Optimal group based testing strategy for detecting infected individuals: comparison of algorithms

Viktor Skorniakov[1], Remigijus Leipus[1]*, Gediminas Juzeliūnas[2], Kęstutis Staliūnas[3,4,5]

[1]Institute of Applied Mathematics, Faculty of Mathematics and Informatics, Vilnius University, Naugarduko 24, Vilnius LT-03225, Lithuania

[2]Institute of Theoretical Physics and Astronomy, Vilnius University, Saulėtekio 3, Vilnius LT-10257, Lithuania,

[3]Laser Reseach Center, Faculty of Physics, Vilnius University, Saulėtekio 3, Vilnius LT-10222, Lithuania

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain,

[5]Departamento de Física, Universitat Politècnica de Catalunya, Campus Diagonal Nord, Edifici B5. C. Jordi Girona, 1-3, Barcelona 08034, Spain

June 29, 2020

Abstract

We investigate group based testing strategy targeted to identify infected patients by making use of a medical test which equally well applies to single and pooled samples. We demonstrate that, under assumed setting, quick sort grounded testing algorithm allows to reduce average costs, and the reduction is very significant when the infection percentage is low. Although the basic idea of test sampling is known, our major novelty is the rigorous treatment of the model. Another interesting insight following rigorous analysis is that an average number of tests per one individual scales like entropy of the prevalence of infection. One more reason for the paper is the context: taking into account the current situation with the coronavirus, dissemination of renowned ideas and the optimisation of algorithms can be of a great importance and of economical benefit.

Keywords: Group testing; Quick Sort algorithm; COVID-19.

1 Introduction

The group testing is devoted to the detection of unknown subset of defective items in a set of objects using the tests in the most efficient way. It turns out that quite often individual testing can be efficiently replaced by testing the groups. In many settings (and, particularly, in a medical one), adoption of such strategy results in the significant time and costs savings and is, therefore, a desirable step.

The concept of group testing was originally introduced by [1] in relation with efficient mass blood testing and later found important applications in molecular biology, quality control, computer science and other fields. In particular, the situation with the current COVID-19 epidemic raises the need of the quick and cheep PCR testing for SARS-COV-2 virus in order to find a rapid exit from the lockdown strategy in many countries. In this paper, we investigate group based testing strategy aimed to identify infected patients by making use of a medical test which equally well applies to single and pooled samples. We show that quick sort grounded (halving) testing algorithm allows to reduce average costs, and the reduction is very significant when the infection percentage is low.

The main objective of the paper is to recall the main schemes of group testing and compare them to each other on the grounds of the rigorous mathematical base. We hope that, taking into account the current situation with the coronavirus, dissemination of renowned ideas and the optimisation of algorithms can be of a great importance and of economical benefit.

*Author is supported by grant S-COV-20-4 from the Research Council of Lithuania

The remaining part of the paper is organized as follows. Section 2 provides the description of two classical schemes of testing together with preliminaries. In Section 3, we consider the main algorithm of the paper (**Scheme C**) together with comparison to the other two schemes. In Section 4, we give an accompanying discussion of the assumptions of our setup and the related literature. Appendices A and B contain mathematical derivations and tables.

2 Preliminaries

Consider the following setup. Assume that the prevalence of some disease (the fraction of the infected individuals) is equal to $p \in (0,1)$ in an infinite (or large enough) population. A cohort, spanning N independent individuals, has to be tested. For this, samples are collected from each individual. The test to be applied performs equally well for individual and for pooled samples: a situation might occur, e.g., when the test indicates the presence of the infection in the blood sample, and there is no difference whether the latter is obtained from a single individual or from a pooled cohort of samples.

For the situation described, physicians can choose different testing strategies. Let us assume that the following are two possible choices.

Scheme A. Test each patient's sample.

Scheme B. Conduct testing of the pooled sample. Test each member of the cohort only in case of detected infection in the pooled sample.

Although it is not obvious at the first glance, the second choice is more efficient. To give a rigorous justification, let us formally define the underlying model corresponding to **Scheme B**.

Consider the sample of N individuals. Put $X_i = 1$, provided the test of i-th individual is positive, and $X_i = 0$, otherwise. Let $S = S_N = X_1 + \cdots + X_N$ be the total number of infected individuals in the sample and let $T = T_N$ be the total number of tests applied to the cohort: the test is applied once if the result is negative, and it is further applied to each of N individuals otherwise, i.e.

$$T \;=\; 1 + N \cdot \mathbf{1}\{S > 0\}.$$

Assume that $X_1, \ldots, X_N$ are independent identically distributed (i.i.d.) random variables, each having Bernoulli distribution $\mathrm{Be}(p)$. Then S has Binomial distribution $\mathrm{Bin}(N,p)$. Therefore, an average number of tests per cohort is

$$\mathrm{E}\,T \;=\; 1 + N\,\mathrm{P}(S>0) \;= 1 + N(1-q^N), \tag{2.1}$$

where $q := 1-p$. An average number of tests per individual, say $t = t(N)$, is

$$t(N) \;=\; \frac{\mathrm{E}\,T}{N} \;=\; \frac{1}{N} + 1 - q^N. \tag{2.2}$$

Consider a function $t : (0,\infty) \mapsto (0,\infty)$ given in (2.2). By equating its derivative to 0, we see that stationary points solve equation

$$\frac{1}{N^2} \;=\; -q^N \ln q \;\text{ or, equivalently, }\; N \;=\; q^{-N/2}\Big(\ln\frac{1}{q}\Big)^{-1/2}, \tag{2.3}$$

which is a fixed point equation for $g(N) = q^{-N/2}\big(\ln\frac{1}{q}\big)^{-1/2}$, $N \in (0,\infty)$, and, hence, can be easily solved iteratively. It is further not difficult to prove that, for p in the region enclosing $(0, 0.2)$, there exists a unique solution $N_p > 0$ of (2.3) which is a minimizer of $t(N)$ (see Proposition A.1 in Appendix B). Then, turning back to economic/biomedical interpretation, we conclude that, having a cohort of $\lfloor N_p \rfloor$ (here and in the sequel, $\lfloor y \rfloor$ stands for an integer part of $y \in \mathbb{R}$) individuals, **Scheme B** results in a lowest average number of tests per person which is possible when applying scheme of this type for a population having prevalence p. **Scheme A**, in contrast, always has a constant number of tests 1 per person. Therefore, the average (absolute) gain attained by applying **Scheme B** versus **Scheme A** is given by the difference

$$G_p \;=\; 1 - t(N_p) \;=\; q^{N_p} - \frac{1}{N_p}. \tag{2.4}$$

Right panel in the Figure 1 shows the graph of $p \mapsto 100 G_p$, $p \in (0, 0.2)$, which is an average gain measured by the number of tests saved per 100 individuals. The corresponding values are provided in Table 1 (see Appendix

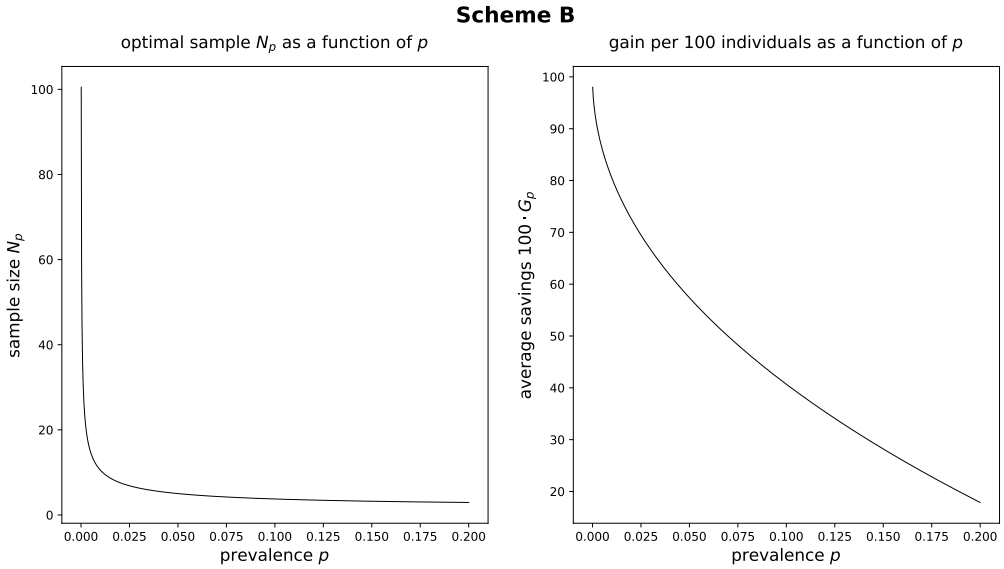

Figure 1: Performance of **Scheme B**.

B). An accompanying graph of $p \mapsto N_p$ (see the left panel of the Figure 1) demonstrates dependence of optimal sample size on p. To obtain a fast numerical evidence, assume that N is bounded away from zero and $pN \to 0$. Then from (2.3) it follows that optimal sample size satisfies

$$N \;\sim\; \frac{1}{\sqrt{p}} \quad \text{and} \quad t(N) = \frac{1}{N} + 1 - (1-p)^N \;\sim\; \frac{1}{N} + pN \;\sim\; 2\sqrt{p}.$$

Hence, assuming that p is small enough for $pN \approx 0$ to hold, the above implies that

$$G_p \;\approx\; 1 - 2\sqrt{p}.$$

For example, if $p = 0.01$, then we have $G_p \approx 0.8$, i.e. an approximate average gain is 80% or so.

3 Main algorithm

In what follows, we retain all the notions and assumptions introduced previously. For the sake of convenience, we additionally assume that the size of the cohort (to be tested) is of the form $N = 2^n$, $n \in \mathbb{N}$. Keeping in mind all the said, below is the recommended algorithm for a group based testing when p is low.

Scheme C

Step 1. Test pooled sample of the whole cohort. Proceed to *Step 2.*

Step 2. If the test is positive, proceed to *Step 3*, otherwise, finish testing cohort.

Step 3. Divide the cohort into two parts consisting of the first and second halves respectively. Apply the whole algorithm to the two obtained parts recursively.

The main features of the above testing algorithm are summarized in Proposition 3.1 below.

Proposition 3.1. *Assume the testing* ***Scheme C****. Then*

(i) *an average number of tests per person is given by*

$$t(N) \;=\; \frac{1}{N} + 2\sum_{k=1}^{\log_2 N} \frac{1-q^{2^k}}{2^k} \;=\; \frac{1}{N} + 2\log_2 N \int_0^1 \Big(\int_q^1 x^{2^{1+\lfloor v \log_2 N \rfloor} - 1} \mathrm{d}x \Big) \mathrm{d}v; \tag{3.1}$$

(ii) *an average number of tests per person in the case of an infinitely large cohort is*

$$t(\infty) \;=\; \lim_{N\to\infty} t(N) \;=\; 2\sum_{k=1}^{\infty} \frac{1-q^{2^k}}{2^k};$$

(iii) *for a fixed* $p \in (0,1)$, *function* $t : \mathbb{N} \mapsto (0,\infty)$ *admits at most two minimizers* N_p*: the value* $N = N_p$ *corresponding to optimal sample size is either* $\lfloor \frac{1}{2\log_2(1/q)} \rfloor$ *or* $\lfloor \frac{1}{2\log_2(1/q)} \rfloor + 1$.

To shed some light on the described testing scheme, we provide several comments on the performance of **Scheme C** vs. **Scheme A** and **Scheme B**.

Consider first the limit in (ii) of Proposition 3.1. Obviously,

$$t(\infty) \quad \leqslant \quad 2\Big(1 - \frac{q^2}{2} - \frac{q^4}{4}\Big).$$

Hence, for $q \approx 1$ (or alternatively $p \approx 0$), $t(\infty) < 1$. The latter means that, when the prevalence is low, the proposed scheme always outperforms common sequential **Scheme A**. To gain a quick quantitative insight, assume that p is small enough for $pN \approx 0$ to hold. Then turning to (iii) and taking a 'continuous' (undiscretized) version of N_p equal to $\frac{\ln 2}{2\ln(1/q)}$ yields relationships (see Remark A.1, eq. (A.3))

$$N_p \quad \approx \quad \frac{\ln 2}{2p} \quad \text{and} \quad t(N_p) \quad \approx \quad \frac{2p}{\ln 2} + 2p\log_2\Big(\frac{\ln 2}{2p}\Big) \quad \approx \quad -2p\log_2 p. \tag{3.2}$$

Therefore, an approximation to an average gain $G_p = 1 - t(N_p)$ is $1 + 2p\log_2 p$. Taking, e.g., $p = 0.01$ results in $G_{0.01} \approx 0.867$. Considering analogous example given in Section 2 for **Scheme B**, we see that the gain has an increase close to 7%. In fact, this does not surprise (for a visual comparison of **Schemes B** and **C** on the linear and the log-log scale, see Figure 2, and, for the numerical one, Table 1 and Table 2 in Appendix B), since for **Scheme B** we had $G_p \approx 1 - 2\sqrt{p}$ and $p\log_2 p/\sqrt{p} \to 0$ as $p \to 0$. Equality (3.2), however, exhibits some magic flavor. To see this, note that, for $p \approx 0$, entropy I_p of $X \sim \mathrm{Be}(p)$ is asymptotically equivalent to $p\log_2 p$ since

$$I_p \quad = \quad p\log_2 p + (1-p)\log_2(1-p) = p\log_2 p - p(1-p) + o(p) = p\log_2 p(1 + o(1)).$$

Consequently, (3.2) means that an average number of tests per one individual scales like entropy of the prevalence of the infection. Keeping in a view the above relationship, it is not surprising that the significant number of works ([2], [3], [4]) have approached the testing problem from the Information Theory perspective.

Another connection with the Information Theory is due to the Quick Sort (QS) algorithm utilized in the **Scheme C**. It is well known that QS yields the best (up to the constant multiple) possible average performance among comparison based algorithms: to sort an array having N non-constant (i.e., random) items, the smallest average number of comparisons is of the order $N\ln N$ [5], and, in fact, all "randomize and conquer" type algorithms (with QS being one among the rest) have expected time asymptotically equivalent to QS, which randomly splits sorted set into two equal subsets [6]. Our formula (3.1) is just a confirmation of this (well known fact). To see this, note that, in the context of sorting task, (3.1) presents an average number of comparisons per item. Though the order is correct, we are inclined to think that the multiplier $\int_0^1 \big(\int_q^1 x^{2^{1+\lfloor v\log_2 N\rfloor}-1}\mathrm{d}x\big)\mathrm{d}v$ can be improved by making use of QS modification (or another comparison-based algorithm) designed to sort items with a small number of possible values (in our case, there are just two: "sick" and "healthy"). On the other hand, we think that the order is optimal since though there are algorithms which can beat the order of QS when sorting integers, e.g., [7], [8], they operate under different, i.e., noncomparisons-based, mode. In our case, however, comparison is predefined by the setting of the problem at hand: we assume that biomedical tests can only be carried by making use of comparison.

4 Discussion

Assumptions. At a first glance, independence assumption seems too restrictive. That is, more natural would be to assume that individuals forming the tested cohort are related. However, an application of the proposed procedure is most effective when the prevalence is low ($p \approx 0$). In such case, under classical assumption $pN \to \lambda > 0$, the number of infected individuals S_N can be well approximated by the Poisson distribution $\mathrm{Pois}(pN)$, and the approximation remains quite accurate irrespectively of the nature of the dependence exhibited by summands (see [9], [10], [11] and references therein for results of this kind with possible extensions of the classical setting).

Another important assumption is that of the constant prevalence p. When the testing is applied in the vibrant pandemic environment, it would be more natural to assume the dependence on time: $p = p(t)$. This, however, requires switching to a more complex process theory model, and it is difficult to say what is more reasonable: to assume validness of our simple setting for a short period of time, or to look for a more complicated yet more realistic model requiring longer and elaborated analysis and suitable data.

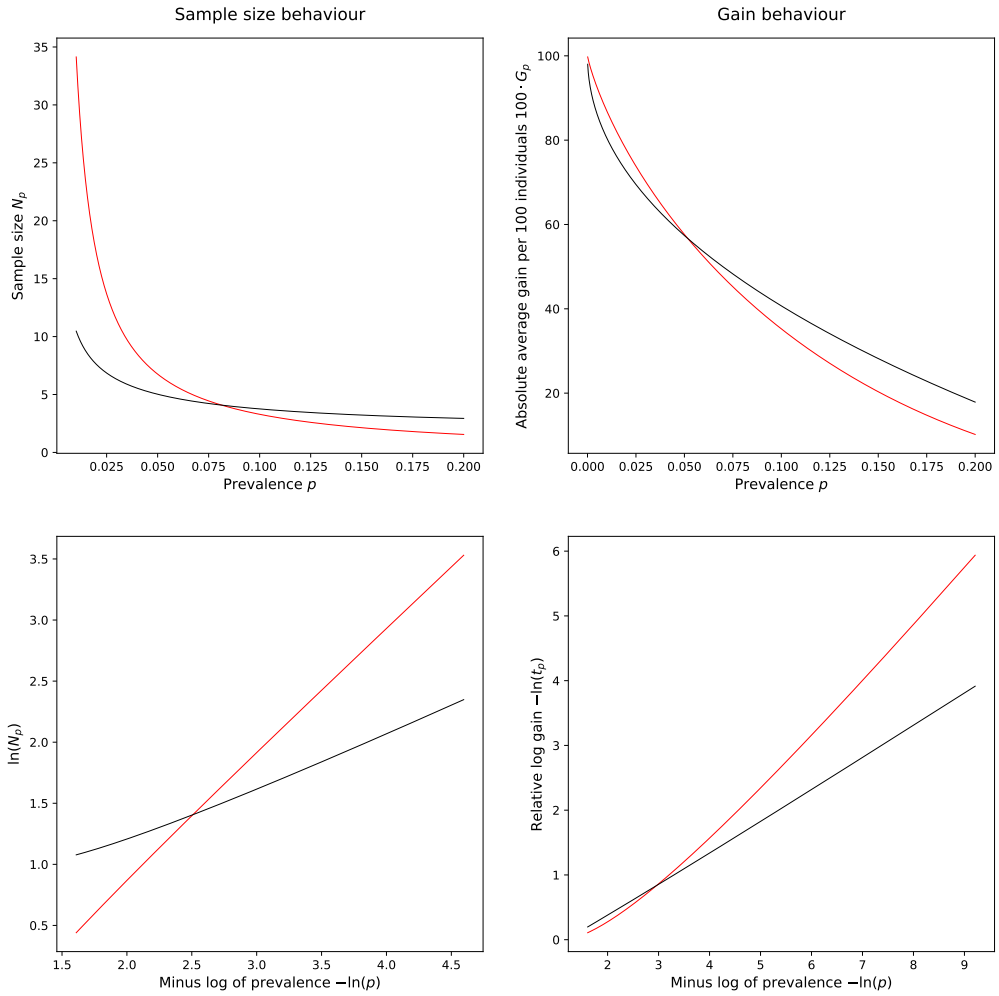

Figure 2: **Scheme C** (red) vs. **Scheme B** (black) on the linear and the log-log scale.

Related works. The ideas underlying the testing problem covered in our paper are far from being new. For example, the **Scheme B** dates back at least to 1943 (see [1]) and then reappears with modifications and further extensions not only in biomedical context but also in the engineering sciences as well: [12], [13], [14]. The **Scheme C** is currently afresh discussed by Gollier and Gossner [15], Mentus et al. [16] and Shani-Narkiss et al. [17]. For an older reference, discussing the case of nonhomogeneous population (i.e., the one in which the probability of being infected p may vary across individuals) and containing quite a large body of applied literature on halving algorithm corresponding to the **Scheme C**, we refer to [18]. Our major input is that, in contrast to the mentioned references discussing the procedure and providing instructions suitable for the practical application of **Scheme C** with a brief theoretical background, we estimate its properties in a rigorous fashion. To our best knowledge, the reference [19] is the only work close to ours both in nature of investigations and results. However, in that paper, the authors focus on the treatment of an asymptotic regime of **Scheme C** when $p \to 0$.

To finish our short discussion, it is also worth to mention that there is a large body of papers of applied nature, prevalent in medical, engineering and other literature and devoted to pooling in various settings (see, e.g., [20], [21] among others). Writing the paper, we aimed to recall the ideas of group testing, highlight directions for improvements, and promote dissemination which seems to be of most importance in the present context forcing a burst of papers devoted to similar problems (see, e.g., [22], [23], [24], [25], [26], [27], [28]). Finally,

note that some countries have already successfully applied pooling methodology for testing of the SARS-CoV-2 virus.[1]

A Appendix. Technical details

PROOF OF PROPOSITION 3.1. (i) For $1 \le i \le j \le N = 2^n$, let $M_{ij} = \{X_i, X_{i+1}, \dots, X_j\}$ and let $S(i,j) = \sum_{k=i}^{j} X_k$ be the number of infected individuals in the cohort M_{ij}. Denote $T(i,j) =$ *the total number of tests applied to the cohort* M_{ij} *after the initial pooled test.* By the description of the testing **Scheme C**,

$$
\begin{aligned}
T &= 1 + T(1,N) = 1 + \mathbf{1}\{S(1,N)>0\}(1+T(1,N/2)) + \mathbf{1}\{S(1,N)>0\}(1+T(N/2+1,N)) \\
&= 1 + 2\cdot\mathbf{1}\{S(1,N)>0\} + \mathbf{1}\{S(1,N)>0\}\big(T(1,N/2)+T(N/2+1,N)\big) \\
&= 1 + 2\cdot\mathbf{1}\{S(1,2^n)>0\} + \mathbf{1}\{S(1,N)>0\}\big(2\cdot\mathbf{1}\{S(1,N/2)>0\} + \mathbf{1}\{S(1,N/2)>0\}\big(T(1,N/4)+T(N/4+1,N/2)\big) \\
&\quad +2\cdot\mathbf{1}\{S(N/2+1,N)>0\} + \mathbf{1}\{S(N/2+1,N)>0\}\big(T(N/2+1,N/2+N/4)+T(N/2+N/4+1,N)\big)\big) \\
&= 1 + 2\cdot\mathbf{1}\{S(1,N)>0\} + 2(\mathbf{1}\{S(1,N/2)>0\} + \mathbf{1}\{S(N/2+1,N)>0\}) \\
&\quad + \mathbf{1}\{S(1,N/2)>0\}\big(T(1,N/4)+T(N/4+1,N/2)\big) \\
&\quad + \mathbf{1}\{S(N/2+1,N)>0\}\big(T(N/2+1,N/2+N/4)+T(N/2+N/4+1,N)\big) \\
&= \dots = 1 + 2\cdot\mathbf{1}\{S(1,N)>0\} + 2\big(\mathbf{1}\{S(1,N/2)>0\} + \mathbf{1}\{S(N/2+1,N)>0\}\big) \\
&\quad + 2\big(\mathbf{1}\{S(1,N/4)>0\} + \mathbf{1}\{S(N/4+1,N/2)>0\} + \mathbf{1}\{S(N/2+1,N/2+N/4)>0\} \\
&\quad + \mathbf{1}\{S(N/2+N/4+1,N)>0\}\big) + \dots + 2\big(\mathbf{1}\{S(1,2)>0\} + \mathbf{1}\{S(3,4)>0\} + \dots + \mathbf{1}\{S(N-1,N)>0\}\big).
\end{aligned}
$$

Taking expectations yields

$$
\mathrm{E}\,T = 1 + 2\sum_{j=0}^{n-1} 2^j\,\mathrm{P}\{S(1,2^{n-j})>0\} = 1 + 2\sum_{j=0}^{n-1} 2^j(1-q^{2^{n-j}}) = 1 + 2\cdot 2^n \sum_{k=1}^{n} \frac{1-q^{2^k}}{2^k}, \tag{A.1}
$$

hence, the first equality in (i). For the second one, take the last sum above and continue as follows:

$$
\begin{aligned}
\sum_{k=1}^{n} \frac{1-q^{2^k}}{2^k} &= \sum_{k=1}^{n}\int_q^1 x^{2^k-1}\mathrm{d}x = \sum_{k=1}^{n}\int_k^{k+1}\Big(\int_q^1 x^{2^{\lfloor y\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}y = \int_1^{n+1}\Big(\int_q^1 x^{2^{\lfloor y\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}y \\
&= n\int_0^1\Big(\int_q^1 x^{2^{1+\lfloor v n\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}v = \log_2 N\int_0^1\Big(\int_q^1 x^{2^{1+\lfloor v\log_2 N\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}v.
\end{aligned}
$$

(ii) By (A.1),

$$
t(N) = \frac{\mathrm{E}\,T}{N} = \frac{1}{N} + 2\sum_{k=1}^{\log_2 N}\frac{1-q^{2^k}}{2^k} \to 2\sum_{k=1}^{\infty}\frac{1-q^{2^k}}{2^k} \quad \text{as} \quad N\to\infty. \tag{A.2}
$$

(iii) Since $N = N(n) = 2^n$, by the second equality in (A.2),

$$
\Delta_n := t(N(n+1)) - t(N(n)) = \frac{1}{2N} - \frac{1}{N} + 2\frac{1-q^{2N}}{2N} = \frac{1-2q^{2N}}{2N}.
$$

Clearly, $q^{2N}\downarrow 0$ as $N\to\infty$. Therefore, there exist no more than two $N_p\in\mathbb{N}$ such that $\forall N\le N_p\ \Delta_n\le 0$, $\forall N\ge N_p\ \Delta_n\ge 0$, and $t(N_p)$ attains its minimal value at N_p. To find N_p, we first solve $1-2q^{2N}=0$ with respect to N, and then choose from the two nearest integers (i.e., $\lfloor N\rfloor, \lfloor N\rfloor+1$) the one which minimizes t_N. □

REMARK A.1. Note that, if $N\geqslant 1$ and $pN\to 0$, then for $t(N)$ in (3.1) it holds

$$
t(N) = \frac{1}{N} + 2p\log_2 N\Big(1+O\Big(\frac{pN}{\log_2 N}\Big)\Big). \tag{A.3}
$$

To see this, it suffices to use the following bounds

$$
1-pn \leqslant (1-p)^n \leqslant 1-pn+\frac{n(n-1)}{2}p^2, \quad n\geqslant 1.
$$

□

[1] According to Wikipedia, "In Israel, researchers at Technion and Rambam Hospital developed and tested a method for testing samples from 64 patients simultaneously, by pooling the samples and only testing further if the combined sample is found to be positive. Pool testing was then adopted in Israel, Germany, South Korea, and Nebraska, and the Indian states Uttar Pradesh, West Bengal, Punjab, Chhattisgarh, and Maharashtra." [29]

The next proposition justifies the discussion on **Scheme B** presented in the introductory Section 2.

Proposition A.1. *Let $p \in A := (0, 1 - \mathrm{e}^{-4/\mathrm{e}^2})$ be fixed. Consider function $g_p(N) = g(N) = q^{-N/2}(\ln \frac{1}{q})^{-1/2}$, $N > 0$. It admits a unique fixed point N_p which minimizes $t(N)$ given by (2.2).*

Proof. Step 1 (fixed points). Define

$$v \;:=\; \frac{2}{\ln(1/\sqrt{q})}, \quad f(N) \;:=\; N - g(N) = N - \frac{\sqrt{v}}{2}\mathrm{e}^{2N/v}. \tag{A.4}$$

Then $f'(N) = 1 - \frac{\mathrm{e}^{2N/v}}{\sqrt{v}} = 0 \iff N = \frac{v \ln \sqrt{v}}{2}$. Note that $f'(N) \to -\infty$ as $N \to \infty$. Moreover, $f'(0) > 0$ since $1 - \frac{1}{\sqrt{v}} > 0 \iff p < 1 - \mathrm{e}^{-4}$ which is satisfied for any $p \in A$. Therefore, f attains maximal value at

$$N_{\max} \;:=\; \frac{v \ln \sqrt{v}}{2} \quad \text{and} \quad f_{\max} \;=\; N_{\max} - \frac{\sqrt{v}}{2}\mathrm{e}^{\frac{2}{v}N_{\max}} \;=\; \frac{v}{2}\big(\ln \sqrt{v} - 1\big) > 0 \tag{A.5}$$

since $\ln \sqrt{v} - 1 > 0 \iff p < 1 - \mathrm{e}^{-4/\mathrm{e}^2}$. On the other hand, $f(0) = -\frac{\sqrt{v}}{2} < 0$ and $f(N) \xrightarrow[N\to\infty]{} -\infty$. Consequently, f has two zeroes: $N_p \in (0, N_{\max})$ and $\widetilde{N}_p \in (N_{\max}, \infty)$. The latter means that g has two fixed points.

Step 2 (minimizer). In this step, we show that N_p from Step 1 is the minimizer for $t(N)$ given in (2.2). By (2.3),

$$t''(N_p) \;=\; \frac{2}{N_p^3} - q^{N_p}(\ln q)^2 \;=\; -q^{N_p} \ln q\big(\frac{2}{N_p} + \ln q\big).$$

Hence,

$$t''(N_p) > 0 \iff \frac{2}{N_p} + \ln q > 0 \iff \frac{v}{2} > N_p. \tag{A.6}$$

From Step 1 (see (A.5)) it follows that $\frac{N_{\max}}{v/2} = \ln \sqrt{v} > 1$, i.e. $v/2 \in (0, N_{\max})$. Therefore, (A.6) holds if and only if $f(v/2) > 0$. The latter reads as $\frac{v}{2} - \frac{\mathrm{e}\sqrt{v}}{2} > 0$ and is equivalent to $p < 1 - \mathrm{e}^{-4/\mathrm{e}^2}$ showing that N_p (being the critical point of t) is indeed the announced minimizer. Finally, note that the above analysis also implies that $\widetilde{N}_p$ from Step 1 is the maximizer of $t(N)$ which affirms the uniqueness of the minimizer. □

Remark A.2. One can also show that $p \mapsto N_p$ is strictly decreasing and continuous on A. However, the latter properties seem to be of less importance and we omit the details. □

B Appendix. Tables

In the tables below, the following information is provided:

- Column 'N_p' shows an optimal sample size corresponding to p ranging in an interval given in the column 'Range of p'.
- Column 'Range of $100G_p$' shows an average gain (as defined in the main body of the paper) per 100 individuals corresponding to values of p and N_p given in the two leading columns. The highest gain corresponds to the lowest p in the corresponding interval. For example, in Table 1, the first line should be interpreted as follows: for $p \in [0.1865, 0.2000]$, optimal sample size N_p is equal to 2; if $p = 0.2000$, then average gain per 100 individuals $100G_p$ is equal to 16.1782; if $p = 0.1865$, then $100G_p = 14.000$; for intermediate values of p, the value of $100G_p$ lies in $[14.0000, 16.1782]$.

Table 1: Performance of **Scheme B**

N_p	Range of p	Range of $100G_p$	N_p	Range of p	Range of $100G_p$
2	0.1865 – 0.2000	14.0000 – 16.1782	22	0.0020 – 0.0021	90.9350 – 91.1457
3	0.0855 – 0.1864	20.5225 – 43.1472	23	0.0019 – 0.0019	91.3723 – 91.3723
4	0.0506 – 0.0854	44.9721 – 56.2450	24	0.0017 – 0.0018	91.6016 – 91.8321
5	0.0336 – 0.0505	57.1747 – 64.2917	25	0.0016 – 0.0016	92.0759 – 92.0759
6	0.0239 – 0.0335	64.8434 – 69.8233	26	0.0015 – 0.0015	92.3261 – 92.3261
7	0.0179 – 0.0238	70.1977 – 73.8374	27	0.0014 – 0.0014	92.5843 – 92.5843
8	0.0140 – 0.0178	74.1163 – 76.8337	28	0.0013 – 0.0013	92.8517 – 92.8517
9	0.0112 – 0.0139	77.0523 – 79.2489	29	0.0012 – 0.0012	93.1296 – 93.1296
10	0.0091 – 0.0111	79.4383 – 81.2637	30	0.0011 – 0.0011	93.4188 – 93.4188
11	0.0076 – 0.0090	81.4428 – 82.8596	32	0.0010 – 0.0010	93.7241 – 93.7241
12	0.0065 – 0.0075	83.0288 – 84.1396	33	0.0009 – 0.0009	94.0421 – 94.0421
13	0.0055 – 0.0064	84.2998 – 85.3889	35	0.0008 – 0.0008	94.3806 – 94.3806
14	0.0048 – 0.0054	85.5569 – 86.3428	38	0.0007 – 0.0007	94.7426 – 94.7426
15	0.0042 – 0.0047	86.5106 – 87.2152	41	0.0006 – 0.0006	95.1303 – 95.1303
16	0.0037 – 0.0041	87.3879 – 87.9915	45	0.0005 – 0.0005	95.5524 – 95.5524
17	0.0033 – 0.0036	88.1708 – 88.6533	50	0.0004 – 0.0004	96.0195 – 96.0195
18	0.0030 – 0.0032	88.8385 – 89.1800	58	0.0003 – 0.0003	96.5507 – 96.5507
19	0.0027 – 0.0029	89.3683 – 89.7296	71	0.0002 – 0.0002	97.1814 – 97.1814
20	0.0024 – 0.0026	89.9265 – 90.3079	100	0.0001 – 0.0001	98.0049 – 98.0049
21	0.0022 – 0.0023	90.5176 – 90.7183			

Table 2: Performance of **Scheme C**

N_p	Range of p	Range of $100G_p$	N_p	Range of p	Range of $100G_p$
1	0.1592 – 0.2000	10.2068 – 18.1573	59	0.0058 – 0.0058	91.4813 – 91.4813
2	0.1092 – 0.1591	18.1801 – 32.0493	60	0.0057 – 0.0057	91.5996 – 91.5996
3	0.0830 – 0.1091	32.0828 – 41.7999	61	0.0056 – 0.0056	91.7184 – 91.7184
4	0.0670 – 0.0829	41.8413 – 48.8858	62	0.0055 – 0.0055	91.8377 – 91.8377
5	0.0562 – 0.0669	48.9333 – 54.2777	64	0.0054 – 0.0054	91.9575 – 91.9575
6	0.0484 – 0.0561	54.3303 – 58.5384	65	0.0053 – 0.0053	92.0779 – 92.0779
7	0.0424 – 0.0483	58.5953 – 62.0600	66	0.0052 – 0.0052	92.1987 – 92.1987
8	0.0378 – 0.0423	62.1207 – 64.9241	67	0.0051 – 0.0051	92.3202 – 92.3202
9	0.0341 – 0.0377	64.9881 – 67.3443	69	0.0050 – 0.0050	92.4422 – 92.4422
10	0.0311 – 0.0340	67.4112 – 69.3911	70	0.0049 – 0.0049	92.5648 – 92.5648
11	0.0285 – 0.0310	69.4607 – 71.2322	72	0.0048 – 0.0048	92.6880 – 92.6880
12	0.0264 – 0.0284	71.3044 – 72.7691	73	0.0047 – 0.0047	92.8118 – 92.8118
13	0.0245 – 0.0263	72.8434 – 74.2009	75	0.0046 – 0.0046	92.9362 – 92.9362
14	0.0229 – 0.0244	74.2775 – 75.4396	76	0.0045 – 0.0045	93.0612 – 93.0612
15	0.0215 – 0.0228	75.5181 – 76.5498	78	0.0044 – 0.0044	93.1869 – 93.1869
16	0.0202 – 0.0214	76.6301 – 77.6042	80	0.0043 – 0.0043	93.3132 – 93.3132
17	0.0191 – 0.0201	77.6863 – 78.5153	82	0.0042 – 0.0042	93.4402 – 93.4402
18	0.0181 – 0.0190	78.5990 – 79.3593	84	0.0041 – 0.0041	93.5678 – 93.5678
19	0.0172 – 0.0180	79.4445 – 80.1325	86	0.0040 – 0.0040	93.6962 – 93.6962
20	0.0164 – 0.0171	80.2193 – 80.8312	88	0.0039 – 0.0039	93.8253 – 93.8253
21	0.0157 – 0.0163	80.9193 – 81.4518	91	0.0038 – 0.0038	93.9552 – 93.9552
22	0.0150 – 0.0156	81.5412 – 82.0814	93	0.0037 – 0.0037	94.0858 – 94.0858
23	0.0144 – 0.0149	82.1721 – 82.6285	96	0.0036 – 0.0036	94.2172 – 94.2172
24	0.0138 – 0.0143	82.7204 – 83.1829	98	0.0035 – 0.0035	94.3493 – 94.3493
25	0.0133 – 0.0137	83.2760 – 83.6506	101	0.0034 – 0.0034	94.4824 – 94.4824
26	0.0128 – 0.0132	83.7448 – 84.1237	104	0.0033 – 0.0033	94.6162 – 94.6162
27	0.0124 – 0.0127	84.2190 – 84.5063	108	0.0032 – 0.0032	94.7509 – 94.7509
28	0.0119 – 0.0123	84.6025 – 84.9897	111	0.0031 – 0.0031	94.8866 – 94.8866
29	0.0115 – 0.0118	85.0871 – 85.3808	115	0.0030 – 0.0030	95.0231 – 95.0231
30	0.0112 – 0.0114	85.4792 – 85.6767	119	0.0029 – 0.0029	95.1607 – 95.1607
31	0.0108 – 0.0111	85.7759 – 86.0750	123	0.0028 – 0.0028	95.2992 – 95.2992
32	0.0105 – 0.0107	86.1752 – 86.3764	128	0.0027 – 0.0027	95.4388 – 95.4388
33	0.0102 – 0.0104	86.4775 – 86.6804	133	0.0026 – 0.0026	95.5794 – 95.5794
34	0.0099 – 0.0101	86.7822 – 86.9868	138	0.0025 – 0.0025	95.7211 – 95.7211
35	0.0096 – 0.0098	87.0896 – 87.2959	144	0.0024 – 0.0024	95.8640 – 95.8640
36	0.0094 – 0.0095	87.3996 – 87.5035	150	0.0023 – 0.0023	96.0081 – 96.0081
37	0.0091 – 0.0093	87.6078 – 87.8172	157	0.0022 – 0.0022	96.1534 – 96.1534
38	0.0089 – 0.0090	87.9224 – 88.0279	164	0.0021 – 0.0021	96.3001 – 96.3001
39	0.0087 – 0.0088	88.1337 – 88.2398	173	0.0020 – 0.0020	96.4481 – 96.4481
40	0.0085 – 0.0086	88.3463 – 88.4532	182	0.0019 – 0.0019	96.5976 – 96.5976
41	0.0083 – 0.0084	88.5603 – 88.6678	192	0.0018 – 0.0018	96.7486 – 96.7486
42	0.0081 – 0.0082	88.7757 – 88.8839	203	0.0017 – 0.0017	96.9012 – 96.9012
43	0.0079 – 0.0080	88.9924 – 89.1014	216	0.0016 – 0.0016	97.0555 – 97.0555
44	0.0077 – 0.0078	89.2107 – 89.3203	230	0.0015 – 0.0015	97.2116 – 97.2116
45	0.0076 – 0.0076	89.4303 – 89.4303	247	0.0014 – 0.0014	97.3696 – 97.3696
46	0.0074 – 0.0075	89.5408 – 89.6515	266	0.0013 – 0.0013	97.5297 – 97.5297
47	0.0072 – 0.0073	89.7627 – 89.8743	288	0.0012 – 0.0012	97.6920 – 97.6920
48	0.0071 – 0.0071	89.9863 – 89.9863	314	0.0011 – 0.0011	97.8567 – 97.8567
49	0.0070 – 0.0070	90.0987 – 90.0987	346	0.0010 – 0.0010	98.0241 – 98.0241
50	0.0068 – 0.0069	90.2115 – 90.3247	384	0.0009 – 0.0009	98.1943 – 98.1943
51	0.0067 – 0.0067	90.4383 – 90.4383	433	0.0008 – 0.0008	98.3677 – 98.3677
52	0.0066 – 0.0066	90.5524 – 90.5524	494	0.0007 – 0.0007	98.5448 – 98.5448
53	0.0064 – 0.0065	90.6669 – 90.7819	577	0.0006 – 0.0006	98.7260 – 98.7260
54	0.0063 – 0.0063	90.8973 – 90.8973	692	0.0005 – 0.0005	98.9120 – 98.9120
55	0.0062 – 0.0062	91.0132 – 91.0132	866	0.0004 – 0.0004	99.1039 – 99.1039
56	0.0061 – 0.0061	91.1295 – 91.1295	1155	0.0003 – 0.0003	99.3030 – 99.3030
57	0.0060 – 0.0060	91.2463 – 91.2463	1732	0.0002 – 0.0002	99.5119 – 99.5119
58	0.0059 – 0.0059	91.3636 – 91.3636	3465	0.0001 – 0.0001	99.7360 – 99.7360

References

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

[2] P. Chen, L. Hsu, and M. Sobel. Entropy-based optimal group-testing procedures. *Probability in the Engineering and Informational Sciences*, 1:497–509, 1987.

[3] L. Hsu. New procedures for group-testing based on the Huffman lower bound and Shannon entropy criteria. In N. Flournoy and W. F. Rosenberger, editors, *Adaptive Designs*, volume 25 of *Lecture Notes - Monograph Series*, pages 249–262. Institute of Mathematical Statistics, 1995.

[4] M. Aldridge, O. Johnson, and J. Scarlett. *Group testing: An information theory perspective*, volume 15 of *Foundations and Trends in Communications and Information Theory Series*, pages 196–392. Now Publishers, 2019.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 3rd ed. edition, 2009.

[6] B. C. Dean. A simple expected running time analysis for randomized "divide and conquer" algorithms. *Discrete Applied Mathematics*, 154(1):1–5, 2006.

[7] A. Andersson, T. Hagerup, S. Nilsson, and R. Raman. Sorting in linear time? *Journal of Computer and System Sciences*, 57(1):74–93, 1998.

[8] M. Thorup. Randomized sorting in $O(n \log\log n)$ time and linear space using addition, shift, and bit-wise Boolean operations. *Journal of Algorithms*, 42(2):205–230, 2002.

[9] L. H. Y. Chen. Poisson approximation for dependent trials. *Annals of Probability*, 3:534–545, 1975.

[10] V. Čekanavičius. *Approximation Methods in Probability Theory*. Springer, New York, 2016.

[11] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Clarendon Press, Oxford, 1992.

[12] M. Sobel and P. A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal*, 38:1179–1252, 1959.

[13] M. Sobel. Optimal group testing. In *Proceedings of the Colloquium on Information Theory*, pages 411–488, Debrecen (Hungary), 1967. Organized by the Bolyai Mathematical Society.

[14] J.-K. Lee and M. Sobel. Dorfman and $R1$-type procedures for a generalized group-testing problem. *Mathematical Biosciences*, 15:317–340, 1972.

[15] C. Gollier and O. Gossner. Group testing against Covid-19. *Covid Economics*, Issue 2:32–42, 2020.

[16] C. Mentus, M. Romeo, and C. DiPaola. Analysis and applications of adaptive group testing method for COVID-19. https://www.medrxiv.org/content/10.1101/2020.04.05.20050245v2, 2020.

[17] H. Shani-Narkiss, O. D. Gilday, N. Yayon, and I. D. Landau. Efficient and practical sample pooling for High-Throughput PCR diagnosis of COVID-19. https://www.medrxiv.org/content/10.1101/2020.04.06.20052159v2, 2020.

[18] M. S. Black, C. R. Bilder, and J. M. Tebbs. Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(2):277–290, 2012.

[19] N. Zaman and N. Pippenger. Asymptotic analysis of optimal nested group-testing procedures. *Probability in the Engineering and Informational Sciences*, 30:547–552, 2016.

[20] E. Litvak, X. M. Tu, and M. Pagano. Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, 89(426):424–434, 1994.

[21] S. May, A. Gamst, R. Haubrich, C. Benson, and D. M. Smith. Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 53(2):194–201, 2010.

[22] C. Gollier. Optimal group testing to exit the Covid confinement. Preprint, Toulouse School of Economics, 2020.

[23] E. Seifried and S. Ciesek. Pool testing of SARS-CoV-2 samples increases test capacity. https://eurekalert.org/pub_releases/2020-03/guf-pto033020.php, 2020.

[24] A. Z. Broder and R. Kumar. A note on Double Pooling Tests. arXiv:2004.01684 [cs, math, stat], 2020.

[25] I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony. Evaluation of COVID-19 RT-qPCR test in multi-sample pools. http://medrxiv.org/lookup/doi/10.1101/2020.03.26.20039438, 2020.

[26] K. R. Narayanan, A. Heidarzadeh, and R. Laxminarayan. Accelerated testing for COVID-19 using group testing. arXiv:2004.04785v1 [cs.IT], 2020.

[27] N. Sinnott-Armstrong, D. L. Klein, and B. Hickey. Evaluation of group testing for SARS-CoV-2 RNA. https://www.medrxiv.org/content/10.1101/2020.03.27.20043968v1.

[28] J. Žilinskas, A. Lančinskas, and M. R. Guarracino. Pooled testing with replication: a mass testing strategy for the COVID-19 pandemics. https://www.medrxiv.org/content/10.1101/2020.04.27.20076422v1.

[29] COVID-19 testing. https://en.wikipedia.org/wiki/COVID-19_testing.