Optimal Testing of Mixed Samples

Hadi Habibzadeh

April 2020

Abstract

In the absence of an effective vaccine or medication, the so-called "flattening the curve" is the most viable strategy to contain an outbreak. The efficacy of this strategy hinges on widespread and regular testing to detect both infected and immune individuals. The high demand for testing during an outbreak, however, puts an enormous strain on our healthcare infrastructure, which results in the extreme scarcity of test kits. This work uses the principles of information theory to mitigate this shortage. Because less than half of the tests are generally positive, the information content of each test is less than one bit. Therefore, this work borrows from existing symbol-based and context-based coding schemes to propose new testing methods for mixed samples that reduce the average number of tests. The proposed methods can increase the testing capability of the US by at least 30% (when compared with the standard method that uses one kit per sample). This increase is further boosted when the probability of the infection decreases (e.g., it increases the testing capacity by a factor of five in South Korea). This increase requires no extra infrastructure or additional cost.

1 Introduction

With the global outbreak of the COVID-19, governments and research institutes have galvanized their resources to fund and expedite innovative solutions that can help contain the pandemic. Among the proposed solutions, the so-called "*flattening the curve*" is believed to be the single most effective strategy to depress the fatalities of the disease. As of April 9th, the number of deaths in the US alone reached 16,444 [1], which although staggering, is far below the projected 100,000–240,000 figures [2]. This lowering in the number of deaths is mostly attributed to the country's success in flatting the curve of the daily number of cases.

In the absence of an effective vaccine or medication, widespread testing remains the most viable solution for flattening the curve [3]. The advantages of extensive testing are twofold. First, it leads to the quick detection of new cases. When coupled with separation and isolation, this can substantially reduce the infection rate of the disease. Second, testing for the related antibodies helps with identifying already-recovered (and thus immune) individuals who can resume their normal social activities and routines; thereby limiting the financial toll of the outbreak. Despite these advantages, the US, along with many other countries, has struggled to supply adequate testing capacity to regularly test the majority of the population [4]. Thus the country has resorted to alternative approaches such as *social-distancing* and *shelter-in-place* orders. These policies, however, are known to have a limited effect on harnessing the spread of the virus and incur a significant financial burden on the economy [5].

Two different tests are generally prescribed for COVID-19. The first method uses reverse transcription polymerase chain reaction (RT-PCR) [6] to detect the genetic material (SARS-Cov2 only has RNA) of the virus. This test can confirm the infection in its early stages. Samples of RT-PCR are generally collected from infected tissues (e.g., throat, nose, or lungs) using a conventional swab. After some processing (details of which can be found in [7]), an RT-PCR machine enriches the viral DNA in the sample to the level where it can be identified using a fluorescent dye. This process can take up to 2 days to complete; nonetheless, real-time RT-PCR machines are now available that can make a diagnosis in about three hours [7]. The second method of testing is serology that can identify the exposure of an individual to a given pathogen. This not only confirms the infection but also tests whether the individual has an active immune system against SARS-Cov2 [8]. Serology tests use blood samples.

This project uses the principles of the information theory to significantly expand the national testing capacity. The proposed techniques are based on two fundamental premises. (i) Both existing testing solutions

use samples that can be mixed and combined for multiple individuals and (ii) based on the latest US data, only a small proportion of tests are positive (19.5% as of April 11th [9]). The second statement is particularly important as it hints at a data compression opportunity. This work proposes four different testing methods that borrow from the standard symbol-based and context-based coding schemes to significantly reduce the average number of tests.

2 Analysis

Let the random binary code $X_n = (x_1, x_2, ..., x_n)$ represent a population of samples with size n, where for each $i \in \{1, 2, ..., n\}$, sample $x_i \in \{0, 1\}$. A test, $\mathcal{T}(X_n)$, returns 1 (positive) if there exists at least one $i \in \{1, 2, ..., n\}$ for which sample $x_i = 1$. Otherwise, it returns 0 (negative). Therefore, we define test $\mathcal{T}$ as:

$$\begin{aligned}\mathcal{T}(X_n) &\triangleq \mathcal{T}(x_1, x_2, ..., x_n) \\ &= \mathcal{T}(x_1) \mid \mathcal{T}(x_2) \mid ... \mid \mathcal{T}(x_n) \\ &= \begin{cases} 1 & \text{if } \exists\ i \in \{1, 2, ..., n\} \text{ s.t. } x_i = 1 \\ 0 & \text{therwise} \end{cases}\end{aligned} \tag{1}$$

where $\mid$ represents the logical *or* operator. Let $p_i = P(\mathcal{T}(x_i) = 1)$ be the probability that the test of x_i is positive. We assume $p \triangleq p_1 = p_2 = ... = p_n$. Likewise, define $q \triangleq q_i = 1 - p_i$ to be the probability of the test being negative. For an unknown sample population $X_n = (x_1, x_2, ..., x_n)$, our objective is to determine all x_i's using as few tests as possible. To this end, we propose four methods with a shared algorithmic backbone that is derived form the traditional binary search algorithm. Pick any 2-adic integer $m \leq n$ and let $|\mathcal{T}(x_1, x_2, ..., x_m)|$ be the number of tests that we need to perform to resolve $x_1, x_2, ..., x_m$ (i.e., to determine the value of each x_i). Then,

$$|\mathcal{T}(x_1, x_2, ..., x_m)| = \begin{cases} 1 & \text{if } \mathcal{T}(x_1, x_2, ..., x_m) = 0 \\ 1 & \text{if } m = 1 \\ 1 + |\mathcal{T}(x_1, x_2, ..., x_{m/2})| + |\mathcal{T}(x_{m/2+1}, x_{m/2+2}, ..., x_m)| & \text{Otherwise.} \end{cases} \tag{2}$$

Observe that $\mathcal{T}(x_1, x_2, ..., x_m) = 0$ immediately resolves all x_i's as it holds if and only if $x_1 = x_2 = ... = x_m = 0$. All four methods introduced in this work utilize the recursive formulation in Eq. (2). They only start to diverge at the last iteration of the recursion. Hence, we only study this iteration for each method.

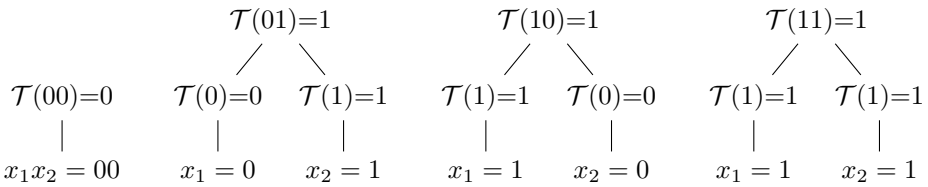
Let $Y_2 = (x_i, x_j)$ for any $i, j \in \{1, 2, ..., n\}$ such that $i \neq j$. In this case, $Y_2 \in C \triangleq \{(00), (01), (10), (11)\}$ with the probability mass function (PMF) of $P_Y = \{q^2, qp, pq, p^2\}$ (Y_3 is also defined similarly). The average information content of Y_2, denoted by $H(Y_2)$ is:

$$\begin{aligned} H(Y_2) = H(P_Y) &= -\sum_{y \in C} P_Y(y) \times \log(P_Y(y)). \\ &= -2 \times (q \log(q) + p \log(p)) \\ &= 2 \times H(p) \end{aligned} \tag{3}$$

Equation (3) represents the minimum number of tests that we need to perform to resolve Y_2. In practice, a symbol coding algorithm such as Huffman's leads to an expected number of tests between $2H(p)$ and $2H(p) + 2$.

2.1 Method I

The most expedient solution when m reaches 2 is to continue the recursion per Eq. (2). The forest shown below depicts the four possible cases in this scenario. The best case happens when both x_1 and x_2 are zero, where both can be resolved using only a single test. In other cases, at least three tests must be performed.



The expected number of tests $E|\mathcal{T}|$ in this method can be computed as:

$$\begin{aligned} E|\mathcal{T}| &= q^2 \times 1 + qp \times 3 + pq \times 3 + p^2 \times 3 \\ &= -2p^2 + 4p + 1. \end{aligned} \tag{4}$$

It is evident that $E|\mathcal{T}| \leq 2$ for $p \leq 0.29$. Therefore, in this range, Method I outperforms the existing testing method.

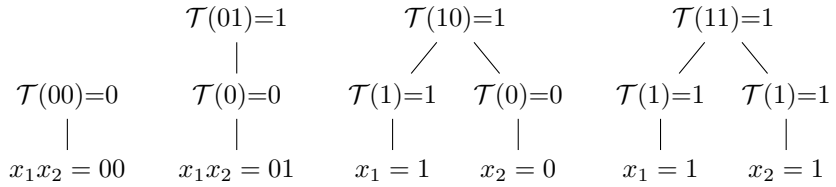
To observe the inefficiency of Method I, let $p = 0.2$. Then, the PMF of Y_2 is $\{0.64, 0.16, 0.16, 0.04\}$. Applying the Huffman coding algorithm to this distribution gives the encoded words $\{0, 10, 110, 111\}$. This code clearly shows that we can resolve the second case ($x_1x_2 = (01)$) using only two tests. Method II implements this optimization.

2.2 Method II

The previous method can optimally resolve the values of x_1 and x_2 when they are both zero. However, it is inefficient for the case when $Y_2 = (x_1x_2) = 01$. In this case, it can be observed that the value of x_1x_2 given $\mathcal{T}(x_1x_2) = 1$ and $\mathcal{T}(x_1) = 0$ is deterministic. Hence,

$$H\left(x_1x_2|\mathcal{T}(x_1x_2) = 1, \mathcal{T}(x_1) = 0\right) = 0. \tag{5}$$

Method II utilizes this property to reduce the number of tests for this case.



The expected number of tests $E|\mathcal{T}|$ in this method can then be computed as:

$$\begin{aligned} E|\mathcal{T}| &= q^2 \times 1 + qp \times 2 + pq \times 3 + p^2 \times 3 \\ &= -p^2 + 3p + 1. \end{aligned} \tag{6}$$

It is evident that $E|\mathcal{T}| \leq 2$ for $p \leq 0.38$. Therefore, in this range, Method II outperforms the existing testing method. The expected number of tests in Method II is always less than the expected number of tests in Method I.

The number of tests for each outcome of x_1x_2 is consistent with the symbol lengths in Huffman coding (at least for $p = 0.2$). Therefore, there is no room for further improvement. However, switching from symbol-based coding to context-based coding can further decrease the expected number of tests. Method III implements this improvement.

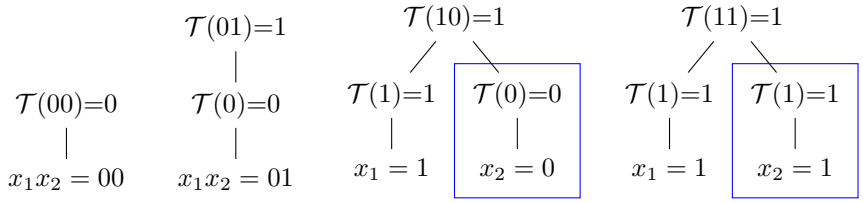
2.3 Method III

Method II requires three tests for cases when $x_1x_2 = (10)$ or $x_1x_2 = (11)$. That is, Method II first completes $\mathcal{T}(x_1)$ in the left branch and then proceeds with $\mathcal{T}(x_2)$ in the right branch (See the forest below). Therefore,

it requires, on average, exactly one test for determining x_2 given that $x_1 = 1$. Observe that due to the independence of x_1 and x_2, it can be shown that $P(x_2 = 1|x_1 = 1) = P(x_2 = 1) = p$. Therefore,

$$H\big(x_2|\mathcal{T}(x_1) = 1\big) = H(p) \ll 1 \qquad \text{for } p \ll \frac{1}{2}. \tag{7}$$

Equation (7) shows that because there are only two cases to resolve (equivalent to having only two symbols in Huffman coding), no actual compression can be achieved regardless of the disparity in probabilities. Method III mitigates this drawback by adding context (similar to Arithmetic coding). After resolving $x_1 = 1$, this method refrains from resolving x_2. Instead, it saves the unresolved value of x_2 in a *bag* and resumes the recursions until all n samples are processed. At this point, the bag, on average, contains $r = p \times n$ unresolved samples. These remaining samples can now be tested using Method II, resulting in an average number of tests of $\frac{1}{2}(-p^2 + 3p + 1)$ (which can be higher or lower than 1 depending on the value of p).



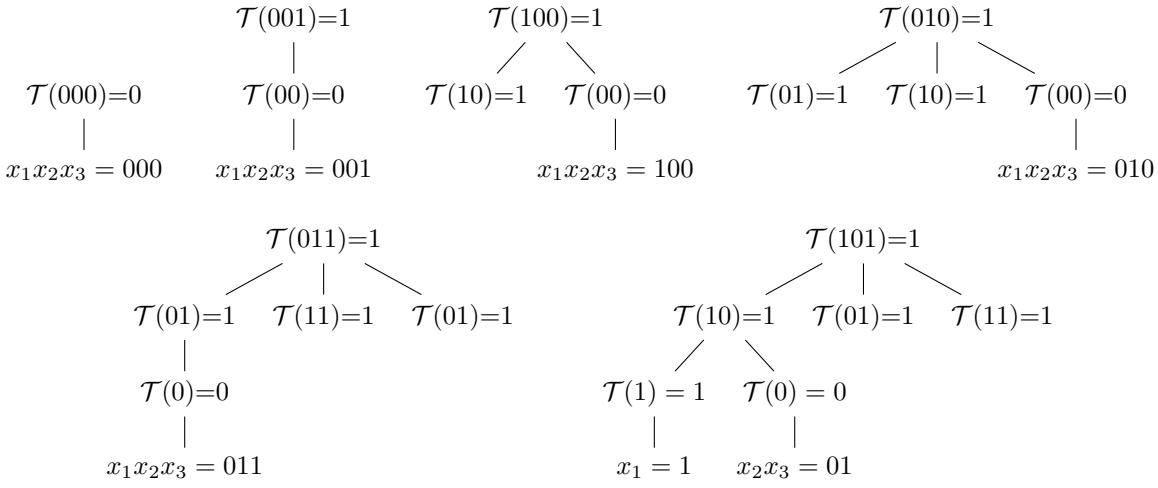
The expected number of tests $E|\mathcal{T}|$ in this method can then be computed as:

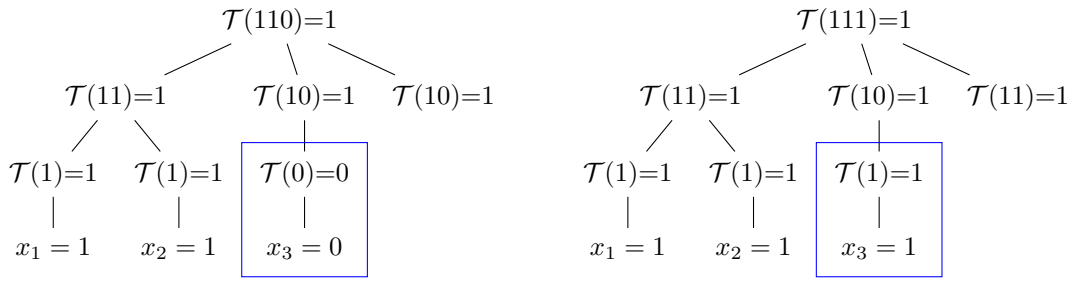
$$\begin{aligned} E|\mathcal{T}| &= q^2 \times 1 + qp \times 2 + (pq + p^2) \times \Big(2 + \frac{1}{2}(-p^2 + 3p + 1)\Big) \\ &= -\frac{1}{2}(p^3 - p^2 - 5p - 2). \end{aligned} \tag{8}$$

Solving $E|\mathcal{T}| \leq 2$ for p yields $p \leq 0.38$ implying that Method III outperforms the existing testing methods for these probabilities. Additionally, comparing $E|\mathcal{T}|$ of Method II and Method III shows that Method III requires a smaller number of tests for $p \leq 0.38$.

2.4 Method IV

Method IV extends Method III to process three samples at a time. This further reduces the expected number of tests for lower values of p. In this method, the testing trees (shown in the forest below) can have at most three levels. The first level tests only $x_1x_2x_3$. The second level tests possible pairs of samples; starting with x_1x_2 and then testing x_2x_3 and x_1x_3 if necessary. Finally, the last level tests all samples individually. This method also uses the same bagging strategy implemented in Method III.





The expected number of tests $E|\mathcal{T}|$ in this method can then be computed as:

$$E|\mathcal{T}| = q^3 + 9q^2p + 17qp^2 + 6p^3 + \frac{1}{3}p^2(q^3 + 9q^2p + 18qp^2 + 7p^3)$$
$$= \frac{1}{3}\Big(- 3p^3 + 3p^4 - 3p^3 + 7p^2 + 18p + 3\Big). \tag{9}$$

3 Numerical Results

The simulations are based on how lab technicians perform the tests in a real-world scenario. For Method I, Method II, and Method III, m must be a 2-adic number. For Method IV, m must be a multiple 3 of a 2-adic number. Therefore, the first three methods use the same population size (n), while method IV uses the first $^3/_4$ of the samples. For a large enough number of samples, this does not skew the simulation results.

Figure 1 shows the average number of tests for each method for the fixed value of $m = 2$ and the population size of $n \approx 1.5 \times 10^6$. The left plot is obtained by simulation while the right plot is based on Eq. (4), Eq. (6), Eq. (8), and Eq. (9). Figure 1 shows that Method III slightly outperforms Method II for $p \leq 0.38$. Nonetheless, both methods are almost optimal for $0.2 \leq p \leq 0.38$. The probability $p \approx 0.38$ is the switching point, where Huffman coding leads to codewords with the same size (size 2). Hence, for these probabilities, the optimal approach is to perform one test per sample. As the probability approaches zero, the gap between the optimal solution (H) and the proposed methods widens. Increasing the value of m remedies this problem, as shown by Method IV.

As the probability p approaches 0, the number of cases where $\mathcal{T}(x_1x_2...x_m) = 0$ increases. This, in turn, reduces the average number of tests for relatively larger values of m. On the other hand, increasing m also increases the number of tests for resolving the worst-case scenario. Figure 2 depicts the trade-off between m and p. This figure also serves as a provisioning tool that enables us to choose the best value of m based on our estimated probability p. For example, if $0.17 \leq p \leq 0.28$, the best performance of Method I is obtained for $m = 2$ whereas for $p > 0.38$, the best choice is to pick $m = 1$.

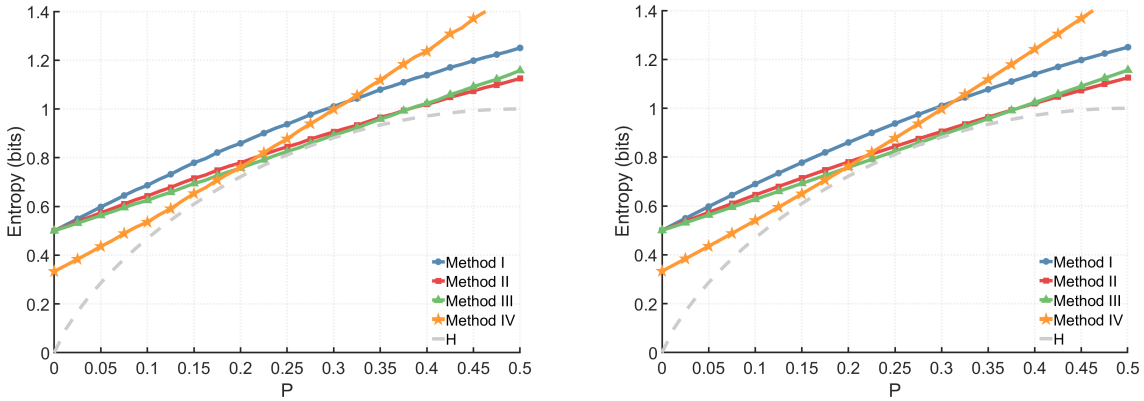


Figure 1: The average number of tests required to resolve a string with two bits ($m = 2$) for the first three methods and a string with three bits ($m = 3$) for Method IV. (left) simulated data (right) analytic results.

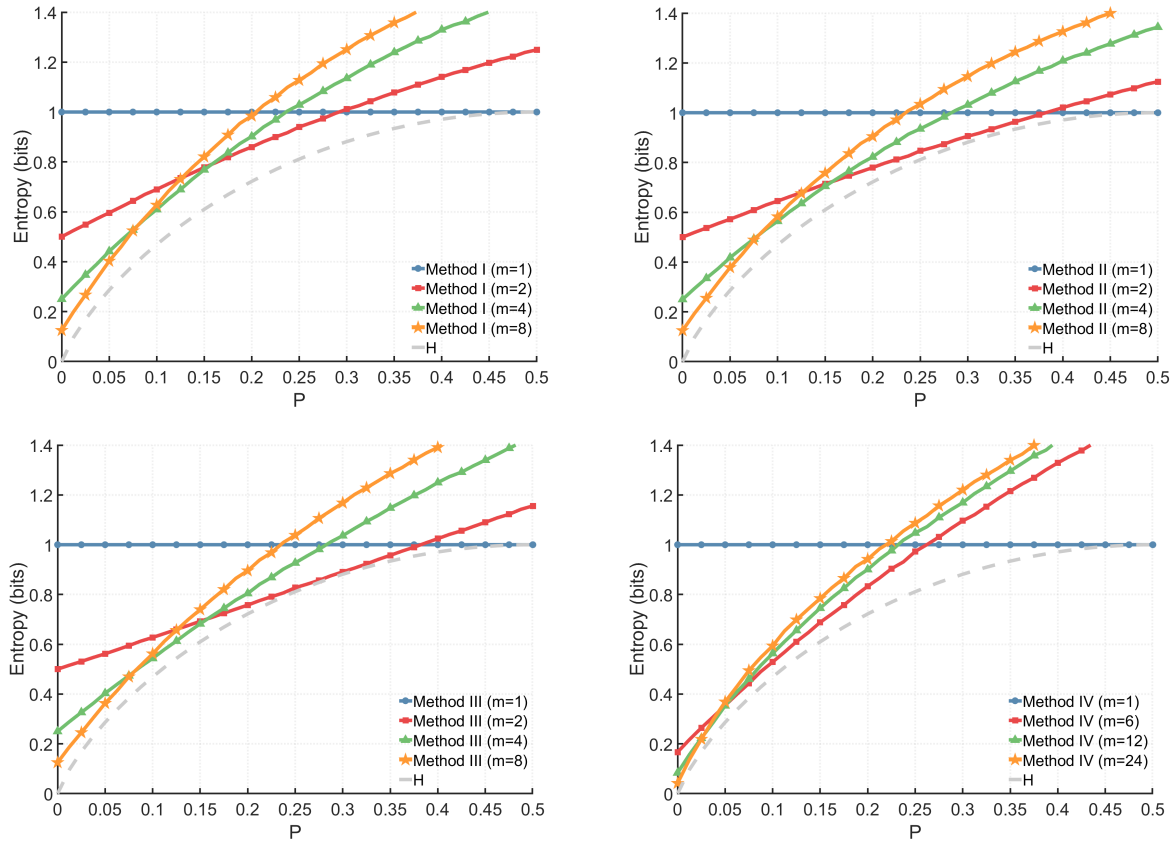


Figure 2: The effect of changing m on the average number of tests of each method. Larger values form m perform better for lower values of p and vice versa.

Using the aforementioned provisioning tool, Fig. 3 (left) shows the average number of tests for each method, when the best value of m is selected from $\{1, 2, 4, 8\}$ for the first three methods and from $\{1, 6, 12, 24\}$ for Method IV, based on the probability p. Figure 3 (right) further combines all four methods into a single curve. All tests numbers that are on the curve are achievable.

4 Discussion

This work shows that the number of required tests using mixed samples can reach length of codewords in Huffman coding. This section discusses some of the advantages and limitations of this approach.

4.1 Making a Difference

In the US, the probability of a COVID-19 test to be positive is about $p = 0.19$ [9]. Currently, one test kit is used for each sample and more than 2.6 million tests are dispersed. Figure 3 (right) shows that for this probability, we can reduce the number of tests to 0.76 tests per sample. This would enable the nation to perform more 3.4 million tests in the same period.

The probability of a positive test in the US, however, is highly inflated by the sever scarcity in test kits. Therefore, it can be confidently concluded that the population of infected people is far below 19%. For example, in South Korea (a country which has been far more successful in sourcing ample testing capability), the probability of a positive test is about $p = 0.02$ [10]. With this probability, we can use 0.21 tests for each sample, on average. This is equivalent to increasing the testing capability by a factor of 5. Additionally, the proposed methods are particularly effective for early stages of an outbreak, where the probability of a

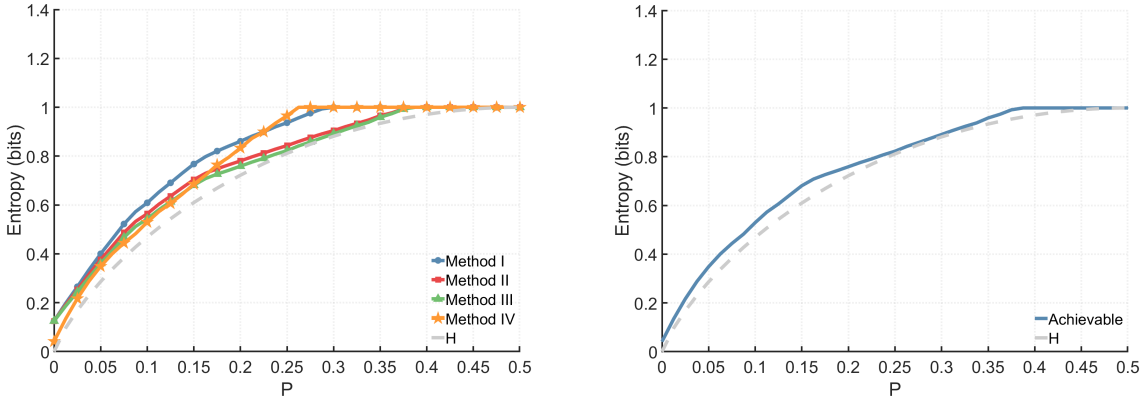


Figure 3: The average number of tests for (left) the best choice of m for each method and (right) the best choice of m and best method at each probability.

positive test is very low and there is an urgent need to test large numbers of samples.

The proposed methods are expected to be even more rewarding for immunity tests. To slacken the suspension of normal social and financial activities, immunity tests must be conducted extensively and regularly. However, as an infection affects a relatively small portion of society, chances for positive immunity tests remain fairly small. For example, out of 327.2 million people who live in the US, only 522,843 confirmed cases of the disease are recorded [9]. This amounts to a ratio of $\approx 0.16\%$. This ratio in South Korea is only $\approx 0.02\%$.

4.2 Timing Overhead

Let $\mathcal{T}(X_n)$ take w seconds to resolve X_n. If we use one test per sample, we can resolve each sample in exactly w seconds. However, if we use Method I or Method II, the expected amount of time to resolve a population with size 2 (Ew) is:

$$\begin{aligned} Ew &= w\left(q^2 + 2qp + 2pq + 2p^2\right) \\ &= w(-p^2 + 2p + 1). \end{aligned} \tag{10}$$

Similarly, for Method III,

$$\begin{aligned} Ew &= w\left(q^2 + 2qp + 2pq + 2p^2 + p(-p^2 + 2p + 1)(pq + p^2)\right) \\ &= w(-p^3 + p^2 + 3p + 1). \end{aligned} \tag{11}$$

and for Method IV:

$$\begin{aligned} Ew &= w\left(q^3 + 9q^2p + 17qp^2 + 6p^3 + p^2(q^3 + 9q^2p + 18qp^2 + 7p^3)\right) \\ &= w(-3p^5 + 3p^4 + 3p^3 + 3p^2 + 6p + 1) \end{aligned} \tag{12}$$

Thus, we have $Ew \geq 1$ for all p.

4.3 Test Accuracy

We can surmise that using Shannon's channel capacity theorem, we can achieve any arbitrary low probability of error by adding redundancy. To this end, we can model $X_n = (x_1, x_2, ..., x_n)$ as the transmitted signal and $R_n = (\mathcal{T}(x_1), \mathcal{T}(x_2), ..., \mathcal{T}(x_n))$ as the received signal. Every error in testing can then be modeled as a random vector $f = (f_1, f_2, ..., f_n)$, where $f_i = 1$ indicates an error in testing result of sample i and $f_i = 0$ indicates otherwise. Then, we can apply (say) Hamming coding to detect and possibly correct errors based on characteristics of our channel (i.e., our testing kit).

References

[1] IHME. (2020, April) Covid-19 projections assuming full social distancing through may 2020. [Online]. Available: https://covid19.healthdata.org/united-states-of-america

[2] J. Acosta and E. Cohen. (2020, April) Top public health official says number of dead could be lower as Americans practice social distancing. CNN. [Online]. Available: https://www.cnn.com/2020/04/07/politics/white-house-coronavirus-death-projections/index.html

[3] WHO. (2020, March) WHO Director-General's opening remarks at the media briefing on COVID-19–16 march 2020. World Health Organization. [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—16-march-2020

[4] J. Cohen. (2020, February) The United States badly bungled coronavirus testing–but things may soon improve. [Online]. Available: https://www.sciencemag.org/news/2020/02/united-states-badly-bungled-coronavirus-testing-things-may-soon-improve

[5] E. Werner, M. DeBonis, and P. Kane. (2020, March) Senate approves $2.2 trillion coronavirus bill aimed at slowing economic free fall. [Online]. Available: https://www.washingtonpost.com/business/2020/03/25/trump-senate-coronavirus-economic-stimulus-2-trillion/

[6] D. C. Rio, "Reverse transcription–polymerase chain reaction," *Cold Spring Harbor Protocols*, vol. 2014, no. 11, pp. pdb–prot080 887, 2014.

[7] N. Jawerth. (2020, March) How is the COVID-19 virus detected using real time RT-PCR? International Atomic Energy Agency (IAEA). [Online]. Available: https://www.iaea.org/newscenter/news/how-is-the-covid-19-virus-detected-using-real-time-rt-pcr

[8] Bloomberg School of Public Health. (2020, February) Serology testing for COVID-19. Johns Hopkins University. [Online]. Available: http://www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-fact-sheets/200228-Serology-testing-COVID.pdf

[9] The COVID Tracking Project. (2020, March) US historical data. [Online]. Available: https://covidtracking.com/data/us-daily

[10] Wikipedia. (2020, April) 2020 coronavirus pandemic in South Korea. [Online]. Available: https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_South_Korea