items in that set. In particular, we let S_ℓ be uniformly distributed over all such sets. This set can now be seen as a set of items whose labels have not yet been revealed.
Moreover, its complement S_ℓ^c can similarly be seen as a set of items that have already been identified. Therefore, having S_ℓ^c being equal to $[n]$, i.e. when S_ℓ is the empty set, amounts to knowing the label of each item.

As the labels of several items are revealed, some possibilities of $\hat{\beta} \in \mathcal{B}(\pi)$, the estimate of the label vector β, can be neglected. To describe this more formally, we define

$$\beta_{S_\ell^c} = \begin{cases} \beta_j, & j \in S_\ell^c, \\ \star, & \text{otherwise.} \end{cases} \tag{3.5}$$

Here, $\star$ represents an unknown value. This vector label describes the labels of the items that still need to be identified. Similarly to the original label vectors, we define $\mathcal{B}_\ell(\pi)$ to be the set of all label vectors that coincide with the known entries of $\beta_{S_\ell^c}$, i.e.

$$\mathcal{B}_\ell(\pi) = \{\gamma \in [d]^n \mid \gamma = (\beta_{S_\ell^c})_i \text{ for all } i = 1, ..., n \text{ where } (\beta_{S_\ell^c})_i \neq \star\}. \tag{3.6}$$

Consequently, $|\mathcal{B}_\ell(\pi)|$ is the number of possible sequences left after the types of the items in S_ℓ have been revealed.
The norm $\|\cdot\|_0$ on a vector indicates how many entries are non-zero. Specifically, $\|\ell\|_0$ indicates of how many types there are items that still need to be recovered.

3.4 Estimating the label vector

Scarlett and Cevher (2017) have proven the following theorem that gives a lower bound for the number of tests needed to achieve a certain level of P_e.

Theorem 1 (Lower bound (Scarlett & Cevher, 2017))**.** *For any decoder, to achieve $P_e \leq \delta$ for any $\delta \in (0, 1)$, it is necessary that*

$$m \geq \max_{\ell : \|\ell\|_0 \geq 2} \frac{(\log |\mathcal{B}_\ell(\pi)|)(1-\delta) - \log 2}{\frac{1}{m}\sum_{i=1}^m I(\beta, Y^{(i)} \mid \beta_{S_\ell^c}, X^{(i)})}. \tag{3.7}$$

We will apply this theorem to find a lower bound for m. First, we will consider the binary case, i.e. $d = 2$, and then further develop the results to a more general case.

3.5 Selecting pooled items in the Bernoulli setting

Finding a lower bound for m requires some knowledge of the test sizes, and in particular the values of $N_t(\beta, X^{(i)})$ for all labels $t \in [d]$ and tests $i \in [m]$. Although we have the obvious bounds $0 \leq N_t(\beta, X^{(i)}) \leq n\pi_t$, more precise bounds are required to obtain more accurate results. Because the test design is a tedious process, we also consider letting this be in the form of a Bernoulli setting. For any test $i \in [m]$, an item is present with a constant probability $\alpha \in (0, 1)$, independently of all other items.

3.6 Simplification of the revealed items

There are many different possibilities of the progress vector ℓ, therefore computing the lower bound of equation (3.7) is rather difficult. Scarlett and Cevher (2017) have shown that the following two specific choices of ℓ are fruitful.

1. For the first option, let for all $t = 1, ..., d$, ℓ_t be either the total number of items with label t, i.e. $\ell_t = n\pi_t$, or 0. Given this, we let $G \subseteq \{1, ..., d\}$ be the set of labels t where ℓ_t is maximal and G^c its complement, i.e.

$$G = \{t \in \{1, ..., d\} : \ell_t = \pi_t n\}.$$

   As a result, the definition of Y as in equation (3.3) can be simplified to

$$(Y^{(i)} | \beta_{S_\ell^c}, X^{(i)}) = \sum_{t \in G} M_t^{(i)}, \tag{3.8}$$

   where M_t for $t \in G$ are similarly defined as in section 3.2, although only when conditional on $\beta_{S_\ell^c}$ and $X^{(i)}$. Analogously to the definition of the distribution vector π, we let

$$\pi_G := (\tfrac{\pi_t}{\sum_{t' \in G} \pi_{t'}})_{t \in G} \in [0, 1]^{|G|}$$

   be the distribution vector where only the types in G are considered.

2. The second simplification is letting $\ell_1 = n\pi_1$, $\ell_2 = 1$ and $\ell_t = 0$ for all $t = 3, ..., d$. In this scenario, all items of type one and a single item with label two are yet to be identified. All other items have been revealed.

As these choices of ℓ have already been proven to be useful in an other example, they are explored in this research as well.

4 Binary case

We will compute a lower bound for the number of tests m by determining the terms in equation (3.7). This is done for both choices of ℓ, as described in section 3.6, and the test vectors $\{X^{(i)}\}_{i\in[m]}$ will assumed to be deterministic. Each of these cases will then be revisited under the assumption that they are not deterministic anymore, but instead in the Bernoullis setting.

The analysis of each of the cases follows a similar strategy. First, the simplification of ℓ is investigated a bit closer, which allows us to find an expression for $|\mathcal{B}_\ell(\pi)|$. The mutual information term as as given in the denominator in Theorem 1, can be written as

$$I(\beta, Y^{(i)}|\beta_{S_l^c}, X^{(i)}) = H(Y^{(i)}|\beta_{S_l^c}, X) - H(Y^{(i)}|\beta, X). \tag{4.1}$$

Hence, bounding this expression entails finding an upper and lower bound for $H(Y^{(i)}|\beta_{S_\ell^c}, X)$, which will be referred to as the progressed entropy, and $H(Y^{(i)}|\beta, X)$, the so-called final entropy, respectively.
We are required to determine the average of these mutual information terms over the m tests (see equation (3.7)). Therefore, we will focus in general on an arbitrary choice of test $i = 1, ..., m$ and derive bounds for the final and progressed entropy independent on the choice of test. To ease the notations, we will omit the superscript $(\cdot)^{(i)}$ when denoting the test vectors $X^{(i)}$ and $Y^{(i)}$ and abbreviate $N_t(\beta, X^{(i)})$ to N_t.

4.1 Simplification 1

The simplification of the choice of ℓ, as described in section 3.6, is limited to having $l_t \in \{0, \pi_t n\}$ for $t \in \{1, 2\}$. Having both values set equal to 0 is equivalent to knowing the label of each item. In that case, the mutual information is undefined, as β is deterministic given this specific $\beta_{S_\ell^c}$, more precisely since these vectors are the same. Moreover, having exactly one of π_1 or π_2 equal to its maximum value entails knowing exactly which items are of one specific type, and hence also the items of the other type. Consequently, β is again deterministic given this specific choice of $\beta_{S_\ell^c}$. This leaves the only option of this form of simplification to be $\ell = (\pi_1 n, \pi_2 n)$. From this it follows that $G = \{1, 2\}$, so that by equation (3.8), we still have $Y = M_1 + M_2$.

First, we will focus on finding a lower bound for the final entropy. This is done by first using the characterization of Y, given β and X, as stated in (3.3), and then reducing the problem to computing the lower bound of a binomial random variable.
Then, to determine an upper bound for the progressed entropy, we first reduce the problem such that we are left with finding the conditional entropy of the entries of Y given $\beta_{S_\ell^c}$ and X, separately. We then derive the conditional distributions of these random variables, specifically given $N_t(\beta_{S_\ell^c}, X)$ for $t = 1, 2$, again using the characterization of Y as used to determine the final entropy. Note that N_t is not deterministic anymore, yet under the given simplifications of ℓ and knowing $\beta_{S_\ell^c}$ and X, its distribution can be derived. These results can then be used to derive the variance of $(Y_t|N_t)$, which in turn are used to find an upper bound for the progressed entropy.
The entropies can be combined in equation (4.1) to bound the mutual information terms, and hence the average of these terms over the tests. Finally, the last component $|\mathcal{B}_\ell(\pi)|$ is determined using a characterization by Scarlett and Cevher (2017).

4.1.1 Lower bound initial entropy

To bound the initial entropy, several steps need to be taken. Instead of directly computing the entropy $H(Y|\beta, X)$, we first split this up into several entropies which can be computed more easily. This reduces the problem to computing the entropy of a binomial random variable, whose size parameter is still random.

Recall that Y can be written as a sum of multinomial random variables. In particular, as in the binary case $\ell = (\pi_1 n, \pi_2 n)$, this gives

$$Y = M_1 + M_2 \tag{4.2}$$

where $(M_t|\beta, X) \sim \text{Multin}(N_t, \boldsymbol{p}^{(t)})$ for $t = 1, 2$ are assumed to be independent. The following lemma avoids having to determine the distribution of Y by splitting the entropies.

Lemma 2. *Let $X_1, ..., X_n$ be $n \in \mathbb{N}$ independent discrete random variables. Then*

$$H\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} H(X_i) - H\left(X_i|\sum_{j=1}^{i} X_j\right). \tag{4.3}$$

For the proof of this lemma, see Appendix A. In the binary case, this implies that

$$H(Y|\beta, X) = H(M_1 + M_2|\beta, X) = H(M_1|\beta, X) + H(M_2|\beta, X) - H(M_1|M_1 + M_2, \beta, X). \tag{4.4}$$

Whereas the conditional distributions of M_1 and M_2 given β and X are relatively straightforward, their conditional distribution also given $M_1 + M_2$ is not. To deal with this, the latter conditional entropy can be simplified further using the following lemma.

Lemma 3 (Conditioning reduces entropy (Cover & Thomas, 2006))**.** *Let X and Y be discrete random variables. Then*

$$H(X|Y) \leq H(X) \tag{4.5}$$

with equality if and only if X and Y are independent.

The proof of this lemma can be found in Appendix A. We apply this lemma to the last conditional entropy in (4.4), yielding

$$H(Y|\beta, X) \geq H(M_1|\beta, X) + H(M_2|\beta, X) - H(M_1|\beta, X) = H(M_2|\beta, X). \tag{4.6}$$

In the binary case, the second entry of M_t is deterministic given its first (and vice-versa). Therefore, this amounts to computing the entropy of a binomial random variable with size parameter N_2 and probability $1 - p$.
However, as addition is commutative, M_1 and M_2 can be interchanged in equation (4.6). As a result,

$$H(Y|\beta, X) \geq \max_{t \in \{1,2\}} H(M_t|\beta, X). \tag{4.7}$$

Now the problem has been reduced to computing the entropy of a binomial random variable. However, its size parameter is random, as it is determined by the random variable β. We will focus on the conditional entropy of M_t, given β and X for an arbitrary choice of $t = 1, 2$. The definition of conditional entropy in (1.2) can be used to compute this expression. It involves averaging over the entropies of the conditional variable M_t, given a specific β and X, with respect to β, i.e.

$$H(M_t|\beta, X) = \sum_{\gamma \in \mathcal{B}(\pi)} f_\beta(\gamma) H(M_t|\beta = \gamma, X).$$

However, the many different possibilities of β make these computations complicated. Since the (conditional) entropy of a random variable is always non-negative (e.g. see Cover and Thomas (2006)), we can bound this expression by summing over a limited choice of γ. In particular, only consider the values of γ such that $N_t(\gamma, X) = \pi_t|X|$. Because β was drawn uniformly from $\mathcal{B}(\pi)$, this expression can now be simplified to

$$\begin{aligned}\sum_{\gamma \in \mathcal{B}(\pi)} f_\beta(\gamma) H(M_t|\beta = \gamma, X) &= \sum_{\gamma \in \mathcal{B}(\pi)} \frac{1}{|\mathcal{B}(\pi)|} H(\text{Bin}(N_t(\gamma, X), p) \\ &\geq \frac{|\{\gamma : N_t(\gamma, X) = \pi_t|X|\}|}{|\mathcal{B}(\pi)|} H(\text{Bin}(\pi_t|X|, \beta)).\end{aligned}$$

There are three terms that need to be computed. Firstly, Scarlett and Cevher (2017) have shown that

$$|\mathcal{B}(\pi)| = e^{n(H(\pi)+o(1))}. \tag{4.8}$$

Moreover, the entropy of a binomial variable is formulated in the following lemma.

Lemma 4 (Entropy of a binomial random variable)**.** *Let $Z \sim \mathrm{Bin}(n,p)$ be a binomial random variable. Then its entropy is given by*

$$H(Z) = \frac{1}{2}\log_2\left(2\pi enp(1-p)\right) + O\left(\frac{1}{n}\right). \tag{4.9}$$

The proof of this lemma is given in Appendix A.
Now we will examine $|\{\gamma \in \mathcal{B}(\pi) : \pi_t|X|\}|$. There are a total of $|X|$ items considered in a test. For any γ in this set, the number of items of type t that are considered in the test is constant, and their arrangement is independent. These items can be sorted in $\binom{|X|}{\pi_t|X|}$ ways. Similarly, the number of items that are not take into account during the test are $n-|X|$ and can be arranged in a total of $\binom{n-|X|}{\pi_t n-\pi_t|X|}$ ways. Therefore,

$$|\{\gamma \in \mathcal{B}(\pi) : \pi_t|X|\}| = \binom{|X|}{\pi_t|X|}\binom{n-|X|}{\pi_t n-\pi_t|X|} = e^{nH(\pi)}.$$

These results can be combined to find

$$H(M_t|\beta,X) \geq \frac{e^n H(\pi)}{e^{n(H(\pi)+o(1))}} H(\mathrm{Bin}(\pi_t|X|,\beta)) \geq \frac{1}{2}\log_2\left(2\pi e\pi_t|X|p(1-p)\right).$$

Now maximizing over t gives the final result

$$H(Y|\beta,X) \geq \max_{t\in\{1,2\}} \frac{1}{2}\log_2\left(2\pi e\pi_t|X|p(1-p)\right). \tag{4.10}$$

4.1.2 Upper bound progressed entropy

To compute an upper bound for the progressed entropy, first the term $H(Y|\beta_{S_\ell^c},X)$ is simplified using some computational rules for entropies. This reduces the problem to computing the conditional entropy of a random variable, of which the distribution can be more easily derived, though an explicit expression is difficult to give. Therefore a lemma, which bounds an random variable's entropy based on its variance, is applied, allowing to avoid having to determine this explicit expression of the distribution.

First, recall that $\ell = (\pi_1 n, \pi_2 n)$, meaning that $\beta_{S_\ell^c}$ only contains $\star$, as defined in (3.5). Therefore, this variable does not introduce any new information, and hence

$$H(Y|\beta_{S_\ell^c},X) = H(Y|X).$$

To find an upper bound for $H(Y|X)$, the following lemma from Cover and Thomas (2006) can be applied.

Lemma 5 (Chain rule for entropy (Cover & Thomas, 2006))**.** *Let $(X_1,...,X_n)$ be a random vector. Then its entropy is equal to*

$$H(X_1,...,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},...,X_1). \tag{4.11}$$

A proof of this lemma is given in Appendix A. In the binary case, this would result in

$$H(Y|X) = H(Y_1|X) + H(Y_2|X,Y_1) = H(Y_1|X) \tag{4.12}$$

as Y_2 is determinate given Y_1 (here, Lemma 3 is applied). Therefore, the characterization of Y_1 given X is of interest.

The multinomial distribution is a generalization of the binomial distribution. Given that N_1 nodes actually have label one, the test will reveal that each of these nodes indeed is labelled one with a constant probability p. This clearly is related to the random variable $B_1 \sim \text{Binomial}(N_1, p)$. Moreover, the nodes with an actual label than one, i.e. here two, will be identified as type one with a probability $\frac{1-p}{d-1}$. In the binary case, this probability simplifies to $1-p$, and there are a total of N_2 such nodes. Similarly to before, this is related to the random variable $B_2 \sim \text{Binomial}(N_2, 1-p)$. The random variable Y_1 can be now expressed as a sum of B_1 and B_2.
However, given X, the value of N_t for any $t \in [d]$ is not deterministic. Yet, its distribution can be derived as follows. For a test, since no item has been revealed yet, $|X|$ items will be randomly selected out of a total of n items. Exactly $n\pi_t$ of all items are of type t. This describes a hypergeometric distribution, specifically $N_t \sim \text{Hyp}(|X|, \pi_t n, n)$ for all $t = 1, 2$.

A closed expression of the distribution of $(Y_1|X)$ is difficult to determine as N_1 and N_2 are not deterministic. The following lemma allows to surpass this problem by instead computing its variance.

Lemma 6 (Massey, 1988)**.** *Let Z be an integer valued random variable with finite variance σ^2. Then its entropy is bounded from above by*

$$H(Z) < \frac{1}{2}\log\left(2\pi e\left(\sigma^2 + \frac{1}{12}\right)\right). \tag{4.13}$$

To compute the variance, traditionally an exact distribution needs to be known. However, the following Lemma allows us to condition on N_t.

Lemma 7 (Law of total variance)**.** *Let U and V be random variables on the same probability space such that the variance of Y is finite. Then*

$$\text{Var}(U) = \mathbb{E}[\text{Var}(U|V)] + \text{Var}(\mathbb{E}[U|V]). \tag{4.14}$$

The proof of this lemma is given in Appendix A. The variance and mean of a hypergeometric random variable are given and proven in Appendix B. Recall that $|X| = N_1 + N_2$ is deterministic given X. Therefore, it can be found that

$$\begin{aligned}
\text{Var}\left(B_1 + B_2|X\right) &= \mathbb{E}[\text{Var}\left(B_1 + B_2|N_1\right)|X] + \text{Var}\left(\mathbb{E}[B_1 + B_2|N_1, N_2)\right)|X] \\
&= \mathbb{E}[\text{Var}\left(B_1|N_1\right) + \text{Var}\left(B_2|N_2\right)|X] + \text{Var}\left(\mathbb{E}[B_1|N_1] + \mathbb{E}[B_2|N_2]|X\right) \\
&= \mathbb{E}[N_1 p(1-p) + N_2 p(1-p)|X] + \text{Var}\left(N_1 p + N_2(1-p)|X\right) \\
&= |X|p(1-p) + \text{Var}\left(N_1 p + (|X| - N_1)(1-p)|X\right) \\
&= |X|p(1-p) + \text{Var}\left(N_1(2p-1)|X\right) \\
&= |X|p(1-p) + (2p-1)^2|X|\frac{\pi_1 n}{n}\cdot\frac{n - n\pi_1}{n}\cdot\frac{n-\pi_1 n}{n-1} \\
&\leq |X|p(1-p) + (2p-1)^2|X|\pi_1(1-\pi_1).
\end{aligned}$$

Now by applying Lemma 6, it can be concluded that

$$H(Y|\beta_{S_\ell^c}, X) \leq \frac{1}{2}\log_2\left(2\pi e\left(|X|p(1-p) + (2p-1)^2|X|\pi_1(1-\pi_1) + \frac{1}{12}\right)\right). \tag{4.15}$$

4.1.3 Number of label vectors

The final component to compute is $|\mathcal{B}_\ell(\boldsymbol{\pi})|$, i.e. the number of label vectors β that have an empirical distribution $\boldsymbol{\pi}$ and coincide with $\beta_{S_\ell^c}$. In the binary case, we have already concluded that the simplification of the choice of ℓ must be $\ell = (\pi_1 n, \pi_2 n)$. As a result, this component is equivalent to $|\mathcal{B}(\boldsymbol{\pi})|$. Moreover, Scarlett and Cevher (2017) have shown that

$$|\mathcal{B}(\boldsymbol{\pi})| = e^{n(H(\boldsymbol{\pi}) + o(1))}. \tag{4.16}$$

4.1.4 Lower bound number of tests

The main results from the previous sections, that can be found in equations (4.10) and (4.15), can be combined to find the mutual information $I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)})$ for any test $i = 1, ..., m$. Using these results, it follows that

$$\begin{aligned} I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)}) &= H(Y^{(i)}|\beta_{S_\ell^c}, X^{(i)}) - H(Y^{(i)}|\beta, X^{(i)}) \\ &\leq \frac{1}{2}\log_2\left(2\pi e|X^{(i)}|\left(p(1-p) + (2p-1)^2\pi_1(1-\pi_1) + \frac{1}{12|X^{(i)}|}\right)\right) \\ &\quad - \max_{t\in\{1,2\}} \frac{1}{2}\log_2\left(2\pi e\pi_t|X^{(i)}|p(1-p)\right) \\ &\leq \min_{t\in\{1,2\}} \frac{1}{2}\log_2\left(1 + \frac{(2p-1)^2}{p(1-p)}\pi_1(1-\pi_1) + \frac{1}{12|X^{(i)}|p(1-p)}\right). \end{aligned}$$

Observe that this bound is dependent on the choice of test. As we are required to average over these mutual-information terms with respect to the tests, a bound on $|X^{(i)}|$ needs to be used. Because $|X^{(i)}| \geq 1$, as otherwise the test would be redundant, we now find

$$\frac{1}{m}\sum_{i=1}^{m} I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)}) \leq \frac{1}{2}\log_2\left(1 + \frac{(2p-1)^2}{p(1-p)}\pi_1(1-\pi_1) + \frac{1}{12p(1-p)}\right). \tag{4.17}$$

Then, combining this with the result in (4.16) and applying Theorem 1 yields the condition that for $P_e \leq \delta$ for any $\delta \in (0,1)$, it is necessary that,

$$m \geq \frac{n(H(\pi) + o(1))(1-\delta) - \log 2}{\frac{1}{2}\log_2\left(1 + \frac{(2p-1)^2}{p(1-p)}\pi_1(1-\pi_1) + \frac{1}{12p(1-p)}\right)} \tag{4.18}$$

as $n \to \infty$.

4.2 Simplification 2

The computations regarding the lower bound for m using the second simplification of ℓ are based on the same principles as in the previous sections. In particular, the mutual terms are again split up as in (4.1). The only terms that remain dependent on the choice of ℓ are the number of potential label vectors $|\mathcal{B}_\ell(\pi)|$ and the progressed entropy $H(Y|\beta_{S_\ell^c}, X)$. We will first consider the latter and bound it using lemma 6, which states that the entropy of a variable can be bounded when its variance is known.

4.2.1 Bounding the mutual information

The lower bound of the final entropy remains the same. For the progressed entropy, the results in equation (4.12) still hold, although there the condition on $\beta_{S_\ell^c}$ is missing.
Recall that $|X|$ is the total test size, i.e. $|X| = N_1 + N_2$. This random variable is deterministic given and X. Now it is possible to rewrite $N_1 = |X| - N_2$. Moreover, recall that Y_1 can be written as a sum of binomial variables B_1 and B_2, where $(B_1|\beta_{S_\ell^c}, X) \sim \mathrm{Bin}(|X| - N_2, p)$ and $(B_2|\beta_{S_\ell^c}, X) \sim \mathrm{Binomial}(N_2, 1-p)$. Using the definition of variance, it can be found that

$$\begin{aligned} \mathrm{Var}\left(Y_1|\beta_{S_\ell^c}, X\right) &= \mathbb{E}[Y_1^2|\beta_{S_\ell^c}, X] - \mathbb{E}[Y_1|\beta_{S_\ell^c}, X]^2 \\ &= \mathbb{E}[(B_1+B_2)^2|\beta_{S_\ell^c}, X] - \mathbb{E}[B_1+B_2|\beta_{S_\ell^c}, X]^2. \end{aligned}$$

Now given the simplification of ℓ, the random variable N_2 takes values in $\{0,1\}$ and indicates whether the single item of type two that has not been revealed yet is taken into account. Let

$B_{\text{tot}} \sim \text{Bin}(|X|, p)$. Now this expression can be bounded by

$$\begin{aligned}\text{Var}\left(Y_1|\beta_{S_\ell^c}, X\right) &\leq \mathbb{E}[(B_{\text{tot}}+1)^2|\beta_{S_\ell^c}, X] - (\mathbb{E}[B_{\text{tot}}-1|\beta_{S_\ell^c}, X])^2 \\ &= \mathbb{E}[B_{\text{tot}}^2 + 2B_{\text{tot}} + 1|\beta_{S_\ell^c}, X] - (\mathbb{E}[B_{\text{tot}}|\beta_{S_\ell^c}, X]^2 - 2\mathbb{E}[B_{\text{tot}}|\beta_{S_\ell^c}, X] + 1) \\ &= |X|p(1-p) + |X|^2p^2 + 2|X|p + 1 - (|X|^2p^2 - 2|X|p + 1) \\ &= |X|p(5-p).\end{aligned}$$

Now by applying Lemma 6, it follows that

$$H(Y|\beta_{S_\ell^c}, X) \leq \frac{1}{2}\log_2\left(2\pi e\left(|X|p(5-p) + \frac{1}{12}\right)\right). \tag{4.19}$$

Combining this result with the bound of $H(Y|\beta, X)$ given in (4.10) yields

$$\begin{aligned}I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\ &\leq \frac{1}{2}\log_2\left(2\pi e\left(|X|p(5-p) + \frac{1}{12}\right)\right) - \max_{t\in\{1,2\}}\frac{1}{2}\log_2\left(2\pi e\pi_t|X|p(1-p)\right) \\ &\leq \min_{t\in\{1,2\}}\frac{1}{2}\log_2\left(\frac{5-p}{\pi_t(1-p)} + \frac{1}{12\pi_t p(1-p)}\right).\end{aligned} \tag{4.20}$$

In the last inequality, it is used that $|X| \geq 1$. This is necessary to compute the average of these mutual-information terms with respect to the tests.

4.2.2 Number of label vectors

The last component to determine is $|\mathcal{B}_\ell(\pi)|$. With this specific choice of ℓ, only $n\pi_1 + 1$ items remain unlabeled. Specifically, when the single item of type two is identified, all other items will be of type one. Therefore, it follows that

$$|\mathcal{B}_\ell(\pi)| = n\pi_1 + 1. \tag{4.21}$$

4.2.3 Lower bound number of tests

We collect the results from the previous two sections to find a lower bound on m. As $n \to \infty$, to obtain $P_e \leq \delta$ for any $\delta \in (0,1)$, it is necessary that

$$m \geq \frac{\log(n\pi_1+1)(1-\delta) - \log 2}{\min_{t\in\{1,2\}}\frac{1}{2}\log_2\left(\frac{5-p}{\pi_t(1-p)} + \frac{1}{12\pi_t p(1-p)}\right)}. \tag{4.22}$$

4.3 Bernoulli setting

In the Bernoulli setting, each item is pooled with a certain probability, independent of all other items. Therefore, a test size $|X|$ is not deterministic anymore. Nevertheless, most computations from the deterministic setting can be replicated in this new context. For both simplifications, the progressed and final entropies are revisited and adapted to the new situation. The number of label vectors given a simplification of ℓ remains the same and can be directly referred to the deterministic setting.

4.3.1 Simplification 1

To compute a lower bound for $H(Y|\beta, X)$ when the test's design concerns a Bernoulli process, the result in equation (4.10) still holds. However as X is now random, we also average over this

variable. Using the results from before, it follows that

$$\begin{aligned} H(M_t|\beta, X) &= \sum_{x\in\mathcal{X}} f_X(x) H(M_t|\beta, X = x) \\ &\geq \sum_{x\in\mathcal{X}} f_X(x) \frac{1}{2}\log_2\left(2\pi e \pi_t |x| p(1-p)\right) \\ &= \frac{1}{2}\log_2\left(2\pi e \pi_t p(1-p)\right) + \frac{1}{2}\mathbb{E}[\log(|X|)]. \end{aligned}$$

As an item is present in a test with a probability $\alpha \in (0,1)$, $|X|$ has a binomial distribution with size n and probability α. Flajolet (1999) has shown that

$$\mathbb{E}\left[\log(|X|)\right] = \log(\alpha n) + O\left(\frac{1}{(\alpha n)^2}\right). \tag{4.23}$$

Substituting this result and maximizing over $t \in [d]$, as is done in (4.7), finally gives

$$H(Y|\beta, X) \geq \max_{t\in[d]} \frac{1}{2}\log_2\left(2\pi e \pi_t \alpha n p(1-p)\right) + O\left(\frac{1}{n^2}\right). \tag{4.24}$$

Here we used that α does not grow with n but is constant.
The same strategy can be used to compute $H(Y|\beta_{S_\ell^c}, X)$. Recall the result given in (4.15), now

$$\begin{aligned} H(Y|\beta_{S_\ell^c}, X) &= \sum_{x\in\mathcal{X}} f_X(x) H(Y|\beta_{S_\ell^c}, X = x) \\ &\leq \sum_{x\in\mathcal{X}} f_X(x) \frac{1}{2}\log_2\left(2\pi e\left(|x|p(1-p) + (2p-1)^2|x|\pi_1(1-\pi_1) + \frac{1}{12}\right)\right) \\ &\leq \sum_{x\in\mathcal{X}} f_X(x) \frac{1}{2}\log_2\left(2\pi e|X|\left(p(1-p) + (2p-1)^2\pi_1(1-\pi_1) + \frac{1}{12}\right)\right) \\ &= \frac{1}{2}\log_2\left(2\pi e\alpha n\left(p(1-p) + (2p-1)^2\pi_1(1-\pi_1) + \frac{1}{12}\right)\right) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

The last equality follows from the result in (4.23) by Flajolet (1999). These expressions can be used to derive a bound for the mutual information term by computing

$$\begin{aligned} I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\ &\leq \frac{1}{2}\log_2\left(2\pi e\alpha n\left(p(1-p) + (2p-1)^2\pi_1(1-\pi_1) + \frac{1}{12}\right)\right) \\ &\quad - \max_{t\in[d]} \frac{1}{2}\log_2\left(2\pi e\pi_t\alpha n p(1-p)\right) + O\left(\frac{1}{n^2}\right) \\ &= \min_{t\in[d]} \frac{1}{2}\log_2\left(\frac{1}{\pi_t}\left(1 + \frac{(2p-1)^2}{p(1-p)}\pi_1(1-\pi_1) + \frac{1}{12p(1-p)}\right)\right) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

This bound matches the result in (4.17). Since $\mathcal{B}_\ell(\pi)$ is independent of X, the lower bound on the number of tests is the same as in the deterministic case in (4.18).

4.3.2 Simplification 2

The final entropy is independent of ℓ and can therefore be directly taken from (4.24). As in the first simplification of ℓ, here the progressed entropy need to be revisited. Recall the results given in (4.19), which gives an upper bound for $H(Y|\beta_{S_\ell^c}, X)$ in the deterministic setting. By again

averaging over X, as it has become random, we obtain

$$\begin{aligned} H(Y|\beta_{S_\ell^c}, X) &= \sum_{x\in\mathcal{X}} f_X(x) H(Y|\beta_{S_\ell^c}, X = x) \\ &\leq \sum_{x\in\mathcal{X}} f_X(x) \frac{1}{2}\log_2\left(2\pi e|X|\left(p(5-p)+\frac{1}{12}\right)\right) \\ &= \frac{1}{2}\log_2\left(2\pi e\alpha n\left(p(5-p)+\frac{1}{12}\right)\right) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

The last equality follows from (4.23). The bound for the final entropy from the first simplification can be combined with this result to find

$$\begin{aligned} I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\ &\leq \frac{1}{2}\log_2\left(2\pi e\alpha n\left(p(5-p)+\frac{1}{12}\right)\right) - \max_{t\in\{1,2\}} \frac{1}{2}\log_2\left(2\pi e\pi_t\alpha np(1-p)\right) \\ &= \min_{t\in\{1,2\}} \frac{1}{2}\log_2\left(\frac{5-p}{\pi_t(1-p)} + \frac{1}{12\pi_t p(1-p)}\right), \end{aligned}$$

which exactly matches the mutual information term found in the deterministic setting in (4.20). As $|\mathcal{B}_\ell(\pi)|$ is similar in both regimes, the bound in (4.22). for the number of tests m needed for recovery also holds in the Bernoulli setting.

4.4 Results

This obtained lower bound is still a rather complex equation. The results given in table 1 indicate the orders of the lower bound in terms of n for several choices of p, i.e. constant or inversely proportionate to n. When no assumptions are made on the tests, the orders are still the same despite the differences in p. This holds for both simplifications of ℓ. However, when the tests design concerns a Bernoulli process, some differences appear. As p approaches 1, the order of the information-theoretic bound on m becomes smaller.

Equation	$p = \Theta(1)$	$p = 1 - \frac{1}{n}$	$p = 1 - \frac{1}{\sqrt{n}}$
(4.18)	$\Omega(n)$	$\Omega(\frac{n}{\log n})$	$\Omega(\frac{n}{\log n})$
(4.22)	$\Omega(\log n)$	$\Omega(1)$	$\Omega(1)$

Table 1: Orders of the number of tests m for several choices of p in the binary case under the deterministic and Bernoulli setting

5 General case

The following computations largely follow the setup given in the binary case. We again start with the first and second simplifications in the deterministic setting, and then revisit them under the Bernoulli setting. The computations mainly revolve around determining the initial and progressed entropy, which will be covered first for each section.

5.1 Simplification 1

The main components that need to be determined are the entropies as given in equation (4.1). The strategy for the computations of these terms is similarly to the structure followed in the binary case in section 4.1.
The average of the mutual terms $I(Y^{(i)}, \beta|\beta_{S_\ell^c}, X^{(i)})$ over the tests $i = 1, ..., m$ is determined by bounding the mutual term for an arbitrary test. This can again be achieved by splitting it up into a final and progressed entropy as defined in (4.1). A lower bound for the final entropy can be found by first simplifying the entropy to that of a multinomial variable. The progressed entropy can be bounded by first splitting Y up into its components and then by applying lemma 6. Because we will focus on an arbitrary choice of $i = 1, ..., m$, we will omit the superscript $(\cdot)^{(i)}$ to simplify the notation.
Finally, the numer of label vectors $|\mathcal{B}_\ell(\pi)|$ after progression can be determined using the a characterization analogously to that in equation (4.16).

5.1.1 Lower bound initial entropy

The same strategy as in the binary case can be used to bound $H(Y|\beta, X)$. This involves first simplifying this term using the characterization of Y as given in (3.3), simplifying the problem to determining the entropy of a multinomial random variable, where its size parameter is also random. Then the definition of conditional entropy can be used to find an expression for this resulting term. First recall that Y can be written as a sum of d independent multinomial random variables $M_1 + ... + M_d$. Now again using Lemma 2, that characterizes the entropy of a sum of independent random variables, we now find that

$$H(Y|\beta, X) = H(\sum_{i=1}^{d} M_i|\beta, X) = \sum_{i=1}^{d} H(M_i|\beta, X) - H(M_i|\beta, X, \sum_{j=1}^{i} M_i).$$

By carefully selecting the conditional random variables analogously to Lemma 3, it follows that

$$\begin{aligned} H(Y|\beta, X) &= \sum_{i=1}^{d} H(M_i|\beta, X) - H(M_i|\beta, X, \sum_{j=1}^{i} M_i) && (5.1)\\ &= H(M_1|\beta, X) - H(M_1|\beta, X, M_1) + \\ &\quad + \left[\sum_{i=2}^{d} H(M_i|\beta, X) - H(M_i|\beta, X, \sum_{j=1}^{i} M_i)\right] && (5.2)\\ &\geq H(M_1|\beta, X) - \left[\sum_{i=2}^{d} H(M_i|\beta, X) - H(M_i|\beta, X)\right] && (5.3)\\ &= H(M_1|\beta, X). && (5.4) \end{aligned}$$

The term $H(M_1|\beta, X, M_1)$ disappears in (5.3) since M_1 is clearly deterministic given M_1, resulting in the entropy being equal to 0 as stated in Lemma 3. Since addition is a commutative operation, any of the variables $\{M_t\}_{t\in[d]}$ can be swapped and therefore

$$H(Y|\beta, X) \geq \max_{t\in[d]} H(M_t|\beta, X). \tag{5.5}$$

We will now focus on determining $H(M_t|\beta, X)$ for an arbitrary choice of $t \in [d]$.
Cichon and Golebiewski (2012) have proven that for a multinomial random variable $M \sim \text{Multin}(n, \boldsymbol{p})$ with dimension $d \in \mathbb{N}$, its entropy is equal to

$$H(M) = \frac{1}{2}\log\left((2\pi ne)^{d-1}\prod_{i=1}^{d} p_i\right) + \frac{1}{12n}\left(3d - 2 - \sum_{i=1}^{d}\frac{1}{p_i}\right) + O\left(\frac{1}{n^2}\right). \tag{5.6}$$

Recall that β was chosen uniformly on $\mathcal{B}(\pi)$. Now using the definition of conditional entropy (see (1.2)), we find

$$\begin{aligned} H(M_t|\beta, X) &= \sum_{\gamma\in\mathcal{B}(\pi)} f_\beta(\gamma) H(M_t|\beta = \gamma, X) \\ &= \sum_{\gamma\in\mathcal{B}(\pi)} \frac{1}{|\mathcal{B}(\pi)|} H(M_t|\beta = \gamma, X). \end{aligned}$$

In the binary case, a similar expression could be bounded by summing over a more selective choice of $\gamma \in \mathcal{B}(\pi)$. The same strategy can be applied in this general case. Consider the set defined by

$$\mathcal{B}'(\pi, X) = \{\gamma \in \mathcal{B}(\pi) : N_i(\gamma, X) = \pi_i|X|, i = 1, ..., d\} \tag{5.7}$$

We will sum over the elements in this set rather than $\mathcal{B}(\pi)$, which will result in a weaker bound that can be more easily computed.
First, we will determine the cardinality of $\mathcal{B}'(\pi, X)$. Any test in this set contains exactly $\pi_i|X|$ items of type $i \in [d]$. These can be swapped in $\binom{|X|}{\pi_1|X|,...,\pi_d|X|}$ ways. A total of $n - |X|$ items are not taken into account, which can similarly be ordered in $\binom{n-|X|}{\pi_1(n-|X|),...,\pi_d(n-|X|)}$ ways. As these ways in which these items are sorted is independent of the way the pooled items are arranged, it now follows that

$$|\mathcal{B}'(\pi, X)| = \binom{|X|}{\pi_1|X|, ..., \pi_d|X|}\binom{n - |X|}{\pi_1 n - \pi_1|X|, ..., \pi_d n - \pi_d|X|} = e^{nH(\pi)}. \tag{5.8}$$

Recall that $|\mathcal{B}(\pi)| = e^{n(H(\pi)+o(1))}$. These results can be combined to obtain

$$\begin{aligned} H(M_t|\beta, X) &\geq \sum_{\gamma\in\mathcal{B}'(\pi,X)} \frac{1}{|\mathcal{B}(\pi)|} H(\text{Multin}(\pi_t|X|, \boldsymbol{p}^{(t)})) \\ &= \frac{|\mathcal{B}'(\pi, X)|}{|\mathcal{B}(\pi)|}\left(\frac{1}{2}\log\left((2\pi e\pi_t|X|)^{d-1}pq^{d-1}\right) + \frac{1}{12\pi_t|X|}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right)\right) \\ &= \frac{d-1}{2}\log\left(2\pi e\pi_t|X|q\right) + \frac{1}{2}\log(p) + \frac{1}{12\pi_t|X|}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right). \end{aligned}$$

Recall that $\pi_l := \max_{t\in[d]} \pi_t$. Now maximizing over $t \in [d]$ as is required in (5.5) yields

$$\begin{aligned} H(Y|\beta, X) &\geq \max_{t\in[d]} \frac{d-1}{2}\log\left(2\pi e\pi_t|X|q\right) + \frac{1}{2}\log(p) + \frac{1}{12\pi_t|X|}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right) \\ &\geq \frac{d-1}{2}\log\left(2\pi e\pi_l|X|q\right) + \frac{1}{2}\log(p) + \frac{1}{12\pi_l|X|}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right). \end{aligned}$$

5.1.2 Upper bound progressed entropy

To find an upper bound for $H(Y|\beta_{S_\ell^c}, X)$, the same strategy as the binary case can be applied. However, instead of directly simplifying this to the entropy of a single variable, as in equation (4.12), we need to compute multiple entropies. This is because given $Y_1, ..., Y_{d-1}$, only then is the last

random variable Y_d deterministic.
Using Lemma 3 and the chain rule for entropy (Lemma 5), it can be found that

$$H(Y|\beta_{S_\ell^c}, X) = \sum_{i=1}^{d} H(Y_i|\beta_{S_\ell^c}, X, Y_{i-1}, ..., Y_1) \tag{5.9}$$

$$= \sum_{i=1}^{d-1} H(Y_i|\beta_{S_\ell^c}, X, Y_{i-1}, ..., Y_1) \tag{5.10}$$

$$\leq \sum_{i=1}^{d-1} H(Y_i|\beta_{S_\ell^c}, X). \tag{5.11}$$

We will now focus on the entropy of Y_t for any $t \in \{1, ..., n-1\}$. Analogously to the binary case, we can write Y_t as a sum of binomial variables. However, an item of a type other than t is identified as t with a constant probability of $q = \frac{1-p}{d-1}$. Moreover, the items of type in G^c already have been identified and can therefore be neglected. This results in the simplification

$$Y_t = \sum_{i=1}^{d} B_i^{(t)} = \sum_{i \in G} B_i^{(t)}. \tag{5.12}$$

Here, we define

$$B_i^{(t)} \sim \begin{cases} \mathrm{Bin}(N_t, p) & i = t, \\ \mathrm{Bin}(N_i, q) & i \neq t. \end{cases} \tag{5.13}$$

for $i, t \in [d]$. The latter sum of indices in G follows from $N_i \equiv 0$, and thus B_i as well, for all $i \notin G$, as the items of this type have already been revealed.
To compute the variance of these random variables, the distribution of N_i for any $i \in G$ should be known. Similarly to the binary case, this concerns a hypergeometric distribution where the parameters can be derived as follows. Let S_G be the set of items that in $\beta_{S_\ell^c}$ equal $\star$ and are present in test X. Hence, this set contains the items that are yet to be identified in this test. Let $m_g := |S_G|$ be its cardinality. Let $p_G := \sum_{t \in G} \pi_t n$ be the total number of items that still need to be identified. Now we are trying to select m_G items out of a collection of p_G items, where $\pi_t n$ are of type t. From this it follows that $N_t \sim \mathrm{Hyper}(m_G, \pi_t n, p_G)$ for all $t \in G$.
Now by invoking the law of total variance (Lemma 7),

$$\begin{aligned}
\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right) &= \mathrm{Var}\left(\sum_{i \in G} B_i|\beta_{S_\ell^c}, X\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\sum_{i \in G} B_i|N_1, ..., N_d\right) \mid \beta_{S_\ell^c}, X\right] + \mathrm{Var}\left(\mathbb{E}\left[\sum_{i \in G} B_i|N_1, ..., N_d\right] \mid \beta_{S_\ell^c}, X\right) \\
&= \mathbb{E}\left[\sum_{i \in G} \mathrm{Var}\left(B_i|N_i\right)|\beta_{S_\ell^c}, X\right] + \mathrm{Var}\left(\sum_{i \in G} \mathbb{E}[B_i|N_i] \mid \beta_{S_\ell^c}, X\right).
\end{aligned}$$

Here it is used that $\{B_i\}_{i \in G}$ are independent given $\{N_i\}_{i \in G}$. Recall that $|X| = \sum_{i \in G} N_i$ is deterministic given X. Then this further simplifies to

$$\begin{aligned}
\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right) &= \mathbb{E}\left[N_t p(1-p) + \sum_{i \in G \setminus \{t\}} N_i q(1-q) \mid \beta_{S_\ell^c}, X\right] + \mathrm{Var}\left(N_t p + \sum_{i \in G \setminus \{t\}} N_i q \mid \beta_{S_\ell^c}, X\right) \\
&= \mathbb{E}[N_t p(1-p) + (|X| - N_t)q(1-q) \mid \beta_{S_\ell^c}, X] + \mathrm{Var}\left(N_t p + (|X| - N_t)q \mid \beta_{S_\ell^c}, X\right) \\
&= \mathbb{E}[N_t|\beta_{S_\ell^c}, X](p(1-p) - q(1-q)) + |X|q(1-q) + (p-q)^2 \mathrm{Var}\left(N_t|\beta_{S_\ell^c}, X\right) \\
&\leq \pi_t np\left(1 - \frac{3}{4}p\right) + q(1-q)\left(|X| + \frac{5}{4}\pi_t n\right) \\
&\leq \pi_t np + q(1-q)\left(|X| + \frac{5}{4}\pi_t n\right)
\end{aligned}$$

Here we used the property that $\mathrm{Var}\,(N_t) \leq \frac{\pi_t n}{4}$ (for the proof of the properties of a hypergeometric random variable, see appendix B). Now by applying lemma 7 we get

$$H(Y_t|\beta_{S_\ell^c}, X) \leq \frac{1}{2}\log\left(2\pi e\left(\pi_t np + q(1-q)\left(|X| + \frac{5}{4}\pi_t n\right) + \frac{1}{12}\right)\right).$$

Now by maximizing over $t \in [d]$ and substituting back into (5.11), we get

$$\begin{aligned} H(Y|\beta_{S_\ell^c}, X) &\leq \sum_{t=1}^{d-1} \frac{1}{2}\log\left(2\pi e\left(\pi_t np + q(1-q)\left(|X| + \frac{5}{4}\pi_t n\right) + \frac{1}{12}\right)\right) \\ &\leq \frac{d-1}{2}\log\left(2\pi e\left(\pi_l np + q(1-q)\left(|X| + \frac{5}{4}\pi_l n\right) + \frac{1}{12}\right)\right). \end{aligned} \tag{5.14}$$

5.1.3 Number of label vectors

Analogously to the binary case, Scarlett and Cevher (2017) have shown that

$$\log|\mathcal{B}_\ell(\pi)| = p_G(H(\pi_G) + o(1)), \tag{5.15}$$

where π_G is the distribution vector whose entries are the scaled equivalent to the positive entries of π, such that $\sum_{t\in G}(\pi_G)_t = 1$. Besides, observing that $p_G = \sum_{t\in G}\pi_t n$, it follows that

$$\begin{aligned} H(\pi_G) &= -\sum_{t\in G}\frac{\pi_t n}{p_G}\log_2\left(\frac{\pi_t n}{p_G}\right) \\ &= -\log\left(\frac{n}{p_G}\right) - \frac{n}{p_G}\sum_{t\in G}\pi_t\log(\pi_t) \\ &\geq -\log\left(\frac{n}{p_G}\right) - \frac{n}{p_G}\sum_{t\in G}\pi_t\log(\pi_m) \\ &= -\log\left(\frac{n}{p_G}\right) - \log(\pi_m) \\ &= -\log\left(\frac{n\pi_m}{p_G}\right). \end{aligned}$$

Here, $\pi_m := \max_{t\in[d]}\pi_t$. Substituting this result back yields

$$\log|\mathcal{B}_\ell(\pi)| \geq p_G\left(-\log\left(\frac{n\pi_m}{p_G}\right) + o(1)\right). \tag{5.16}$$

5.1.4 Lower bound number of tests

The results from sections 5.1.1 and 5.1.2 can be combined to find an upper bound for the mutual information term $I(\beta, Y|\beta_{S_\ell^c}, X)$. Using the characterization of the mutual information, it follows that

$$\begin{aligned} I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\ &\leq \frac{d-1}{2}\log\left(2\pi e\left(\pi_l np + q(1-q)\left(|X| + \frac{5}{4}\pi_l n\right) + \frac{1}{12}\right)\right) \\ &\quad - \frac{d-1}{2}\log\left(2\pi e\pi_l|X|q\right) - \frac{1}{2}\log(p) - \frac{1}{12\pi_l|X|}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right) \\ &= \frac{d-1}{2}\log\left(\frac{np}{|X|q} + (1-q)\left(\frac{1}{\pi_l} + \frac{5n}{|X|}\right) + \frac{1}{12\pi_l|X|q}\right) - \frac{1}{2}\log(p) \\ &\quad + \frac{1}{12\pi_l|X|}\left(\frac{1}{p} + \frac{d-1}{q} - 3d + 2\right) \end{aligned}$$

This term is dependent on the test size. Using $1 \leq |X| \leq n$, this can be surpassed, and therefore

$$\frac{1}{m}\sum_{i=1}^{m} I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)}) \leq \frac{d-1}{2}\log\left(\frac{np}{q} + (1-q)\left(\frac{1}{\pi_l} + 5n\right) + \frac{1}{12\pi_l q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l}\left(\frac{1}{p} + \frac{d-1}{q} - \frac{3d-2}{n}\right). \tag{5.17}$$

The lower bound in theorem 1 is expressed as a maximization over the vector ℓ. With the simplification as described in section 3.6, this amounts to maximizing over the choice of G. As the only term that is dependent on this choice is $\log|\mathcal{B}_\ell(\pi)|$, we are tasked with finding the set G such that this term is maximal. This occurs when p_G takes the largest value possible, i.e. when $G = \{1, ..., d\}$ and $p_G = n$ (see Appendix C for the full proof). Consequently, for any $\delta \in (0, 1)$, to satisfy the condition that $P_e \leq \delta$, the number of tests m needs to be bounded by

$$m \geq \frac{n\left(-\log(\pi_l) + o(1)\right)(1-\delta) - \log 2}{\frac{d-1}{2}\log\left(\frac{np}{q} + (1-q)\left(\frac{1}{\pi_l} + 5n\right) + \frac{1}{12\pi_l q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l}\left(\frac{1}{p} + \frac{d-1}{q} - \frac{3d-2}{n}\right)}. \tag{5.18}$$

5.2 Simplification 2

The second simplification in the generalised case has a lot in common with the binary case. By letting $\ell = (n\pi_1, 1, 0, ..., 0)$, we will also consider only two types of items. Nevertheless, as one of these items can still be identified as a type $t \geq 3$, there are some small notable differences. The final entropy remains the same with respect to the first simplification, yet the progressed entropy changes. First, recall that any of the entries of Y is deterministic given all of the others. Therefore, using the result in (5.11), but swapping Y_1 and Y_d, and Lemma 6, it holds that

$$H(Y|\beta_{S_\ell^c}, X) \leq \sum_{t=2}^{d} H(Y_t|\beta_{S_\ell^c}, X) \tag{5.19}$$

$$\leq \sum_{t=2}^{d} \frac{1}{2}\log\left(2\pi e\left(\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right) + \frac{1}{12}\right)\right). \tag{5.20}$$

We will compute $\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right)$ for an arbitrary choice of $t = 2, ..., d$. First, recall that $Y_t = \sum_{i\in[d]} B_i^{(t)}$, where these random variables are given in (5.13). Given the simplification of ℓ, we may assume that $B_t^{(i)} \equiv 0$ for $t = 3, ..., d$, as the items of these types have already been identified. This matches the binary case, although now the binomial variables have a different probability parameter. Therefore, it can quickly be seen that

$$\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right) \leq |X|q(5-q)$$

for $t = 2, ..., d$. Since this bound is independent of t, it now follows that

$$H(Y|\beta_{S_\ell^c}, X) \leq \frac{d-1}{2}\log\left(2\pi e\left(|X|q(5-q) + \frac{1}{12}\right)\right). \tag{5.21}$$

The final entropy remains the same with respect to the first simplification, therefore the mutual information terms can be bounded by

$$\begin{aligned} I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\ &\leq \frac{d-1}{2}\log\left(2\pi e\left(|X|q(5-q) + \frac{1}{12}\right)\right) \\ &\quad - \frac{d-1}{2}\log\left(2\pi e\pi_l|X|q\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l|X|}\left(\frac{1}{p} + \frac{d-1}{q} - 3d + 2\right) \\ &= \frac{d-1}{2}\log\left(\frac{5-q}{\pi_l} + \frac{1}{12\pi_l|X|q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l|X|}\left(\frac{1}{p} + \frac{d-1}{q} - 3d + 2\right). \end{aligned}$$

This expression is dependent on the test size. In the deterministic setting, there are no other bounds for $|X|$ than $1 \leq |X| \leq n$. Applying this yields

$$\frac{1}{m}\sum_{i=1}^{m} I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)}) \leq \frac{d-1}{2}\log\left(\frac{5-q}{\pi_l} + \frac{1}{12\pi_l q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l}\left(\frac{1}{p} + \frac{d-1}{q} - \frac{3d-2}{n}\right). \tag{5.22}$$

Having already identified the items of all types other than one and two, the number of possible label vectors is similarly as given in the binary case in (4.21). Therefore, in the deterministic setting, the number of tests required for $P_e \leq \delta$ for any $\delta \in (0,1)$ is bounded from below by

$$m \geq \frac{\log(n\pi_1 + 1)(1-\delta) - \log 2}{\frac{d-1}{2}\log\left(\frac{5-q}{\pi_l} + \frac{1}{12\pi_l q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\pi_l}\left(\frac{1}{p} + \frac{d-1}{q} - \frac{3d-2}{n}\right)} \tag{5.23}$$

as $n \to \infty$.

5.3 Bernoulli setting

In the Bernoulli setting, the tests, and in particular the test sizes, are not deterministic anymore. Instead, the latter follows a binomial distribution $\mathrm{Bin}(\alpha, n)$. Therefore, there will be some changes to the progressed and final entropies with respect to the deterministic setting.

Since the final entropy is similar for both simplifications of ℓ, we will first investigate the changes here. Using the definition of conditional entropy and the results from the deterministic setting, it now follows that

$$\begin{aligned} H(Y|\beta, X) &= \sum_{x\in\mathcal{X}} f_X(x) H(Y|\beta, X = x) \\ &\geq \frac{d-1}{2}\left(\log\left(2\pi e \pi_l q\right) + \mathbb{E}[\log|X|]\right) + \frac{1}{2}\log(p) + \mathbb{E}\left[\tfrac{1}{|X|}\right]\cdot\frac{1}{12\pi_l}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right) \\ &= \frac{d-1}{2}\log\left(2\pi e\alpha n q\right) + \frac{1}{2}\log(p) + \frac{1}{12\alpha n\pi_l}\left(3d - 2 - \frac{1}{p} - \frac{d-1}{q}\right) + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{5.24}$$

It remains to compute the progressed entropies for both cases of ℓ separately. The number of potential label vectors a decoder can yield given ℓ remains the same.

5.3.1 Simplification 1

Similarly, the progressed entropy can now be found using the results from the deterministic setting and the definition of conditional entropy. This results in

$$H(Y|\beta_{S_\ell^c}, X) = \sum_{x\in\mathcal{X}} f_X(x) H(Y|\beta_{S_\ell^c}, X = x) \tag{5.25}$$

$$\leq \sum_{x\in\mathcal{X}} f_X(x)\frac{d-1}{2}\log\left(2\pi e\left(\pi_l np + q(1-q)\left(|x| + \frac{5}{4}\pi_l n\right) + \frac{1}{12}\right)\right) \tag{5.26}$$

$$\leq \frac{d-1}{2}\log\left(2\pi e\alpha n\left(\pi_l np + q(1-q)\left(1 + \frac{5}{4}\pi_l n\right) + \frac{1}{12}\right)\right) + O\left(\frac{1}{n^2}\right). \tag{5.27}$$

The final entropy can be directly taken from (5.24). The mutual information terms can therefore be bounded by

$$\begin{aligned}
I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\
&\leq \frac{d-1}{2}\log\left(2\pi e\alpha n\left(\pi_l np + q(1-q)\left(1+\frac{5}{4}\pi_l n\right)+\frac{1}{12}\right)\right) \\
&\quad - \frac{d-1}{2}\log(2\pi e\alpha nq) - \frac{1}{2}\log(p) + \frac{1}{12\alpha n\pi_l}\left(\frac{1}{p}+\frac{d-1}{q}-3d+2\right) \\
&= \frac{d-1}{2}\log\left(\frac{\pi_l np}{q}+(1-q)\left(1+\frac{5}{4}\pi_l n\right)+\frac{1}{12q}\right) - \frac{1}{2}\log(p) \\
&\quad + \frac{1}{12\alpha n\pi_l}\left(\frac{q}{p}+\frac{d-1}{q}-3d+2\right).
\end{aligned}$$

Once again, these are independent on the tests, and therefore averaging over the tests is not required. It can be concluded that

$$m \geq \frac{n(H(\pi)+o(1))(1-\delta)-\log 2}{\frac{d-1}{2}\log\left(\frac{\pi_l np}{q}+1+\frac{5}{4}\pi_l n+\frac{1}{12q}\right)-\frac{1}{2}\log(p)+\frac{1}{12\alpha n\pi_l}\left(\frac{q}{p}+\frac{d-1}{q}-3d+2\right)}. \tag{5.28}$$

to obtain $P_e \leq \delta$ for any $\delta \in (0,1)$ as $n \to \infty$.

5.3.2 Simplification 2

The progressed entropy with the second simplification of ℓ is given in (5.21). By following the same computations as before,

$$\begin{aligned}
H(Y|\beta_{S_\ell^c}, X) &= \sum_{x\in\mathcal{X}} f_X(x) H(Y|\beta_{S_\ell^c}, X=x) && (5.29)\\
&\leq \sum_{x\in\mathcal{X}} f_X(x)\frac{d-1}{2}\log\left(2\pi e\left(|x|q(5-q)+\frac{1}{12}\right)\right) && (5.30)\\
&\leq \mathbb{E}[\log|X|]\ \frac{d-1}{2}\log\left(2\pi e\left(q(5-q)+\frac{1}{12}\right)\right) && (5.31)\\
&= \frac{d-1}{2}\log\left(2\pi e\alpha n\left(q(5-q)+\frac{1}{12}\right)\right)+O\left(\frac{1}{n^2}\right). && (5.32)
\end{aligned}$$

The final entropy in this setting as given in (5.24) can be directly taken to compute the mutual information. Now it follows that

$$\begin{aligned}
I(\beta, Y|\beta_{S_\ell^c}, X) &= H(Y|\beta_{S_\ell^c}, X) - H(Y|\beta, X) \\
&\leq \frac{d-1}{2}\log\left(2\pi e\alpha n\left(q(5-q)+\frac{1}{12}\right)\right) - \frac{d-1}{2}\log(2\pi e\alpha nq) - \frac{1}{2}\log(p) \\
&\quad + \frac{1}{12\alpha n\pi_l}\left(3d-2-\frac{1}{p}-\frac{d-1}{p}\right) \\
&= \frac{d-1}{2}\log\left(\frac{5-q}{\pi_m}+\frac{1}{12q}\right) - \frac{1}{2}\log(p) + \frac{1}{12\alpha n\pi_m}\left(\frac{1}{p}+\frac{d-1}{q}-3d+2\right).
\end{aligned}$$

These are independent of the test and therefore averaging over these mutual information terms with respect to the tests is not required. The number of tests can therefore be bounded by

$$m \geq \frac{\log(n\pi_1+1)(1-\delta)-\log 2}{\frac{d-1}{2}\log\left(\frac{5-q}{\pi_m}+\frac{1}{12q}\right)-\frac{1}{2}\log(p)+\frac{1}{12\alpha n\pi_m}\left(\frac{1}{p}+\frac{d-1}{q}-3d+2\right)}, \tag{5.33}$$

under the usual requirements.

5.4 Results

Table 2 indicates the orders of the bounds on m for several choices of p. The orders differ under the deterministic and Bernoulli setting. When p is taken as a constant, the bounds are (sub)-linear and different for the two simplifications of ℓ.

Setting	Equation	$p = \Theta(1)$	$p = 1 - \frac{1}{n}$	$p = 1 - \frac{1}{\sqrt{n}}$
Deterministic	(5.18)	$\Omega(n)$	$\Omega(\frac{n}{\log n})$	$\Omega(\frac{n}{\log n})$
	(5.23)	$\Omega(\log n)$	$\Omega(\frac{\log n}{n})$	$\Omega(\frac{\log n}{\sqrt{n}})$
Bernoulli	(5.28)	$\Omega(\frac{n}{\log n})$	$\Omega(\frac{n}{\log n})$	$\Omega(\frac{n}{\log n})$
	(5.33)	$\Omega(\log n)$	$\Omega(1)$	$\Omega(1)$

Table 2: Necessary conditions on the number of tests m for several choices of p in the general case

6 Discussion

6.1 Lower bounds on the number of tests

Table 1 indicates the order of the lower bounds on m for the two different simplifications of ℓ. Notably, the bounds are independent of the deterministic or Bernoulli setting. The bound on m in Theorem 1 is an extreme minimum, i.e. even when the decoder works optimally, this bound is valid. Under the Bernoulli setting, the tests may not be chosen optimally as they are randomly generated, in contrast to the deterministic regime. Therefore, it is expected that the orders of m are smaller, or at least not as large, in the deterministic case than the Bernoulli setting. The results in the binary case indicate the latter situation. The general case, however, exhibit larger differences between the two settings, though only per simplification of ℓ. For example, in the first choice of ℓ, there are no differences, whereas for the second choice, the lower bounds on m are smaller in the Bernoulli setting.

El Alaoui et al. (2019) have shown that in the noiseless case in the Bernoulli setting, an optimal decoder yields a recovery with order at most $O(\frac{n}{\log n})$. Scarlett and Cevher (2017) on the other hand, have shown that the lower bound is of the same order in this scenario. When p converges to one, the noise decreases and eventually a noiseless scenario is approximated. Therefore, it is expected that the order will not vary much from the result, or is at least at most $\Omega(\frac{n}{\log n})$. The results, for both the binary and general cases, confirm these expectations.
Besides, when comparing the bounds found for a particular setting and and choice of ℓ, letting p converge to one would yield a smaller lower bound on m, as the noise reduces and the reliability of the test results increases. Particularly, the faster p converges, the lower order of m is expected. The binary case coincides with the expectations, whereas for the general case this is mostly valid. The orders indeed decrease when p converges to one, however, a faster convergence does not always yield a smaller order of m, e.g. see the result in (5.23).
The binary case, on the contrary, indicates that the order of m is independent on the rate of converges for p. This difference can be justified by assessing the entropy of binomial and multinomial random variables in (4.9) and (5.6) respectively. These are both used to directly compute the final entropies. Note that the entropy of a binomial variable has an error term of $O\left(\frac{1}{n}\right)$, whereas the entropy of the multinomial variable has the error term $O\left(\frac{1}{n^2}\right)$. Therefore, the latter is a more exact expression. Specifically, when the order of the error term would be reduced, the entropy of a multinomial variable would then be given by

$$H\left(\mathrm{Multin}\left(n, \boldsymbol{p}^{(t)}\right)\right) = \frac{1}{2}\log\left((2\pi neq)^{d-1}p\right) + O\left(\frac{1}{n}\right).$$

Here, $\boldsymbol{p}^{(t)}$ for any $t = 1, ..., d$ is defined as in (3.4). This exactly matches the entropy of a binomial variable in (4.9). Now the term linear in $\frac{1}{n}$ has vanished, which originally caused the difference between the two final entropies.

The lower bounds depend on the distribution π. In the first choice of ℓ, either the items of a specific type are all revealed or not before the decoder starts. As the choice of these labels is not specified -though it is limited in the binary case,- the bounds in the general case are symmetric in terms of the distribution. More precisely, here they only depend on the maximum proportion π_l. The lower bounds obtain a maximum value when π_l is at its minimum, i.e. if π is uniform, and vice versa. When π_l takes a value close to one, any queried pool will mostly contain items of that label. As a result, it requires little effort to identify the items of all other types.
The second choice of ℓ is more specific. In this case, all items of type one and a single item of type two remain to be identified. However, a same strategy can be applied by assuming that the items of a type other than one are yet to be revealed. By that reasoning, the bounds in this choice of ℓ can be maximized over π. Then, the bounds would also be symmetric in terms of the distribution π. Besides, the bounds take a maximum value when this is again a uniform distribution.
In the general case, the lower bounds also depend on the number of labels d. Particularly, the bounds decrease when d takes a larger value.

These results coincide with the outcome of studies about the noiseless model. Particularly, El Alaoui et al. (2019) have shown that an upper bound on m is also maximum when π is uniform. Besides, they argue that a larger number of labels also results in fewer tests. The results in the study by Scarlett and Cevher (2017) prove the same for the lower bounds in the noiseless setting.

6.2 Gaps in the computations

The found lower bound of m may not actually be a sufficient number of tests to recover the labels of each item. There is a gap that arises from several simplifications and inequalities that have been used throughout the computations, especially of the mutual terms $I(\beta, Y^{(i)}|\beta_{S_\ell^c}, X^{(i)})$.

Firstly, Lemma 6 by Massey (1988) gives an upper bound for an integer valued random variable based on its variance. More precisely, there is a strict inequality. This lemma has been widely investigated and improved upon, although under some additional conditions. For example, Rioul (2022) has proven that a non-negative integer-valued random variable X with finite mean μ and variance σ^2, such that $\frac{\mu}{\sigma^2}$ is bounded from above by a factor $\gamma < 2\pi$, then for large enough σ^2,

$$H(X) \leq \frac{1}{2}\log(2\pi e\sigma^2).$$

In particular, for a random variable $X \sim \mathrm{Bin}(n, p)$ the conditions are satisfied. The main difference between this result and lemma 6 is the term $\frac{1}{12}$ that now has disappeared. This new variant of Massey's lemma is not applied as it is still relatively new and not thoroughly checked. Nevertheless, it illustrates that the results regarding to the progressed entropy can be improved upon.

Another computation where a gap could have come from is that in (5.19). Here, the term $H(Y|\beta_{S_\ell^c}, X)$ is bounded from above by computing the conditional entropies $H(Y_t|\beta_{S_\ell^c}, X)$ for $t = 2, ..., d$. The term $H(Y_1|\beta_{S_\ell^c}, X)$ is left out as Y_1 is deterministic given all other components of Y. Since this generally holds for any other component as well, this bound could have been optimized by rather computing

$$H(Y|\beta_{S_\ell^c}, X) \leq \min_{t\in[d]} \sum_{i\in[d]\setminus\{t\}} H(Y_i|\beta_{S_\ell^c}, X). \tag{6.1}$$

In this setting, only the elements of type one and a single item of type two are yet to be identified. To compute the expression on the right hand side, their variances are determined. Therefore this bound is optimal when $t \in [d]$ is chosen such that $\mathrm{Var}\left(Y_t|\beta_{S_\ell^c}, X\right) \geq \mathrm{Var}\left(Y_i|\beta_{S_\ell^c}, X\right)$ for all $i = 1, ..., d$. Since most items that need to be identified are of type one, it is reasonable that $t = 1$ in (6.1).

In several occasions, we are required to compute an expectation of the form $\mathbb{E}[\log(a+b|X|)]$, where $|X|$ is a binomial random variable and $a, b > 0$ are some positive constants, for example in (5.26). However, there are no computational rules for a logarithm that allows to split up this expression. Flajolet (1999) has given an asymptotic expression for $\mathbb{E}[\log|X|]$, but this is without the constants. One possible way to tackle this problem is by rewriting this to $\log(a)\mathbb{E}[\log(1 + \frac{b}{a}|X|)]$ noting that such a logarithmic expression can be expanded by

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + ... = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}x^k$$

for any $x > -1$. However, then we are required to give a closed expression for the k-th moment of $|X|$, and also compute this sum. As this is a heavy task, another strategy needed to be implemented. The expectations are computed by using the inequality $\log(a + b|X|) \leq \log(|X|(a + b))$, and then by applying the result from Flajolet (1999). This is done in several occasions, such as (5.26) and (5.30), and could drastically change the outcome. We will illustrate these consequences using the example in equation (5.26), where an attempt is made to find a good upper bound on $H(Y|\beta_{S_\ell^c}, X)$ in the general case for the first choice of ℓ. Note that the expression in (5.26) is of the form

$$\mathbb{E}[\alpha\log(\beta_1 n + \beta_2|X| + \beta_3)]$$

for some constants $\alpha, \beta_1, ..., \beta_3 > 0$, all independent of n. Since $|X|$ is a binomial random variable, this is expected to be linear in terms of n, which would result in an expression of the form

$$\alpha \log(\beta_1 n + \beta_2' n + \beta_3)$$

for some new constant $\beta_2' > 0$. However, by applying the strategy as described before, this now becomes

$$\alpha \log(\gamma n(\beta_1 n + \beta_2 + \beta_3))$$

for some $\gamma > 0$. This contains a polynomial of order two instead of one, in contrast to what was expected. The bound on m given in (5.28), which comes forth from these computations, is now also different from what was expected. More precisely, when $|X|$ is indeed linear in terms of n inside the logarithm, then this bound would be of order $\Omega(n)$ instead of $\Omega(\frac{n}{\log n})$ for a constant p.

7 Conclusion

In this research, we have investigated lower bounds in the noisy pooled data problem where elements are correctly identified with a certain probability. This was achieved using the framework by Scarlett and Cevher (2017). Within this framework, we characterized the distribution of the pool test results.
We were interested in the order of these bounds and how they depend on the probability of correct identification. This was first done in the binary case, where only two labels were considered, and then generalized to an arbitrary number of labels. Both cases were analyses in the deterministic setting, i.e. when the queried pools were pre-determined, and the Bernoulli setting, when each item is pooled with a certain probability, independent of all other items.

The results revealed that in the deterministic setting, the number of required tests is at least linear in terms of the population's size when the probability of correct identification is constant. This is much higher than in the noiseless case. Moreover, when this probability converges to one, a noiseless setting is approached and the order of the number of tests is identical to that of the noiseless case. In a Bernoulli setting, the lower bounds on the number of tests are identical in the binary case, but of a lower order in the general case. There against, in both cases the order are the same as in the noiseless setting as the probability of correct identification approaches one.
Moreover, a larger number of labels resulted in a decrease in the required number of tests. Besides, the bounds obtain a maximum when the proportions of the item labels are uniform, and oppositely a minimum when most items are of the same type. These result coincide with the outcome of several studies about the noiseless setting.

Further research is required to obtain a more exact lower bound. This specifically concerns determining the expectation of $\log(a + Z)$ for a binomial distributed random variable Z. Moreover, an exact expression for the entropy of a binomial random variable can be used to find a more accurate bound. Whereas the current research has focused on lower bounds, future research could also investigate upper bounds.

References

Cichon, J., & Golebiewski, Z. (2012, December). On bernoulli sums and bernstein polynomials. *Discrete Mathematics & theoretical computer science DMTCS*. doi: 10.46298/dmtcs.2993

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons, Inc.

de Moivre, A. (2013). *The doctrine of chances*. Cambridge University Press. doi: https://doi-org.dianus.libr.tue.nl/10.1017/CBO9781139833783

Douglas, J. A., Skol, A. D., & Boehnke, M. (2002, February). Probability of detection of genotyping errors and mutations as ineheritance of inconsistencies in nuclear-family data. *The American Journal of Human Genetics*, *70*. doi: https://doi.org/10.1086/338919

El Alaoui, A., Ramdas, A., Krzakala, F., Zdeborova, L., & Jordan, M. (2016). Decoding from pooled data: Sharp information-theoretic bounds. *SIAM Journal on Mathematics of Data Science*, *1*(1). doi: https://doi.org/10.1137/18M1183339

El Alaoui, A., Ramdas, A., Krzakala, F., Zdeborova, L., & Jordan, M. (2019). Decoding from pooled data: Phase transitions of message passing. *IEEE Transactions on Information Theory*, *65*(1). doi: https://doi.org/10.1109/TIT.2018.2855698

Flajolet, P. (1999, February 28). Singularity analysis and asymptotics of bernoulli sums. *Theoretical Computer Science*, *215*, 371–381. doi: https://doi.org/10.1016/S0304-3975(98)00220-5

Hahn-Klimroth, M., & Müller, N. (2021, August). Near optimal efficient decoding from pooled data. *arXiv.org, e-Print Archive, Mathematics*. doi: https://arxiv.org/pdf/2108.04342.pd

Massey, J. L. (1988, July 4-7). On the entropy of integer-valued random variables. *Proceedings Beijing International Workshop on Information Theory*.

Paulsen, W. (2014). *Asymptotic analysis and perturbation theory* (1st ed.). Taylor & Francis Group.

Rioul, O. (2022, May). Variations on a theme by massey. *IEEE Transactions on Information Theory*, *68*. doi: https://doi.org/10.48550/arXiv.2102.04200

Scarlett, J., & Cevher, V. (2017). Phase transitions in the pooled data problem. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., & Owen, M. (2002, November). Dna pooling: a tool for large-scale association studies. *Nature Review Genetics*, *3*. doi: DOI:10.1038/nrg930

van Berkum, E., & Bucchianico, A. D. (2016). *Statistical compendium*. Eindhoven University of Technology.

Wang, C., Zhao, Q., & Chuah, C. (2018, February). Optimal nested test plan for combinatorial quantitative group testing. *IEEE Transactions on Signal Processing*, *66*. doi: https://doi.org/10.1109/TSP.2017.2780053

A Proof of lemmas

Lemma 2. (Entropy of sum of independent random variables). *Let $X_1, ..., X_n$ be $n \in \mathbb{N}$ independent discrete random variables. Then*

$$H\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} H(X_i) - H\left(X_i | \sum_{j=1}^{i} X_j\right).$$

Proof. We proceed by induction. We will first show it for the base case, i.e. $n = 1$. Recall that the entropy of a deterministic random variable is equal to 0, and therefore

$$\sum_{i=1}^{1} H(X_i) - H(X_i | \sum_{j=1}^{i} X_j) = H(X_1) - H(X_1|X_1) = H(X_1) = H\left(\sum_{i=1}^{1} X_i\right).$$

Now assume that the lemma holds for some discrete random variables $X_1, ..., X_k$, where $k \in \mathbb{N}$, i.e. indeed

$$H\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} H(X_i) - H\left(X_i | \sum_{j=1}^{i} X_j\right).$$

Consider the random variable X_{k+1} and define $Y = \sum_{i=1}^{k} X_i$. We first find that

$$H(Y + X_{k+1}, X_{k+1}) = H(Y|X_{k+1}) + H(X_{k+1}) = H(Y) + H(X_{k+1})$$

since $Y = X_1 + ... + X_k$ and X_{k+1} are assumed to be independent (here, lemma 3 is applied). For the proof that the entropy of these random variables together can be written as such two separate entropies, see the proof in lemma 5.
Similarly, since the entropy is symmetric, we find by taking Y as the conditional random variable that also

$$H(Y, X_{k+1}) = H(X_{k+1}, Y) = H(X_{k+1}|Y + X_{k+1}) + H(Y + X_{k+1}).$$

Now $Y + X_{k+1} = X_1 + ... + X_{k+1}$. Combining these results indeed yields

$$H(Y+X_{k+1}) = H(Y)+H(X_{k+1})-H(X_{k+1}|Y+X_{k+1}) = \sum_{i=1}^{2} H(X_i)-H\left(X_i | \sum_{j=1}^{i} X_j\right).$$

So the lemma also holds for $k + 1$, which proves the induction step. □

Lemma 3. (Conditioning reduces entropy (Cover & Thomas, 2006)). *Let X and Y be discrete random variables. Then*

$$H(X|Y) \leq H(X) \tag{A.1}$$

with equality if and only if X and Y are independent.

Proof. Let X and Y be discrete random variables that take values in $\mathcal{X}$ and $\mathcal{Y}$ and have probability mass functions f_X and f_Y respectively. Then by the law of total probability, the mutual information between these terms can be rewritten to

$$\begin{aligned}
I(X|Y) &= H(X) - H(X|Y) \\
&= -\sum_{x\in\mathcal{X}} f_X(x)\log(f_X(x)) + \sum_{y\in\mathcal{Y}} f_Y(y) \sum_{x\in\mathcal{X}} f_{X|Y}(x|y)\log(f_{X|Y}(x,y)) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_{X|Y}(x|y)f_Y(y)\log(f_X(x)) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_{X|Y}(x|y)f_Y(y)\log(f_{X|Y}(x|y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_{X|Y}(x|y)f_Y(y)\log\left(\frac{f_{X|Y}(x|y)}{f_X(x)}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right).
\end{aligned}$$

Now we can apply the inequality

$$\log(x) \geq 1 - \frac{1}{x}, \quad x > 0$$

with equality if and only if $x = 1$. This yields

$$\begin{aligned} I(X|Y) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f_{X,Y}(x,y) \log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) \\ &\geq \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f_{X,Y}(x,y) \left(1 - \frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)}\right) \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f_{X,Y}(x,y) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f_X(x)f_Y(y) \\ &= 1 - 1 \\ &= 0. \end{aligned}$$

Therefore rewriting the definition of the mutual entropy gives

$$H(Y|X) = H(X) - I(X|Y) \leq H(X).$$

To prove equality if and only if X and Y are dependent, note that only under this condition,

$$\frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)} = \frac{f_X(x)f_Y(y)}{f_X(x)f_Y(y)} = 1.$$

This yields the equality regarding the logarithm. □

Lemma 4. (Entropy of a binomial random variable). *Let $Z \sim \mathrm{Binomial}(n,p)$ be a binomial random variable. Then its entropy is given by*

$$H(Z) = \frac{1}{2}\log_2\left(2\pi enp(1-p)\right) + O\left(\frac{1}{n}\right).$$

Proof. The proof relies on the De-Moivre Laplace theorem. The probability mass function of Z can be given by

$$f_Z(x) = \binom{n}{x} p^n (1-p)^{n-x}$$

for all $x \in \mathcal{Z}$ (e.g. see van Berkum and Bucchianico (2016)). The De Moivre-Laplace theorem can be applied to approximate this function.

Theorem 8 (De Moivre-Laplace (2013))**.** *As n grows large, for k in the $\sqrt{npq}$ neighbourhood of np, i.e. $|k - np| < \sqrt{npq}$, we can approximate*

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}} + O\left(\frac{1}{n}\right).$$

For shorter notation, we define $\sigma^2 = np(1-p)$ and $\mu = np$, then it follows that

$$\begin{aligned}
H(B) &= -\sum_x \binom{n}{x} p^x (1-p)^x \log_2\left(\binom{n}{x} p^x (1-p)^x\right) \\
&= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log_2\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx + O\left(\frac{1}{n}\right) \\
&= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\frac{1}{2}\log_2(2\pi\sigma^2) + \log_2\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)\right] dx + O\left(\frac{1}{n}\right) \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{\log(2)} dx + O\left(\frac{1}{n}\right) \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{1}{2\sigma^2\log(2)} \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma^2}} dx + O\left(\frac{1}{n}\right) \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{1}{2\sigma^2}\log_2(e)\sigma^2 + O\left(\frac{1}{n}\right) \\
&= \frac{1}{2}\log_2(2\pi e\sigma^2) + O\left(\frac{1}{n}\right) \\
&= \frac{1}{2}\log_2(2\pi enp(1-p)) + O\left(\frac{1}{n}\right).
\end{aligned}$$

Here, the distribution of a normal random variable with mean μ and variance σ^2 can be recognized. In particular, in the last integral, the definition of its variance can be found. As a result, we find

$$H(B) = \frac{1}{2}\log_2(2\pi enp(1-p)) + O\left(\frac{1}{n}\right).$$

□

Lemma 5. (Chain rule for entropy, Cover and Thomas (2006)). *Let $(X_1, ..., X_n)$ be a random vector. Then its entropy is equal to*

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, ..., X_1).$$

Proof. We use induction to prove the lemma. First consider the random vector (X_1, X_2) with probability mass function f_{X_1,X_2}. Now by the definition of conditional probability,

$$f_{X_1,X_2}(x_1, x_2) = f_{X_2|X_1}(x_2|x_1) \cdot f_{X_1}(x_1)$$

so that

$$\begin{aligned}
H(X_1, X_2) &= \sum_{x_1,x_2} f_{X_1,X_2}(x_1, x_2) \log_2(f_{X_1,X_2}(x_1, x_2)) \\
&= \sum_{x_1,x_2} f_{X_2|X_1}(x_2|x_1) \cdot f_{X_2}(x_2) \log_2(f_{X_2|X_1}(x_2|x_1) \cdot f_{X_1}(x_1)) \\
&= \sum_{x_1,x_2} f_{X_2|X_1}(x_2|x_1) \cdot f_{X_1}(x_1) \left[\log_2(f_{X_2|X_1}(x_2|x_1)) + \log_2(f_{X_1}(x_1))\right] \\
&= \sum_{x_1} f_{X_1}(x_1) \sum_{x_2} f_{X_2|X_1}(x_2|x_1) \left[\log_2(f_{X_2|X_1}(x_2|x_1)) + \log_2(f_{X_1}(x_1))\right] \\
&= \sum_{x_1} f_{X_1}(x_1) \left[H(X_2|X_1) + \log_2(f_{X_1}(x_1))\right] \\
&= H(X_2|X_1) + \sum_{x_2} f_{X_1}(x_1) \log_2(f(x_1)) \\
&= H(X_2|X_1) + H(X_1).
\end{aligned}$$

Now assuming that it holds for some $k \in \mathbb{N}$. Then the induction step follows from taking $Y = (X_1, ..., X_k)$ as a single random variable and applying the same steps as before to find

$$\begin{aligned} H(X_1, ..., X_{k+1}) &= H(Y, X_{k+1}) = H(X_{k+1})|Y) + H(Y) \\ &= H(X_{k+1}|X_1, ..., X_k) + \sum_{i=1}^{k} H(X_i|X_1, ..., X_{i-1}) \\ &= \sum_{i=1}^{k+1} H(X_i|X_1, ..., X_{i-1}). \end{aligned}$$

This concludes the proof of the lemma. □

Lemma 7. (Law of total variance) *Let X and Y be random variables on the same probability space such that the variance of Y is finite. Then*

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X]).$$

Proof. First by the definition of variance, we write

$$\mathrm{Var}\left(Y\right) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2.$$

We then apply the law of expectation to both expectations in this equation and again use the definition of variance to find

$$\begin{aligned} \mathrm{Var}\left(Y\right) &= \mathbb{E}[\mathbb{E}[Y^2|X]] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\ &= \mathbb{E}[\mathrm{Var}\left(Y|X\right) + \mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\ &= \mathbb{E}[\mathrm{Var}\left(Y|X\right)] + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\ &= \mathbb{E}[\mathrm{Var}\left(Y|X\right)] + \mathrm{Var}\left(\mathbb{E}[Y|X]\right). \end{aligned}$$

□

B Properties Hypergeometric Random Variable

Consider a hypergeometric random variable $H \sim \mathrm{Hyp}(k, m, n)$. Then the expected value and variance of this random variable are respectively

$$\mathbb{E}[H] = \frac{km}{n}, \quad \mathrm{Var}\,[H] = k \cdot \frac{m}{n} \cdot \frac{n-m}{n} \cdot \frac{n-k}{n-1}.$$

Proof. For the proof, we first define H' to be the hypergeometric random variable with parameters $k-1$, $m-1$ and $n-1$, i.e. $H' \sim \mathrm{Hyp}(k-1, m-1, n-1)$. Moreover, for any non-negative integers $n, k \in \mathbb{N}$ such that $1 \leq k \leq n$, we note that a binomial coefficient can be re-written as

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n}{k} \cdot \frac{(n-1)!}{(k-1)! \cdot ((n-1)-(k-1))!} = \frac{n}{k}\binom{n-1}{k-1}.$$

The probability mass function of such a hypergeometric random variables is defined as

$$\mathbb{P}(H = x) = \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}$$

for all $x \in \{0, ..., \min(m, n)\}$, e.g. as stated by van Berkum and Bucchianico (2016). Then we see that

$$\begin{aligned}
\mathbb{E}[H] &= \sum_{x \in \mathcal{H}} x \cdot \mathbb{P}(H = x) \\
&= \sum_{x=0}^{n} x \cdot \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}} \\
&= \sum_{x=1}^{n} x \cdot \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}} \\
&= \sum_{x=1}^{n} x \cdot \frac{\frac{m}{x} \cdot \binom{m-1}{x-1}\binom{n-m}{k-x}}{\frac{n}{k} \cdot \binom{n-1}{k-1}} \\
&= \frac{km}{n} \sum_{x=1}^{n} \frac{\binom{m-1}{x-1}\binom{(n-1)-(m-1)}{(k-1)-(x-1)}}{\binom{n-1}{k-1}} \\
&= \frac{km}{n} \sum_{x=0}^{n-1} \frac{\binom{m-1}{x}\binom{(n-1)-(m-1)}{(k-1)-x}}{\binom{n-1}{k-1}} \\
&= \frac{km}{n} \sum_{x \in \mathcal{H}'} \mathbb{P}(H' = x) \\
&= \frac{km}{n}.
\end{aligned}$$

For the variance, the same principle can be applied. Using the expected value, we first find

$$\begin{aligned}
\mathbb{E}[H^2] &= \sum_{x\in\mathcal{H}} x^2 \cdot \mathbb{P}(H = x) \\
&= \sum_{x=1}^{n} \frac{km}{n} \cdot x \cdot \frac{\binom{m-1}{x-1}\binom{(n-1)-(m-1)}{(k-1)-(x-1)}}{\binom{n-1}{k-1}} \\
&= \frac{km}{n} \sum_{x=0}^{n-1} x \cdot \frac{\binom{m-1}{x}\binom{(n-1)-(m-1)}{(k-1)-x}}{\binom{n-1}{k-1}} \\
&= \frac{km}{n} \sum_{x=0}^{n-1} (x+1) \cdot \mathbb{P}(H' = x) \\
&= \frac{km}{n} \left(\mathbb{E}[H'] + 1\right) \\
&= \frac{km}{n} \left(\frac{(k-1)(m-1)}{n-1} + 1\right).
\end{aligned}$$

I, Now using the definition of the variance, this gives

$$\begin{aligned}
\operatorname{Var}[H] &= \mathbb{E}[H^2] - \mathbb{E}[H]^2 \\
&= \frac{km}{n} \left(\frac{(k-1)(m-1)}{n-1} + 1\right) - \left(\frac{km}{n}\right)^2 \\
&= k \cdot \frac{m}{n} \left(\frac{km-k-m+1}{n-1} + 1 - \frac{km}{n}\right) \\
&= k \cdot \frac{m}{n} \left(\frac{km-k-m+n}{n-1} - \frac{km}{n}\right) \\
&= k \cdot \frac{m}{n} \cdot \frac{kmn - kn - mn + n^2 - kmn + km}{n(n-1)} \\
&= k \cdot \frac{m}{n} \cdot \frac{km - kn - mn + n^2}{n(n-1)} \\
&= k \cdot \frac{m}{n} \cdot \frac{n-m}{n} \cdot \frac{n-k}{n-1}.
\end{aligned}$$

□

C Maximization of number of label vectors

The objective is to find the choice of G such that

$$-p_G \log_2\left(\frac{n\pi_m}{p_G}\right)$$

is maximal. As the only term depending on G is p_G, this amounts to finding the value of p_G such that this expression is maximal. By considering this expression as a function of p_G and setting its first derivative equal to 0, the extreme values can be found. Now

$$\begin{aligned}\frac{\partial}{\partial p_G} - p_G \log_2\left(\frac{n\pi_m}{p_G}\right) &= -\log_2\left(\frac{n\pi_m}{p_G}\right) - p_G \cdot \frac{1}{\log(2)\cdot\frac{n\pi_m}{p_G}} \cdot \frac{n\pi_m}{p_G^2} \\ &= -\log_2\left(\frac{n\pi_m}{p_G}\right) + \frac{1}{\log(2)} \\ &= -\log_2\left(\frac{n\pi_m}{e\cdot p_G}\right) = 0\end{aligned}$$

is satisfied when $p_G = \frac{n\pi_m}{e}$. To assess whether this gives a maximum or minimum extreme, we compute

$$\frac{\partial^2}{\partial p_G^2} - p_G \log_2\left(\frac{n\pi_m}{p_G}\right)\Big|_{p_G=\frac{n\pi_m}{e}} = \frac{1}{\log(2)\cdot\frac{n\pi_m}{e\cdot p_G}} \cdot \frac{n\pi_m}{e^2 p_G^2}\Big|_{p_G=\frac{n\pi_m}{e}} = \frac{1}{\log(2) n\pi_m} > 0.$$

Hence, this is a minimum. The other two possibilities for p_G are its boundary values. Note that when $p_G = n$, then also

$$-p_G \log_2\left(\frac{n\pi_m}{p_G}\right) = -n\log_2(\pi_m) \to \infty, \quad \text{as } n \to \infty.$$