The illusion of group testing

Teddy Furon

▶ **To cite this version:**

Teddy Furon. The illusion of group testing. [Research Report] RR-9164, Inria Rennes Bretagne Atlantique. 2018, pp.1-19. hal-01744252

HAL Id: hal-01744252
https://inria.hal.science/hal-01744252

Submitted on 27 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

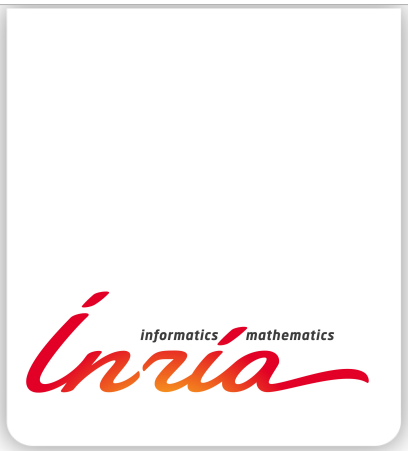
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The illusion of group testing

Teddy Furon

RESEARCH REPORT
N° 9164
March 2018
Project-Team Linkmedia

ISSN 0249-6399 ISRN INRIA/RR--9164--FR+ENG



The illusion of group testing

Teddy Furon

Project-Team Linkmedia

Research Report n° 9164 — March 2018 — 19 pages

Abstract: This report challenges the assumptions usually made in non-adaptive group testing. The test is usually modelled as a probabilistic mechanism prone to false positive and / or false negative errors. However, the models are still too optimistic because the performances of these non ideal tests are assumed to be independent of the size of the groups. Without this condition, the report shows that the promises of group test (a number of tests and a decoding complexity scaling as $c \log N$) do not hold.

Key-words: Group testing, hypothesis testing, identification, information theory

This work has been done for the CHIST-ERA project IDentification for the Internet Of Things - ID_IOT

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

L'illusion des tests par groupe

Résumé : Ce rapport de recherche présente une investigation sur les hypothèses parfois cachées en test par groupe. Pour un nombre c de malades sur une population de taille N, on dit souvent qu'il suffit de $O(c \log N)$ tests pour identifier les malades. Ce résultat est erroné dès que les performances du test s'effondrent avec la taille du groupe.

Mots-clés : Test par groupe, test d'hypothèse, identification, théorie de l'information

1 Introduction

Group testing has recently received a surge of research works mainly due to its connection to binary compressed sensing [3, 1, 13, 14]or to traitor tracing [11, 9]. The usual setup is often described in terms of clinical screening as it was the first application of group testing [5]. Among a population of N individuals, there are c infected people, with c much smaller than N. Screening the whole population by individual blood test is too costly. However, it is possible to mix blood samples from several persons and to perform a test. Ideally, the test is negative if none of these persons are infected, and positive if at least one of them is infected. The application of group testing are nowadays DNA screening [12], signal processing [6], machine learning [17]. Indeed, group testing may be a solution to any 'needles in haystack' problem, *i.e.* aiming at identifying among a large collection the few 'items' sharing a peculiar property detectable by a test, provided that this test can be performed on groups of several items. In this paper, we use the terminology of items and defective items.

The dominant strategy nowadays is called non-adaptive group testing [1, 3]. A first stage pools items into groups and performs the tests. A second stage, so-called decoding, analyses the result of these tests to identify the defective items. Tests and decoding are sequential. If the number of tests M is sufficiently big, the decoding stage has enough information to identify the defective items. In a nutshell, the groups are overlapping in the sense that one item is involved in several tests. Decoding amounts at finding the smallest subset of items which would trigger the observed positive tests.

The promises of group testing are extremely appealing. First, the theoretical number of tests M asymptotically scales as $O(c \log N)$ as N goes to infinity [1, 9, 13]. This result holds even if c increases with N, but at a lower rate [13, 14]. Second, recent papers propose practical schemes non only achieving this efficiency (or almost, *i.e.* $O(c \log c \log N)$) but also within a decoding complexity of $O(c \log N)$ (or almost, *i.e.* $O(c \log c \log N)$) [2, 10].

This paper takes a complete opposite point of view. The number c of defective items is fixed, and we don't propose more efficient design. On contrary, we show that these promises hold only for some specific probabilistic models. These models are well known in the literature of group testing. They do take into account some imperfection in the test process, however, they are somehow optimistic. As group testing becomes popular, people applying this technique to their 'needles in haystack' problems might be disappointed. The promises of group testing (a number of tests in $O(c \log N)$ together with a computational complexity of $O(c \log N)$) fade away for applications not compliant with these models.

The goal of this paper is to investigate what is specific in these models and to better understand the conditions necessary for achieving the promises of group testing. This paper has the following structure. Section 2 describes the recent approaches achieving both the minimum asymptotic number of tests and the minimal decoding complexity. The usual models are introduced together with an information theoretic justification that these approaches are sound. Section 3 introduces some more general models and shows that the total number of tests no longer scales as $O(c \log N)$ in most cases.

2 Previous works

A typical paper about non-adaptive group testing proposes a scheme, which is composed of a pooling design and a decoding algorithm. The pooling is the way M groups are composed from a collection of N items. The decoding receives the M binary test results (positive or negative) to infer the defective items. Under a definition of a successful decoding and some probabilistic models, the paper then shows how the necessary number of tests asymptotically scales. For

instance, if the decoding aims at identifying all the defective items, the authors show how M should scale as $N \to \infty$ to make the probability of success converge to one. The best asymptotical scaling has been proven to be in $O(c \log N)$ in theoretical analysis [1, 9].

2.1 Notations and models

The assumptions of the proof of a typical group testing paper concern the distribution of the defective items in the collection and the model of the test. Denote $\mathbf{x}$ a binary vector of dimension N encoding which items are defective: $x_i = 1$ if the i-th item is defective, 0 otherwise. $\mathbf{X}$ is the random variable associated to this indicator vector. We assume that there are a fixed number c of defective s.t. $\mathbb{P}[\mathbf{X} = \mathbf{x}] = 1/\binom{N}{c}$ if $|\mathbf{x}| = c$, 0 otherwise.

As for the test, the models define the probabilistic behavior of its output. Suppose a group $\mathcal{G}_i$ of n items, and let $0 \leq K_i \leq \max(n, c)$ be the random variable encoding the number of defectives in this group. Denote first by Z_i a binary r.v. s.t. $Z_i = 1$ if $K_i > 0$, 0 otherwise. Now denote by $Y_i \in \{0, 1\}$ the r.v. modeling the output of the test performed on this group. There are four well known models:

1. Noiseless test: $Y_i = Z_i$. The test is positive if and only if there is at least one defective in a group,

2. Noisy test: $Y_i = Z_i \oplus N_i$ with N_i independent and identically distributed as Bernouilli $\mathcal{B}(\epsilon)$ and $\oplus$ the XOR operator.

3. Dilution: $Y_i = \vee_{j \in \mathcal{G}_i}[X_j \wedge W_{i,j}]$, where $\wedge$ and $\vee$ are the AND and OR operators and $W_{i,j}$ a binary r.v. modeling the detectability of the j-th item in the i-th group. These random variables are independent (both along i and j) and identically distributed: $W_{i,j} \sim \mathcal{B}(1-\upsilon)$. For a given defective and test, the probability of being diluted (*i.e.* not detectable) is υ.

4. Threshold: $Y_i = 0$ if $K_i \leq \kappa_L$ and $Y_i = 1$ if $K_i \geq \kappa_U$. There are plenty variants describing what happens for $\kappa_L < K_i < \kappa_U$ [4].

Note that some models can be 'concatenated': we can witness a dilution phenomenon of parameter υ followed by a noise channel of parameter ϵ.

Another way to model a test is through the $c + 1$ parameters $(\theta_0, \cdots, \theta_c)$ defined as the following probabilities:

$$\theta_k := \mathbb{P}[Y_i = 1 | K_i = k]. \tag{1}$$

Parameter θ_0 is thus the probability of a false positive, whereas $1 - \theta_k$ for $0 < k \leq c$ are the probabilities of false negative when k defectives pertain to the test group. For the models above mentioned, we have the equivalent formulation:

1. Noiseless test: $\theta_0 = 0$ and $\theta_k = 1$ for $0 < k \leq c$.

2. Noisy test: $\theta_0 = \epsilon$ and $\theta_k = 1 - \epsilon$ for $0 < k \leq c$.

3. Dilution: $\theta_k = 1 - \upsilon^k$ with the convention that $x^0 = 1$, $\forall x \in \mathbb{R}^+$.

4. Threshold: $\theta_k = 0$ if $0 \leq k \leq \kappa_L$, $\theta_k = 1$ if $\kappa_U \leq k \leq c$.

2.2 Probabilistic group testing

The next step is to create the binary design matrix $A \in \{0,1\}^{M\times N}$. This matrix indicates which items belong to which groups: $A_{i,j} = 1$ if item j is involved in test i, and 0 if not. There are constructions which are deterministic (up to a permutation over the N items) such as those relying on disjunct matrices [7, 12]. Another popular method is the probabilistic construction where $A_{i,j}$ is set to one depending on a coin flip: $\mathbb{P}[A_{i,j} = 1] = p$. These coin flips are independent w.r.t. indices i (groups) and j (items). The sequence $(A_{1,j}, \cdots, A_{M,j})$ is often called the codeword of item j. We shall focus on this last construction.

Theoretical studies [14, 13, 1] shows that there is a phase transition: it is possible to identify all defectives with an asymptotically vanishing probability of error (as $N \to \infty$) if

$$M \geq \max_{\ell\in\{1,\cdots,c\}} \frac{\ell \log_2 \frac{N}{\ell}}{I(Y; A_{\mathcal{G}_{\mathsf{dif}}}|A_{\mathcal{G}_{\mathsf{eq}}})}(1+\eta); \tag{2}$$

whereas the error probability converges to one for any decoding scheme if

$$M \leq \max_{\ell\in\{1,\cdots,c\}} \frac{\ell \log_2 \frac{N}{\ell}}{I(Y; A_{\mathcal{G}_{\mathsf{dif}}}|A_{\mathcal{G}_{\mathsf{eq}}})}(1-\eta). \tag{3}$$

The sets of items $(\mathcal{G}_{\mathsf{dif}}, \mathcal{G}_{\mathsf{eq}})$ compose a partition of the set of defective items such that $|\mathcal{G}_{\mathsf{dif}}| = \ell$ and $|\mathcal{G}_{\mathsf{eq}}| = c - \ell$, and $A_{\mathcal{G}_{\mathsf{dif}}}$ (resp. $A_{\mathcal{G}_{\mathsf{eq}}}$) denote the codewords of the items in $\mathcal{G}_{\mathsf{dif}}$ (resp. $A_{\mathcal{G}_{\mathsf{eq}}}$).

These theoretical results are extremely powerful since they still hold when c is not fixed but slowly increasing with N. They are somehow weakly related to practical decoding schemes. For instance, equation (2) comes from a genie aided setup: a genie reveals to the decoder some defective items $\mathcal{G}_{\mathsf{eq}}$, and the number of tests needed to identify the remaining ones, *i.e.* $\mathcal{G}_{\mathsf{dif}}$, is evaluated. This is done for different sizes of $\mathcal{G}_{\mathsf{eq}}$, from 0 to $c-1$. A decoder without any genie needs more than the supremum of all these quantities.

In the sequel, we consider simpler expressions of the total number of tests but related to practical (or almost) decoders.

2.3 Joint decoder

The joint decoder computes a score per tuple of c items. It spots the tuple of defective items (identifying all of them) with a probability at least $1 - \alpha_J$; and it incorrectly points a tuple of non defective items with probability β_J. Denote $\gamma_J := \log(\alpha_J)/\log(\beta_J/N^c)$. T. Laarhoven [9] showed that a sufficient and necessary number of tests is at least:

$$M_J = \frac{c \log_2 N}{\max_{p\in(0,1)} I_J(p)}(1 + O(\sqrt{\gamma_J})), \tag{4}$$

where $I_J(p) = I(Y_i, (A_{i,j_1}, \cdots, A_{i,j_c})|p)$ is the mutual information between the output of the test and the codeword symbols of the defectives $\{j_1, \cdots, j_c\}$. In other words, this corresponds to the case where the genie reveals no information: $\mathcal{G}_{\mathsf{eq}} = \emptyset$ [14].

Since $\lim_{N\to\infty} \gamma_J = 0$ for fixed (α_J, β_J), this allows to state that M_J scales as $M_J \approx c \log_2 N / I_J(p^\star_J)$ with $p^\star_J = \arg\max_{p\in(0,1)} I_J(p)$. For the equivalent model $(\theta_0, \cdots, \theta_c)$, this amounts to find the maximizer of the following function:

$$I_J(p) := h\left(P(p)\right) - \sum_{k=0}^{c} \pi_k h(\theta_k), \tag{5}$$

with

$$P(p) \quad := \quad \mathbb{P}[Y_i = 1|p] = \sum_{k=0}^{c} \pi_k \theta_k, \tag{6}$$

$$\pi_k \quad := \quad \binom{c}{k} p^k (1-p)^{c-k}, \forall 0 \leq k \leq c, \tag{7}$$

and $h(x)$ is the entropy in bits of a binary r.v. distribution as $\mathcal{B}(1, x)$. Laarhoven gives the expressions of $p_J^\star$ for large c and for the usual models [9]. The maximizer and the maximum are functions of c and of the parameters of the test model (for example, ϵ or υ for the noisy or dilution model).

The drawback is that the decoding is exhaustive: it scans the $\binom{N}{c}$ possible subsets of size c from a set of N items. Therefore its complexity is in $O(N^c)$. This is called a joint decoder as it jointly considers a subset of c items. The joint decoder is mainly of theoretical interest since its complexity is hardly tractable. Some schemes propose approximations of a joint decoder with manageable complexity resorting to Markov Chain Monte Carlo [8], Belief Propagation [15] or iterative joint decoders [11].

2.4 Single decoder

The single decoder analyses the likelihood that a single item is defective. It correctly identifies a defective item with probability $1 - \alpha_S$ while incorrectly suspecting a non defective item with probability less than β_S. Denote $\gamma_S = \log(\beta_S)/\log(\alpha_S/N)$. Laarhoven [9] showed that a sufficient and necessary number of tests is at least:

$$M_S = \frac{\log_2 N}{\max_{p\in(0,1)} I_S(p)}(1 + O(\gamma_S)) \tag{8}$$

where $I_S(p) = I(Y_i, A_{i,j_1}|p)$ is the mutual information between the output of the test and the symbol of the codeword of one defective, say j_1. Again, since $\lim_{N\to\infty} \gamma_S = 0$ for fixed (α_S, β_S), this allows to state that M_S scales as $M_S \approx \log_2 N / I_S(p_S^\star)$ with $p_S^\star = \arg\max_{p\in(0,1)} I_S(p)$. For the equivalent model $(\theta_0, \cdots, \theta_c)$, this amounts to find the maximizer of the following function:

$$I_S(p) := h\left(P(p)\right) - ph(P_1(p)) - (1-p)h(P_0(p)) \tag{9}$$

with

$$P_1(p) \quad := \quad \mathbb{P}[Y_i = 1|A_{i,j_1} = 1, p] = \sum_{k=1}^{c} \binom{c-1}{k-1} p^{k-1}(1-p)^{c-k}\theta_k \tag{10}$$

$$P_0(p) \quad := \quad \mathbb{P}[Y_i = 1|A_{i,j_1} = 0, p] = \sum_{k=0}^{c-1} \binom{c-1}{k} p^{k}(1-p)^{c-1-k}\theta_k \tag{11}$$

Laarhoven gives the expressions of $p_S^\star$ for large c and for the usual models [9]. It always holds that $I_J(p) \geq cI_S(p)$, for any $p \in [0, 1]$. This yields M_S inherently bigger than M_J [9]: Both total numbers of tests scale as $c \log N$, but with a bigger multiplicative constant for M_S. The simple decoder computes a score for each item. Therefore its complexity is linear in $O(N)$.

2.5 Divide and Conquer

Papers [2, 10] have recently proposed schemes meeting the promises of group testing as listed in the introduction: optimal scaling both in the total number of tests and decoding complexity.

Both of them are deploying a 'Divide and Conquer' approach. Identifying c defectives among a collection of N items is too complex. Their strategy splits this problem into S simpler problems. The collection is randomly split into S subsets. S is chosen such that any subset likely contains at most one defective. Indeed, their proof selects S big enough s.t., with high probability, each defective belongs at least to one subset where it is the only defective. Assume that it is possible to detect whether a subset has no, one or more defectives. Then, a group testing approach is applied on each subset containing a single defective (so called 'singleton' subset in [10]): The decoding identifies this defective thanks to the result of tests performed on groups composed of items of that subset. It turns out that identifying defectives in a collection is much simpler when knowing there is only one. In a non-adaptive framework, all group tests are performed in the first stage, but the decoding is only run on subsets deemed as 'singleton'.

We detail here our own view of this 'Divide and Conquer' approach. Papers [2, 10] slightly differ in the way subsets are created. More formally, each subset $\mathcal{S}_k$,$1 \leq k \leq S$, is composed independently by randomly picking N_S items in the collection of N items. Denote by π the probability that an item belongs to a given subset: $\pi = {}^{N_S}/_{N}$. Subset $\mathcal{S}_k$ is not useful for identifying a given defective if:

- it doesn't belong to subset $\mathcal{S}_k$ with probability $1 - \pi$,
- else, if it is not the only defective in this subset with probability $1 - H(0; N, c, N_S)$, where $H(k; N, c, N_S)$ is the hypergeometric distribution,
- else, if the decoding over this subset misses its identification with probability denoted by α.

Over all, this event happens with probability

$$g(N_S) := (1 - \pi) + \pi\left((1 - H(0; N, c, N_S) + H(0; N, c, N_S)\alpha\right) \tag{12}$$

which is minimized by selecting $N_S = {}^{N+1}/_{c+1}$ because $g(N_S) - g(N_S - 1) \leq 0$ iff $N_S \leq {}^{N+1}/_{c+1}$ (we assume that $c + 1$ divides $N + 1$). The probability that $\mathcal{S}_k$ is useless for identifying a given defective simplifies in:

$$\begin{aligned} g(N_S) &= 1 - N_S \frac{(N-c-1)!}{N!}\frac{(N-N_S)!}{(N-N_S-c)!}(1-\alpha), &(13)\\ &= 1 - \left(\frac{c}{c+1}\right)^c . \frac{1-\alpha}{c+1}.(1 + O(1/N)), &(14)\end{aligned}$$

where we use the fact that ${}^{\Gamma(N+a)}/_{\Gamma(N+b)} = N^{a-b}(1 + (a + b - 1)^{(a-b)}/_{2N} + O({}^{1}/_{N^2}))$ [16].

Suppose that the goal of the decoding is to identify on expectation a fraction $(1 - \alpha_S)$ of the defectives. The probability of missing a given defective because none of the subset is useful for identifying it equals α_S:

$$\mathbb{P}[\text{Not identifying a given defective}] = g(N_S)^S = \alpha_S. \tag{15}$$

Since $({}^{c}/_{c+1})^c \geq {}^{1}/_{e}$, it is safe to choose $S = \lceil (c + 1)e(-\log \alpha_S)/(1 - \alpha)\rceil$.

Suppose now that the goal is to identify all the defectives with probability $1 - \alpha_J$. The probability of identifying them all is given by:

$$\mathbb{P}[\text{identifying all of them}] = \left(1 - g(N_S)^S\right)^c = 1 - \alpha_J. \tag{16}$$

This can be achieved with $S \geq \lceil e(c + 1)\log(c/\alpha_J)/(1 - \alpha)\rceil$.

The point of this 'Divide and conquer' approach is that $m = \Theta(\log_2 N_S)$ tests are needed for identifying a unique defective item in a subset of size N_S and with a fixed probability of error α (see Sec. 2.4). Since the sizes of the subsets are all equal to N/c, the total number of tests scales as $M_{DC} = O(c \log c \log N/c)$ to identify all defectives with high probability, which is almost the optimal scaling. In [10], the authors show that the decoding can also exploit subsets containing two defectives (so-called 'doubleton') which reduces $S = O(c)$ for identifying all the defectives.

To discover a fraction of the defectives the total number of tests scales as $M_{DC} = O(c \log_2 N/c)$. This ends up in the optimal scaling achieved by GROTESQUE [2] and the 'singleton' only version of SAFFRON [10].

These schemes have also the following advantages:

- The decoding complexity scales like $O(cm) = O(c \log_2 N/c)$ if a deterministic construction is used as in [2] and [10]. We decode $O(c)$ 'singleton' subsets in total. In the noiseless setup, decoding a 'singleton' amounts to read the outputs of the tests because it exactly corresponds to the codeword of the unique defective of that subset. If the setup is not noiseless, the outputs are a noisy version of this codeword. An error correcting code whose decoding is in $O(m)$ gets rid of these wrong outputs. For instance, the authors of [10] uses a spatially-coupled LDPC error correcting code.

- The decoding complexity scales like $O(cmN_S) = O(N \log_2 N/c)$ if a probabilistic construction is used per subset. We decode $O(c)$ 'singleton' subsets. Decoding a singleton amounts to compute the likelihood scores for N_S items and identifying the defective as the items with the biggest score. The likelihood is a weighted sum of the m test outputs. The next section shows that finding the optimal parameter $p^\star$ of the probabilistic construction is also simple.

The main drawback of the 'Divide and Conquer' strategy is that it doesn't apply when $\theta_1 = \theta_0$. This typically corresponds to the 'threshold' model where one unique defective is not enough to trigger the output of a test. Likewise, if $\theta_1 \approx \theta_0$, any efficient error correcting decoder will fail and the only option is the exhaustive maximum likelihood decoder. At that point, a probabilistic construction is preferable.

2.6 Identifying the unique defective in a 'singleton' subset

This 'Divide and conquer' approach greatly simplifies the model (1). Since there is a single defective, we only need parameters θ_0 and θ_1. By the same token, there is no need of joint decoding since the defective is unique. The mutual information in (8) takes a simple expression:

$$I_{DC}(p) = H(Y_i|p) - H(Y_i|A_{i,j}, p) = h(\theta_0 + p(\theta_1 - \theta_0)) - (1-p)h(\theta_0) - ph(\theta_1), \tag{17}$$

which is strictly positive on $(0,1)$ if $\theta_1 \neq \theta_0$ and whose maximisation is simpler than for the single and joint decoders. This concave function has a null derivative for:

$$p^\star_{DC} = \frac{1}{\theta_1 - \theta_0} \cdot \left(\frac{1}{2^{\frac{h(\theta_1) - h(\theta_0)}{\theta_1 - \theta_0}} + 1} - \theta_0 \right). \tag{18}$$

This gives the following application to the usual models:

1. Noiseless test: $\theta_0 = 1 - \theta_1 = 0$ so that $p^\star_{DC} = 1/2$ and $I_{DC}(p^\star) = 1$.

2. Noisy test: $\theta_0 = 1 - \theta_1 = \epsilon$ so that $p^\star_{DC} = 1/2$ and $I_{DC}(p^\star) = 1 - h(\epsilon)$.

3. Dilution: $\theta_0 = 0$ and $\theta_1 = 1 - \upsilon$ so that

$$p^\star_{DC}(\upsilon) = \frac{1}{1-\upsilon} \cdot \frac{1}{2^{h(\upsilon)/(1-\upsilon)} + 1}, \tag{19}$$

$$I_{DC}(p^\star_{DC}(\upsilon)) = h\left((1-\upsilon)p^\star_{DC}(\upsilon)\right) - p^\star_{DC}(\upsilon)h(\upsilon). \tag{20}$$

Denote $f(\upsilon) = 1/p^\star_{DC}(\upsilon)$. We have:

$$f'(\upsilon) = -1 - 2^{\frac{h(\upsilon)}{1-\upsilon}} \left(1 + \frac{\ln \upsilon}{1-\upsilon}\right) \quad \text{for } \upsilon \in [0,1]. \tag{21}$$

Since $\ln(\upsilon) \leq \upsilon - 1 - (1-\upsilon)^2/2$, we have on one hand $2^{\frac{h(\upsilon)}{1-\upsilon}} \geq 2^\upsilon/(1-\upsilon)$ and on the other hand $(1 + \frac{\ln \upsilon}{1-\upsilon})/(1-\upsilon) \leq -1/2$. This shows that $f'(\upsilon) \geq 0$. Function f is increasing, therefore $p^\star_{DC}$ is a decreasing function of the dilution factor υ. As $\upsilon \to 1$, $h(\upsilon)/(1-\upsilon) = 1/\ln(2) - \log_2(1-\upsilon) + O(1-\upsilon)$, s.t. $p^\star_{DC} \to 1/e \approx 0.37$.

The number of tests for a subset is given by (8) which multiplied by the number of subsets gives

$$M_{DC} \approx \frac{ce(-\log \alpha_S)}{I_{DC}(p^\star)} \log_2(N/c) \tag{22}$$

for identifying a fraction $1 - \alpha_S$ of defectives on expectation.

3 Less optimistic models

We would like to warn the reader that the promises of group testing are due to the simplicity of the models described in Sec. 2.1. These models are not naive since they do encompass the imperfection of the test over a group. The output of the test is modeled as a random variable. Yet the statistics of the test only depend on the number of defective items inside the group, but not on the size of the group.

Consider the noisy setup with parameter ϵ modelling the imperfection of the test. The optimal setting is $p^\star_{DC} = 1/2$ for the 'Divide and Conquer' scheme. This means that if $N = 200$, then the size of the groups is around 100, if $N = 2 \cdot 10^9$, the groups are composed of a billion of items and, still, the reliability of the test is not degraded. There are some chemical applications where tests can detect the presence of one single particular molecule among billions. But this is certainly not the case of all 'needles in haystack' problems.

3.1 Our proposed model

We believe there are many applications where the reliability of the test degrades as the size n of the group increases. Indeed, when the size of the group grows to infinity, the test might become purely random. For the noisy setup, ϵ should be denoted as ϵ_n s.t. $\lim_{n\to+\infty} \epsilon_n = 1/2$ if the test gets asymptotically random. For the dilution model, υ should be denoted as υ_n s.t. $\lim_{n\to+\infty} \upsilon_n = 1$. This captures the fact that the defectives get completely diluted in groups whose size grows to infinity.

Instead of coping with the noisy or dilution setups, we prefer to consider the equivalent model where probabilities $(\theta_{0,n}, \cdots, \theta_{c,n})$ now depend on the size of the group. We make the following assumptions:

- For all n, $\theta_{0,n} \leq \cdots \leq \theta_{c,n}$. Having more defective items in a group increases the probability that the test is positive.

- $\theta_{0,n}$ is a non decreasing function of n. Parameter $\theta_{0,n}$ is the probability of a false positive (the test is positive whereas there is no defective in the group). Increasing the size of the group will not help decreasing the probability of this kind of error.

- For $0 < k \leq c$, $\theta_{k,n}$ is a non increasing function. Again, $1-\theta_{k,n}$ is the probability of a false negative (the test is negative whereas there k defective items in the group). Increasing the size of the group will not help decreasing the probability of this kind of error.

- These probabilities are bounded monotonic functions, therefore they admit a limit as $n \to +\infty$, denoted as $\bar{\theta}_k := \lim_{n\to+\infty} \theta_{k,n}$.

A test is deemed as asymptotically random if $\bar{\theta}_0 = \cdots = \bar{\theta}_c$, whatever the value of this common limit.

3.2 Application to group testing designs

We consider the three schemes above-mentioned: single, joint and 'Divide and conquer'. As described in Sec. 2.5, the 'Divide and conquer' design builds S subsets by randomly picking N_S items out of N. The optimal size N_S of a subset grows linearly with N. The three schemes then compose random groups from a population whose size N, be it $\mathsf{N} = N$ (single and joint schemes) or $\mathsf{N} = N_S$ ('Divide and conquer' design), grows to infinity. We denote the size of the test groups by $n(\mathsf{N})$ to investigate different choices as N goes to infinity. Note that $n(\mathsf{N}) \leq \mathsf{N}$. When we pick up at random $n(\mathsf{N})$ items the probability p that this group contains a given defective is $p = n(\mathsf{N})/\mathsf{N}$.

We analyze two choices concerning the asymptotical size of the test groups: either $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = +\infty$ or $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) < +\infty$. We derive the mutual information at stake for the three schemes ('Divide and Conquer', simple, and joint) and we deduce the asymptotical scaling of the total number of tests.

We introduce the non increasing functions $\delta_{k,n} := \theta_{k,n} - \theta_{0,n} \geq 0$. For the 'Divide and conquer' scheme, the decoding is performed on singleton subset. Therefore $\delta_{1,n}$ shows the speed at which the test gets closer to a random test. For the simple and joint decoder, there are up to c defective items in a group (we suppose that $c < n$) and $\delta_{c,n}$ shows the speed at which the test gets closer to a random test.

The most important factor is the limit $\bar{\delta}_k := \lim_{n\to\infty} \delta_{k,n}$. For the 'Divide and conquer' scheme, $\bar{\delta}_1 > 0$ means that the two probabilities $\theta_{0,n}$ and $\theta_{1,n}$ have distinct limits, and Sect. 2.6 gives the optimum choice replacing θ_0 and θ_1 by their limits $\bar{\theta}_0$ and $\bar{\theta}_1$. When there is a single defective in a subset, Eq. (8) shows that the number of tests to identify it is $\Theta(\log_2 N_S)$, which in turn is $\Theta(\log_2 {}^N\!/_c)$. Because the number of subsets S used by the 'Divide and Conquer' strategy in Sect. 2.5 asymptotically gets independent of N, the total number of tests scales as $\Theta(c \log_2 {}^N\!/_c)$ (identification of a fraction of defective items) or $\Theta(c \log c \log_2 N)$ (identification of all the items).

For the single and joint decoders, $\bar{\delta}_c > 0$ means that the test is not asymptotically random. Packing more and more items in groups always provides informative tests. The strategy of selecting $n(\mathsf{N})$ s.t. $\lim_{\mathsf{N}\to\infty} n(\mathsf{N})/\mathsf{N} = p^\star$ will deliver a non null mutual information. Parameter $p^\star$ is derived from Sec. 2 replacing $(\theta_0, \cdots, \theta_c)$ by their limits $(\bar{\theta}_0, \cdots, \bar{\theta}_c)$. There might be other way giving a lower total number of tests, but at least this strategy delivers the promises of group testing with a total number of tests scaling as $\Theta(c \log_2 N)$.

The next sections investigate our main concern : the case where the test is asymptotically random.

4 First strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \bar{n}$

The first strategy makes the size of the test groups converging to the finite value $\bar{n} := \lim_{\mathsf{N}\to\infty} n(\mathsf{N})$ for which the test is not random: suppose $\theta_{0,\bar{n}} < \theta_{1,\bar{n}}$. On the other hand, the probability of an item being in a given test group vanishes as $p = \bar{n}/\mathsf{N}$.

4.1 The case where $\theta_{0,\bar{n}} \neq 0$

Assume first that $\theta_{0,\bar{n}} \neq 0$, Taylor series give the following asymptotics (See Appendices A.1, B.1, and C.1):

$$I_J(p) \approx c\frac{\bar{n}}{N}\Delta h_{0,1}, \tag{23}$$

$$I_S(p) \approx \frac{\bar{n}}{N}\Delta h_{0,1}, \tag{24}$$

$$I_{DC}(p) \approx \frac{\bar{n}}{N_S}\Delta h_{0,1}, \tag{25}$$

with $\Delta h_{0,1} := [(\theta_{1,\bar{n}} - \theta_{0,\bar{n}})h'(\theta_{0,\bar{n}}) + h(\theta_{0,\bar{n}}) - h(\theta_{1,\bar{n}})]$. These three mutual informations only depend on the first two parameters of the model which is unusual for the joint and the single schemes. As the probability p vanishes, the tests are positive for a unique reason: there is a single defective in the groups. More formally, thanks to L'Hôspital's rule, the probability that there is a single defective knowing that there at least one converges to zero:

$$\lim_{p\to 0} \frac{\pi_1}{1-\pi_0} = \lim_{p\to 0} 1 - (c-1)\frac{p}{1-p} = 1 \tag{26}$$

It is therefore quite normal that $I_{DC}(p)$ and $I_S(p)$ coincide (except that N_S is replaced by N). However, the 'Divide and Conquer' scheme runs a group testing procedure per subset s.t. it asymptotically needs $e(-\log\alpha_S)$ more tests than the single decoder (comparison of (22) with (8)).

$I_J(p)$ is exactly c times bigger than $I_S(p)$, which in the end offers the same scaling for the total number of tests: $M_J \approx M_S$ (comparison of (4) with (8)). This signifies that the joint decoding doesn't perform better than the single decoding. Indeed, the score computed for a tuple of c items by the joint decoder becomes asymptotically equal to the sum of the scores of the c items as computed by the single decoder.

The three schemes need a total number of tests scaling as $O(N\log N)$. It is surprising that it doesn't depend on c, but the most important point is that this is much less appealing than the promise in $O(c\log N)$.

4.2 The case where $\theta_{0,\bar{n}} = 0$

When $\theta_{0,\bar{n}} = 0$, the expressions above are no longer correct because $\lim_{x\to\infty} h'(x) = \infty$. New Taylor series give the following asymptotics (See Appendices A.1, B.1, and C.1):

$$I_J(p) \approx c\theta_{1,\bar{n}}\frac{\bar{n}}{N}\log_2\frac{N}{\bar{n}}, \tag{27}$$

$$I_S(p) \approx \theta_{1,\bar{n}}\frac{\bar{n}}{N}\log_2\frac{N}{\bar{n}}, \tag{28}$$

$$I_{DC}(p) \approx \theta_{1,\bar{n}}\frac{\bar{n}}{N_S}\log_2\frac{N_S}{\bar{n}}. \tag{29}$$

The same comments as above hold except that this time the schemes provide a better total number of tests scaling as $O(N)$. The explanation is that a test s.t. $\theta_{0,\bar{n}} = 0$ has the advantage

of being positive if and only if there is at least one defective in the group. Indeed, there is a single defective in a positive group exactly in the 'Divide and Conquer' scheme and asymptotically for the joint and single decoders. This certainty eases a lot the decoding.

These mutual informations show that the multiplicative factor of this scaling is $1/(\bar{n}\theta_{1,\bar{n}})$ for the single and joint decoders. We thus need to select $\bar{n}$ s.t. $\bar{n}\theta_{1,\bar{n}} > 1$ in order to be, asymptotically at least, preferable to an exhaustive search testing items separately. This raises an even more stringent condition for the 'Divide and Conquer' scheme because we need $\bar{n}\theta_{1,\bar{n}} > e(-\log \alpha_S)$.

5 Second strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \infty$

The second strategy makes the size of the test groups increasing as $\mathsf{N} \to \infty$. Therefore, the rate at which the test becomes random matters. This is reflected by the speed at which $\delta_{1,n}$ ('Divide and Conquer') or $\delta_{c,n}$ (joint and single) converge to zero. Once again, we make the distinction between tests s.t. $\bar{\theta}_0 > 0$ and those for which $\bar{\theta}_0 = 0$.

5.1 The case where $\bar{\theta}_0 \neq 0$

Assume first that $\bar{\theta}_0 \neq 0$, Taylor series give the following asymptotics (See Appendices A.2.1, B.2.1 and C.2.1):

$$I_J(p) \approx -\frac{1}{2}h''(\bar{\theta}_0)\mathsf{Var}[\theta_{K,n}], \tag{30}$$

$$I_S(p) \approx -\frac{1}{2}h''(\bar{\theta}_0)\frac{1}{c^2 p(1-p)}\mathsf{Cov}(K,\theta_{K,n})^2, \tag{31}$$

$$I_{DC}(p) \approx -\frac{1}{2}h''(\bar{\theta}_0)p(1-p)\delta_{1,n}^2, \tag{32}$$

with $K \sim \mathcal{B}(c,p)$.

Since $\mathsf{Cov}(K,\theta_{K,n})^2 \leq \mathsf{Var}[K]\mathsf{Var}[\theta_{K,n}]$ and $\mathsf{Var}[K] = cp(1-p)$, these series comply with the rule that $I_S(p) \leq I_J(p)/c$. We can also check that if $c = 1$, these three series are equal. Another remark: If $\theta_{K,n} = K\delta_{1,n} + \theta_{0,n}\,\forall 0 \leq K \leq c$, then $I_{DC}(p) = I_S(p) = I_J(p)/c$ and the three schemes provide the same scaling of the total number of tests.

Note however that the probability p of belonging to a group equals n/N in the first two expressions, whereas it equals n/N_S in the last expression.

Application to the noisy group testing: In this setup, $\theta_{0,n} = 1 - \theta_{k,n} = \epsilon_n \to 1/2$. This simplifies the expressions above as follows:

$$I_J(p) \approx \frac{2}{\ln 2}(1-p)^c(1-(1-p)^c)\delta_{1,n}^2 \tag{33}$$

$$I_S(p) \approx \frac{1}{\ln 2}p(1-p)^{2c-1}\delta_{1,n}^2 \tag{34}$$

$$I_{DC}(p) \approx \frac{2}{\ln 2}p(1-p)\delta_{1,n}^2. \tag{35}$$

If we suppose that $\delta_{1,n} = O(n^{-a})$ with $a > 0$ and we increase the size of the group s.t. $n \propto N^b$ with $0 \leq b \leq 1$ (or $n \propto N_S^b$ for the 'Divide and Conquer' scheme), then the three schemes offer a mutual information of the same order:

- If $0 < a \leq 1/2$, the best option is to choose $b = 1$, *i.e.* to fix the value of p, to achieve $I = O(N^{-2a})$, which ends up in a total number of test scaling as $\Omega(N^{2a} \log N)$.
- If $a > 1/2$, the best option is to set $b = 0$, *i.e.* to fix the size of the group, to achieve $I = O(N^{-1})$, which ends up in a total number of tests scaling as $\Omega(N \log N)$. We rediscover here the results of Sect. 4.

These scalings are much bigger than the promised $\Theta(c \log N)$. Yet, if the test smoothly becomes random as n increases, *i.e.* when $a < 1/2$, the situation is actually not that bad since the scale of the total number of tests is slower than $\Theta(N)$, *i.e.* the scaling of the exhaustive screening (yet, we need a setup where the Ω becomes a Θ).

The appendices gives upper bounds of the mutual informations of the single and joint decoders in the general case:

$$I_J(p) \lesssim -\frac{1}{2} h''(\bar{\theta}_0)(1-(1-p)^c)\delta_{c,n}^2, \tag{36}$$

$$I_S(p) \lesssim -\frac{1}{2} h''(\bar{\theta}_0)\frac{p}{1-p}\delta_{c,n}^2 \tag{37}$$

These two upper bounds share the same decrease in $O(N^{-2a})$ if $\delta_{1,n} = O(n^{-a})$ with $0 < a \leq 1/2$.

Now to get $M = O((\log N)^d)$ we need to have $I = \Omega((\log N)^{1-d})$ and therefore, for a fixed p, $\delta_{c,n}$ (or $\delta_{1,n}$ for the 'Divide and Conquer' scheme) being $\Omega((\log n)^{1-d/2})$. The point of this chapter is to consider that $\delta_{c,n}$ converges to zero, therefore $d > 1$. We are getting closer to the promise of group testing for tests becoming random at a very low speed.

5.2 The case where $\bar{\theta}_0 = 0$

The appendices A.2.2, B.2.2 and C.2.2 show that:

$$\begin{aligned} I_J(p) &\leq (-\theta_{c,n}\log_2\theta_{c,n})(1-\pi_0-\pi_c) + \theta_{c,n}(-(1-\pi_0)\log_2(1-\pi_0)) + o(\theta_{c,n}) && (38)\\ I_S(p) &\leq (-\theta_{c,n}\log_2\theta_{c,n})(1-\pi_0-\pi_c) + \theta_{c,n}(-(1-\pi_0)\log_2(1-\pi_0) \\ &+ (c-1)p^c\log_2 p) + o(\theta_{c,n}) && (39)\\ I_{DC}(p) &= \theta_{1,n}(-p\log_2 p) + o(\theta_{1,n}), && (40) \end{aligned}$$

with $\pi_0 = (1-p)^c$ and $\pi_c = p^c$.

For a fixed p, the mutual informations of the joint and single decoders are dominated by $-\theta_{c,n}\log_2\theta_{c,n}$. If $\theta_{c,n} = O(n^{-a})$, $a > 0$, the total number of tests scales as $\Omega(N^a)$. This does not hold for the 'Divide and Conquer' scheme: if $\theta_{1,n} = O(n^{-a})$, the total number of tests scales as $\Omega(N^a \log N)$.

If $n \propto N^b$, $0 \leq b \leq 1$ so that $p \propto N^{b-1}$, then the mutual informations of the joint and single decoders are $O(N^{b(1-a)-1}\log N)$. If the test is slowly converging to a random test, *i.e.* $a < 1$, then we should set $b = 1$ and we are back to the option of freezing p. Otherwise, it is better to set $b = 0$ so that the total number of tests scales as $\Omega(N)$, and we find back the first strategy fixing n. The same comment holds for the the 'Divide and Conquer' scheme.

Again, this case is preferable to the case $\bar{\theta}_0 \neq 0$: $M = \Omega(N^a \log N)$ and not $\Omega(N^{2a}\log N)$, and for a longer range $0 < a \leq 1$ (and not $0 < a \leq 1/2$).

Last but not least, for the 'Divide and Conquer' scheme, to get $M = O((\log N)^d)$ we need to have $\theta_{1,n} = \Omega((\log N)^{1-d})$ with $d > 1$ to make $\theta_{1,n}$ vanishing as $n = pN$ increases (p is fixed). With the same setup, $M = O(^{(\log N)^d}/_{\log\log N})$ for the single and joint decoders.

6 Conclusion

The point of this chapter is not to find the best choice concerning the asymptotical size of the test groups. We just show that whatever this choice, group testing fails delivering the promise of a total number of tests scaling as $O(c \log N)$. The condition of utmost importance for such an appealing scaling is to have a test which doesn't become purely random as the size of the group grows to infinity. However, group testing almost keeps its promise, *i.e.* a total number of tests scaling as a power of $\log N$, for setups where the test converges to randomness very slowly, *i.e.* at a rate in $\Omega(1/\log^g n)$ with $g > 0$.

For this kind of setups, it is better to fix p, which means that the size of the groups are proportional to N. However, if the test becomes random too rapidly, *i.e.* as fast as $O(n^{-a})$ with $a \geq 1/2$, it is useful to switch from a fixed p strategy to a fixed n strategy.

Setups where there is no false positive ($\theta_{0,n} = 0$) or no false negative ($\theta_{k,n} = 1$ for $k > 0$) lead to better performances: the total number of tests is lower and the transition from fixed p to fixed n occurs at a higher rate, *i.e.* for $a = 1$.

A 'Divide and conquer'

A.1 First strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \bar{n}$

The first strategy makes the size of the test groups converging to the finite value $\bar{n} := \lim_{N_S\to\infty} n(N_S)$ for which the test is not random, *i.e.* $\theta_{0,\bar{n}} < \theta_{1,\bar{n}}$. On the other hand, the probability of an item being in a given test group vanishes as $p = \bar{n}/N_S$.

Assume that $\theta_{0,\bar{n}} \neq 0$, a Taylor series of (17) gives the following asymptotic:

$$I_{DC}(p) = \frac{\bar{n}}{N_S}\left[(\theta_{1,\bar{n}} - \theta_{0,\bar{n}})h'(\theta_{0,\bar{n}}) + h(\theta_{0,\bar{n}}) - h(\theta_{1,\bar{n}})\right] + o(N_S^{-1}). \tag{41}$$

If $\theta_{0,\bar{n}} = 0$, the result above does not hold because $\lim_{x\to 0} h'(x) = +\infty$. A new Taylor series gives the following asymptotic:

$$I_{DC}(p) = \theta_{1,\bar{n}}\frac{\bar{n}}{N_S}\log_2\frac{N_S}{\bar{n}} + o\left(\frac{1}{N_S}\log(N_S)\right). \tag{42}$$

A.2 Second strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \infty$

A.2.1 When $\bar{\theta}_0 \in]0,1[$

The assumption here is that $\bar{\theta}_1 = \bar{\theta}_0$ which lies in $]0,1[$. We denote $\eta_n := \theta_{0,n} - \bar{\theta}_0$ and $\delta_{1,n} := \theta_{1,n} - \theta_{0,n}$. With these notations, we have

$$I_{DC}(p) = h(\bar{\theta}_0 + \eta_n + p\delta_{1,n}) - (1-p)h(\bar{\theta}_0 + \eta_n) - ph(\bar{\theta}_0 + \eta_n + \delta_{1,n}), \tag{43}$$

Note that $\eta_n \leq \eta_n + p\delta_{1,n} \leq \eta_n + \delta_{1,n}$ because $0 \leq p \leq 1$, which implies that

$$|\eta_n + p_n\delta_{1,n}| \leq \max(|\eta_n|, |\eta_n + \delta_{1,n}|). \tag{44}$$

Both $|\eta_n|$ and $|\delta_{1,n}|$ converges to 0 so that, for $\epsilon > 0$, there exist n_0 big enough s.t. $\forall n \geq n_0$, $\max(|\eta_n|, |\eta_n + \delta_{1,n}|) \leq \epsilon$. We then apply the following Taylor development for $\bar{\theta}_0 \in]0,1[$:

$$h(\bar{\theta}_0 + \epsilon) = h(\bar{\theta}_0) + \epsilon h'(\bar{\theta}_0) + \epsilon^2 h''(\bar{\theta}_0)/2 + o(\epsilon^2), \tag{45}$$

on the three terms of (43) to simplify it to:

$$I_{DC}(p) \quad = \quad -\frac{1}{2}\delta_{1,n}^2 h''(\bar{\theta}_0)p(1-p) + o(\epsilon^2). \tag{46}$$

Since we assume in the text that $\theta_{0,n}$ is non decreasing, it has to converge to $\bar{\theta}_0$ from below s.t. η_n is non positive. In the same way, $\theta_{1,n}$ converges to $\bar{\theta}_0$ from above s.t. $\eta_n + \delta_{1,n}$ is non negative. This shows that $\epsilon \leq \delta_{1,n} \leq 2\epsilon$ and therefore $\delta_{1,n} = \Theta(\epsilon)$. This allows to replace $o(\epsilon^2)$ by $o(\delta_{1,n}^2)$ in (46).

A.2.2 When $\bar{\theta}_0 \in \{0, 1\}$

We detail the case for $\bar{\theta}_1 = \bar{\theta}_0 = 0$. With the same notations as in App. A.2.1, this case implies that $\eta_{0,n} = 0$ because $\theta_{0,n}$ is non decreasing and non negative. The mutual information in this context equals:

$$I_{DC}(p) = h(p\delta_{1,n}) - ph(\delta_{1,n}). \tag{47}$$

For $\epsilon > 0$, there exist n big enough for which $\delta_{1,n} = \epsilon$ and where $h(\epsilon) = -\epsilon \log_2(\epsilon) + \epsilon / \ln 2 + o(\epsilon)$. Applying this development, we obtain:

$$I_{DC}(p) = \delta_{1,n}(-p \log_2 p) + o(\delta_{1,n}). \tag{48}$$

B Joint decoder

We assume that the size of a group is always larger than c. Therefore, $(c+1)$ parameters define the test $(\theta_{0,n}, \cdots, \theta_{c,n})$.

B.1 First strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \bar{n}$

The mutual information for the joint decoder (5) has the following Taylor series when $p = \bar{n}/N \to 0$, for $\theta_{0,\bar{n}} > 0$:

$$I_J(p) = c\frac{\bar{n}}{N}\left[(\theta_{1,\bar{n}} - \theta_{0,\bar{n}})h'(\theta_{0,\bar{n}}) + h(\theta_{0,\bar{n}}) - h(\theta_{1,\bar{n}})\right] + o(1/N), \tag{49}$$

and, for $\theta_{0,\bar{n}} = 0$ and $\theta_{1,\bar{n}} > 0$:

$$I_J(p) = c\theta_{1,\bar{n}}\frac{\bar{n}}{N}\log(N) + o(N^{-1}\log N). \tag{50}$$

We have supposed that $\bar{n}$ is chosen s.t. $\delta_{1,\bar{n}} > 0$. It is possible to relax this constraint and the first non nul parameter $\delta_{k,\bar{n}}$ will appear in the above equations. Yet, we also get a decay in N^{-k} instead of N^{-1}, whence choosing $\bar{n}$ s.t. $\delta_{1,\bar{n}} = 0$ should be avoided if possible. This is a real issue for the threshold group testing model where $\theta_{1,n} = \theta_{0,n}$ (see Sect. 2.1).

B.2 Second strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \infty$

Now suppose that the parameters of the model vary with n s.t. $\theta_{0,n} \leq \theta_{1,n} \leq \cdots \leq \theta_{c,n}$, and that $\delta_{c,n} = \theta_{c,n} - \theta_{0,n}$ vanishes to 0 as n increases. The analysis is made for a fixed $0 < p < 1$. We remind that $\pi_k = \binom{c}{k} p^k (1-p)^{c-k}$, $\forall 0 \leq k \leq c$, the distribution of the binomial $\mathcal{B}(c, p)$.

B.2.1 When $\bar{\theta}_0 \in]0,1[$

As in the previous section, $\eta_n = \theta_{0,n} - \bar{\theta}_0$ and $\delta_{k,n} = \theta_{k,n} - \theta_{0,n}$ converge to zero. For any $\epsilon > 0$, there exists n large enough for which $\max(|\eta_n|, \eta_n + \delta_{c,n}) = \epsilon$. This implies that $|\eta_n + \delta_{k,n}|$, $\forall 0 \leq k \leq c$, and $|\eta_n + P(p) - \theta_{0,n}|$ are smaller than ϵ.

For $0 < \bar{\theta}_0 < 1$, we apply the development (45) to $h(P(p))$ and $h(\theta_{k,n})$:

$$I_J(p) = -\frac{1}{2}h''(\bar{\theta}_0)\left(\sum_{k=0}^{c}\pi_k\theta_{k,n}^2 - P(p)^2\right) + o(\epsilon^2) \quad (51)$$

$$= -\frac{1}{2}h''(\bar{\theta}_0)\mathsf{Var}[\theta_{K,n}] + o(\epsilon^2). \quad (52)$$

Note that $\mathsf{Var}[\theta_{K,n}] \leq \mathbb{E}[(\theta_{K,n} - \theta_{0,n})^2] \leq \delta_{c,n}^2(1-(1-p)^c) < 4\epsilon^2$. On the other hand,

$$\mathsf{Var}[\theta_{K,n}] \geq \pi_0(\theta_{0,n} - P(p))^2 + \pi_c(\theta_{c,n} - P(p))^2 \geq \frac{\pi_0\pi_c}{\pi_0+\pi_c}\delta_{c,n}^2 \geq \epsilon^2. \quad (53)$$

This shows that $\mathsf{Var}[\theta_{K,n}] = \Theta(\epsilon^2)$ so that we can replace $o(\epsilon^2)$ by $o(\mathsf{Var}[\theta_{K,n}])$.

B.2.2 When $\bar{\theta}_0 \in \{0,1\}$

We start by applying the development $h(x) = -x\log_2(x) + {}^{x}/_{\ln 2} + o(x)$ on $h(P(p))$ and $h(\theta_{k,n})$:

$$I_J(p) = (-P(p)\log_2(P(p))) - \sum_{k=0}^{c}\pi_k(-\theta_{k,n}\log_2\theta_{\theta_{k,n}}) + o(\theta_{c,n}). \quad (54)$$

The function $x \mapsto -x\log_2(x)$ is increasing over $[0, {}^{1}/_{e})$ and $P(p) \leq \theta_{c,n}(1-\pi_0)$. On the other hand $\sum_{k=0}^{c}\pi_k(-\theta_{k,n}\log_2\theta_{\theta_{k,n}}) \geq \pi_c(-\theta_{c,n}\log_2\theta_{c,n})$. This inequality follows from these arguments:

$$I_J(p) \leq (-\theta_{c,n}\log_2\theta_{c,n})(1-\pi_0-\pi_c) + \theta_{c,n}(-(1-\pi_0)\log_2(1-\pi_0)) + o(\theta_{c,n}). \quad (55)$$

C Single decoder

We assume that the size of a group is always larger than c. Therefore, $(c+1)$ parameters $(\theta_{0,n}, \cdots, \theta_{c,n})$ define the test.

C.1 Asymptotical analysis for $\lim_{N_S\to\infty} n(N_S) < \infty$

Assume that $\theta_{0,\bar{n}} \neq 0$, a Taylor series of (9) gives the following asymptotic:

$$I_S(p) = \frac{\bar{n}}{N}\left[(\theta_{1,\bar{n}} - \theta_{0,\bar{n}})h'(\theta_{0,\bar{n}}) + h(\theta_{0,\bar{n}}) - h(\theta_{1,\bar{n}})\right] + o(1/N). \quad (56)$$

If $\theta_{0,\bar{n}} = 0$, a Taylor series gives the following asymptotic:

$$I_S(p) = \theta_{1,\bar{n}}\frac{\bar{n}}{N}\log_2\frac{N}{\bar{n}} + o\left(\log(N)/N\right). \quad (57)$$

If $\theta_{1,\bar{n}} = \theta_{0,\bar{n}}$, the Taylor series wil show the role of the first non nul coefficient $\delta_{k,n}$ but fraction ${}^{\bar{n}}/_{N}$ must be replaced by $({}^{\bar{n}}/_{N})^k$. This should be avoided as it

C.2 Second strategy: $\lim_{\mathsf{N}\to\infty} n(\mathsf{N}) = \infty$

We first present some relations between $P(p)$, $P_1(p)$ and $P_0(p)$:

$$\begin{aligned} P_1(p) &= P(p) + (1-p)P'(p)/c, && (58)\\ P_0(p) &= P(p) - pP'(p)/c, && (59)\\ P'(p) &= \frac{\partial P(p)}{\partial p} = \frac{1}{p(1-p)}\sum_{k=0}^{c}\pi_k\theta_{k,n}(k-cp). && (60) \end{aligned}$$

From (10) and (11), it is clear that $\theta_{0,n} \le P_0(p)$ and $P_1(p) \le \theta_{c,n}$. We also have

$$p(1-p)P'(p) = \mathbb{E}[(K-cp)\theta_{K,n}] = \mathbb{E}[K\theta_{K,n}] - \mathbb{E}[K]\mathbb{E}[\theta_{K,n}] = \mathsf{Cov}(K,\theta_{K,n}). \quad (61)$$

with $K \sim \mathcal{B}(c,p)$ because $\mathbb{E}[K] = cp$. We introduce $K' \sim \mathcal{B}(c,p)$ independent of K. Then, on one hand:

$$\mathsf{Cov}(K-K',\theta_{K,n}-\theta_{K',n}) = \mathsf{Cov}(K,\theta_{K,n}) + \mathsf{Cov}(K',\theta_{K',n}) = 2\mathsf{Cov}(K,\theta_{K,n}), \quad (62)$$

while, on the other hand,

$$\begin{aligned} \mathsf{Cov}(K-K',\theta_{K,n}-\theta_{K',n}) &= \mathbb{E}[(K-K')(\theta_{K,n}-\theta_{K',n})] - \mathbb{E}[K-K']\mathbb{E}[\theta_{K,n}-\theta_{K',n}]\\ &= \sum_{0\le k,k'\le c}\pi_k\pi_{k'}(k-k')(\theta_{k,n}-\theta_{k',n}). \quad (63) \end{aligned}$$

Since $\theta_{k,n}$ is increasing with k, the summands in the last equation are all non negative. We can also lower bound this sum by only keeping the terms $|k-k'| = c$. This shows that

$$0 \le \delta_{c,n}cp^{c-1}(1-p)^{c-1} \le P'(p) = \frac{1}{p(1-p)}\mathsf{Cov}(K,\theta_{K,n}). \quad (64)$$

An upper bound is given by noting that $\sum_{k=0}^{c}\pi_k\theta_{k,n}k \le cp\theta_{c,n}$ and $\sum_{k=0}^{c}\pi_k\theta_{k,n} \ge \theta_{0,n}$, s.t. $p(1-p)P'(p) \le \delta_{c,n}cp$. This proves that $P'(p) = \Theta(\delta_{c,n})$.

Since $P'(p) \ge 0$, we have

$$\theta_{0,n} \le P_0(p) \le P(p) \le P_1(p) \le \theta_{c,n}. \quad (65)$$

These five probabilities converge to $\bar{\theta}_0$ as $n \to \infty$.

C.2.1 When $\bar{\theta}_0 \in]0,1[$

For $\epsilon > 0$, there exists n large enough s.t. $\max(|\eta_n|, \delta_{c,n}+\eta_n) = \epsilon$. The Taylor series of (9) leads to:

$$\begin{aligned} I_S(p) &= -\frac{1}{2}h''(\bar{\theta}_0)\frac{p(1-p)}{c^2}P'(p)^2 + o(\epsilon^2) && (66)\\ &= -\frac{1}{2}h''(\bar{\theta}_0)\frac{1}{c^2p(1-p)}\mathsf{Cov}(K,\theta_{K,n})^2 + o(\epsilon^2). && (67) \end{aligned}$$

Now, $\delta_{c,n} = \delta_{c,n} + \eta_n - \eta_n$ s.t. $\epsilon \le \delta_{c,n} \le 2\epsilon$ and $P'(p) = \Theta(\epsilon)$. This allows to replace $o(\epsilon^2)$ by $o(P'(p)^2)$ or $o(\mathsf{Cov}(K,\theta_{K,n})^2)$ in the equations above.

The upper bound on $P'(p)$ is used to bound the mutual information:

$$I_S(p) \le -\frac{1}{2}h''(\bar{\theta}_0)\frac{p}{1-p}\delta_{c,n}^2 + o(\delta_{c,n}^2). \quad (68)$$

C.2.2 When $\bar{\theta}_0 \in \{0, 1\}$

We have:

$$P_1(p) \geq p^{c-1}\theta_{c,n}\ ,\ P_0(p) \geq 0. \tag{69}$$

Following the same rationale as in Sect. B.2.2, we get:

$$\begin{aligned} I_S(p) &\leq -\theta_{c,n}(1-\pi_0)\log_2(\theta_{c,n}(1-\pi_0)) + \pi_c\theta_{c,n}\log_2(p^{c-1}\theta_{c,n}) + o(\theta_{c,n}) \\ &\leq (-\theta_{c,n}\log_2\theta_{c,n})(1-\pi_0-\pi_c) \\ &\quad +\theta_{c,n}(-(1-\pi_0)\log_2(1-\pi_0) + (c-1)p^c\log_2 p) + o(\theta_{c,n}). \end{aligned}$$

References

[1] G. K. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Transactions on Information Theory*, 58(3):1880–1901, March 2012.

[2] Sheng Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi. Grotesque: Noisy group testing (quick and efficient). In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 1234–1241, Oct 2013.

[3] Chun Lam Chan, Pak Hou Che, Sidharth Jaggi, and Venkatesh Saligrama. Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. *CoRR*, abs/1107.4540, 2011.

[4] Mahdi Cheraghchi. Improved constructions for non-adaptive threshold group testing. *Algorithmica*, 67(3):384–417, 2013.

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

[6] A. C. Gilbert, B. Hemenway, A. Rudra, M. J. Strauss, and M. Wootters. Recovering simple signals. In *Information Theory and Applications Workshop (ITA), 2012*, pages 382–391, Feb 2012.

[7] Piotr Indyk, Hung Q. Ngo, and Atri Rudra. *Efficiently Decodable Non-adaptive Group Testing*, chapter 91, pages 1126–1142.

[8] E Knill, A Schliep, and D C Torney. Interpretation of pooling experiments using the Markov chain Monte Carlo method. *J. Comput. Biol.*, 3(3):395–406, 1996.

[9] T. Laarhoven. Asymptotics of fingerprinting and group testing: Tight bounds from channel capacities. *Information Forensics and Security, IEEE Transactions on*, 10(9):1967–1980, Sept 2015.

[10] Kangwook Lee, Ramtin Pedarsani, and Kannan Ramchandran. SAFFRON: A fast, efficient, and robust framework for group testing based on sparse-graph codes. *CoRR*, abs/1508.04485, 2015.

[11] Peter Meerwald and Teddy Furon. Group testing meets traitor tracing. In *ICASSP*, Prague, Czech Republic, May 2011. IEEE.

[12] Hung Q. Ngo and Ding-Zhu Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. In *Discrete mathematical problems with medical applications*, volume 55 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 171–182. Amer. Math. Soc., 2000.

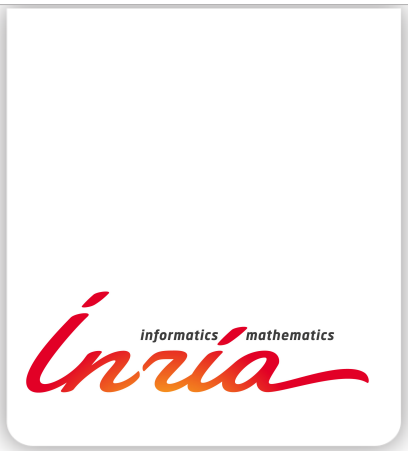
[13] Jonathan Scarlett and Volkan Cevher. Converse bounds for noisy group testing with arbitrary measurement matrices. *CoRR*, abs/1602.00875, 2016.

[14] Jonathan Scarlett and Volkan Cevher. Phase transitions in group testing. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, pages 40–53. SIAM, 2016.

[15] D. Sejdinovic and O. Johnson. Note on noisy group testing: asymptotic bounds and belief propagation reconstruction. In *Proc. 48th Allerton Conf. on Commun., Control and Computing*, Monticello, IL, USA, October 2010. arXiv:1010.2441v1.

[16] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math.*, 1(1):133–142, 1951.

[17] Y. Zhou, U. Porwal, C. Zhang, H. Q Ngo, L. Nguyen, C. Ré, and V. Govindaraju. Parallel feature selection inspired by group testing. In *NIPS*, 2014.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399