


RESEARCH ARTICLE

# A negative binomial approximation in group testing

Letian Yu<sup>1</sup>, Fraser Daly<sup>2</sup> and Oliver Johnson<sup>3</sup> 

<sup>1</sup>Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin NT, Hong Kong. E-mail: [letian.yu@link.cuhk.edu.hk](mailto:letian.yu@link.cuhk.edu.hk)

<sup>2</sup>Department of Actuarial Mathematics and Statistics, Heriot–Watt University, Edinburgh EH14 4AS, UK. E-mail: [f.daly@hw.ac.uk](mailto:f.daly@hw.ac.uk)

<sup>3</sup>School of Mathematics, University of Bristol, Fry Building, Woodland Road, Bristol BS8 1UG, UK. E-mail: [o.johnson@bristol.ac.uk](mailto:o.johnson@bristol.ac.uk)

**Keywords:** 62E17, 60F05, 94A20

## Abstract

We consider the problem of group testing (pooled testing), first introduced by Dorfman. For nonadaptive testing strategies, we refer to a nondefective item as “intruding” if it only appears in positive tests. Such items cause misclassification errors in the well-known COMP algorithm and can make other algorithms produce an error. It is therefore of interest to understand the distribution of the number of intruding items. We show that, under Bernoulli matrix designs, this distribution is well approximated in a variety of senses by a negative binomial distribution, allowing us to understand the performance of the two-stage conservative group testing algorithm of Aldridge.

## 1. Introduction to group testing

The group testing (pooled testing) problem was introduced by Dorfman [14] and provides a way of efficiently finding a small number of infected individuals in a large population. The key construction underlying this method is a so-called pooled test: given a subset  $S$  of the population, we combine samples from each member of  $S$  into a testing pool, and test them all together. We suppose that such a test returns a positive result if and only if at least one person in  $S$  is infected. Hence, a negative test allows us to deduce that every person in  $S$  is not infected, allowing for efficient screening of individuals. Some inference can be drawn from a positive test, but the analysis is typically more involved.

By carrying out a series of pooled tests, we hope to efficiently identify all the infected individuals—using as few tests as possible, and ideally using simple algorithms which do not require intensive computation.

The group testing problem has generated an extensive literature, surveyed for example in [5,15], and a large variety of variations on the problem exist. Group testing has been proposed as a solution to problems in a wide variety of fields (see [5] Sect. 1.7 for a survey of some of these applications). In engineering and computer science in particular, it has been used to efficiently search computer memories [24], to give a novel data compression algorithm [21], to detect high-demand items in databases [11], to bound the performance of multi-access communications channels [32], and in many other problems besides. Due to the shortage of tests early in the COVID-19 pandemic, group testing was a natural solution proposed to find infected individuals, and has indeed been deployed at scale in a variety of countries (see [3] Sect. 6 for a review of some such uses).

One key distinction is whether we are allowed to employ adaptive testing strategies (where the choice of individuals to be tested can depend on the outcome of previous tests) or are restricted to nonadaptive ones (where the test design is chosen in advance). Clearly, the ability to search adaptively cannot harm

us, and indeed it is known (see [5] Sect. 1.5) that in this case Hwang's algorithm [22], based on efficient binary search, requires a number of tests which is asymptotically optimal for a wide range of parameters. However, in many circumstances, adaptive algorithms may not be an option, and there is independent mathematical interest in understanding the performance of algorithms which are nonadaptive or are restricted to a small number of stages.

One motivating example for the study of group testing algorithms with restricted stages is that of COVID testing. It has been known since the early days of the coronavirus pandemic that a sample from one infected individual gives a strong enough PCR positivity signal when mixed in a pool of 32 or 64 samples that group testing is a potentially viable strategy [33]. However, to take advantage of the ability of PCR machines to perform 96 or more tests at a time in parallel [16] we need to use nonadaptive algorithms. Furthermore, as argued in for example [26], if each round of tests takes a few hours to be processed, then multi-stage binary search algorithms can give information about the infection status of individuals too late to be useful, meaning that the virus could have already been passed on before the results are received. For this reason, in this article, we will focus on nonadaptive and two-stage algorithms.

When we work nonadaptively, we must declare our testing strategy in advance, and it is perhaps not immediately obvious which strategy to choose. One simple idea, which we will refer to as Bernoulli testing, consists of randomly placing each member of the population into each pool with the same probability  $p$ , each such choice being taken independently of one another. In fact, if there are  $k$  infected individuals, it is generally a good strategy to choose  $p = 1/k$ , so that there is, on average, one infected individual in each test pool. Of course, Bernoulli testing is not the only strategy, and improved performance can be obtained by placing each person in a fixed number of tests at random [23], or by more advanced test designs [10]. However, Bernoulli testing is simple to describe and analyze, and generally gives performance [5] Sect. 2 within a constant multiple of the best possible, so we will focus on it here.

Having chosen a particular nonadaptive test strategy, a key question is how we find the infected individuals. A variety of algorithms are possible but one particularly simple one is referred to as COMP after its use in the paper [9], but dates back at least to the work of Kautz and Singleton [24]. This algorithm works as follows: as mentioned above, each person who appears in a negative test is guaranteed to be not infected. Hence, we can build a list of noninfected people by collecting together the people from each negative test. For definiteness, we simply assume that everyone else is infected.

Given enough tests, each noninfected person should appear in at least one negative test, but using arguments based on the coupon-collector problem we can deduce that this may require more tests than we would otherwise hope. If we use insufficiently many tests, the algorithm is likely to fail, with a clear single source of error. That is, if some noninfected person only appears in positive tests, then they will be incorrectly classified as infected. We refer to such a person as “intruding,” and the focus of this paper will be to count the number of intruding individuals, which we will refer to as  $G$ . (Each person declared to be noninfected will definitely be so—see [5] Lem. 2.3.)

The COMP algorithm succeeds in deducing every individual's infection status exactly if and only if  $G = 0$ , but by understanding the distribution of  $G$  we can also consider some related issues. First, as mentioned previously, given perfect testing, COMP never classifies a noninfected person as infected, and can be used to provide a quick screening of the population. Knowledge of  $G$  tells us precisely how many healthy people would be wrongly quarantined as a result of this screening, so given a particular tolerance of this effect we could choose the number of tests accordingly.

Second, we can regard COMP as the first stage of a “conservative two-stage group testing” algorithm as described by Aldridge [2], where following an initial screening using COMP we choose to test each person who has not received a clean bill of health using individual testing. [2] Thm. 1 describes the expected number of tests for such a procedure to succeed. If there are  $k$  infected people, clearly the second stage requires  $G + k$  tests to succeed, so by understanding the distribution of  $G$  we can approximate the probability this two-stage algorithm will succeed, which can give more information than the expected value. Furthermore, given an overall budget of  $T$  tests, we might wish to know the optimal number of tests  $T_1$  to use for the initial COMP stage, and analyzing the distribution of  $G$  will give insight into this.

Finally, the purely nonadaptive DD algorithm introduced in [4] uses COMP as a first stage of the analysis, and performs a further analysis based on looking for positive tests that contain exactly one “nonscreened” item. Informally, we know that DD will succeed if the number of intruding items  $G$  is much less than the number of defectives  $k$ , so again by understanding the distribution of  $G$  we gain insight into the performance of DD.

The structure of the remainder of the paper is as follows. In Section 2, we give a more formal introduction to the group testing problem, including introducing notation, defining the COMP algorithm, and proving some simple properties of the number of intruding items  $G$ . In Section 3, we show that  $G$  can be well approximated by a negative binomial distribution, first by considering a limiting argument that shows convergence of all falling moments in an asymptotic limit and then giving a more detailed bound based on a novel adaptation of the Stein–Chen method which gives bounds in finite blocklength settings as well. Section 4 discusses the implications of these results for various group testing algorithms, before a brief conclusion is given in Section 5.

## 2. Notation and definitions

### 2.1. Group testing setup

We now state the group testing problem in slightly more formal language and introduce some notation similar to that of [5]. We will write  $n$  for the total population of individuals, which we will refer to as “items.” Instead of referring to infected and healthy individuals, we will follow standard group testing terminology by calling them “defective” and “nondefective,” respectively. We write  $\mathcal{K}$  for the set of defective items (or defective set for short) and  $k = |\mathcal{K}|$  for the total number of defective items, and  $T$  for the number of tests.

As is standard, we can represent a nonadaptive testing strategy by a binary  $T \times n$  test matrix  $X$ , with rows corresponding to tests and columns corresponding to items. Here, the entry  $X_{ti} = 1$  means that item  $i$  appears in test  $t$ . In this paper, we focus on Bernoulli testing, where the  $(X_{ti})$  are independent Bernoulli random variables with parameter  $p$ . We will refer to this as a “Bernoulli test design with parameter  $p$ ,” and this design will apply throughout. Of particular interest will be the case  $p = 1/k$ .

The outcome of test  $t$  is represented as a binary value  $Y_t$  (where  $Y_t = 1$  means a positive test) which can be calculated for this test matrix as

$$Y_t = \bigvee_{i \in \mathcal{K}} X_{ti}, \quad (1)$$

where  $\bigvee$  represents a standard binary OR, capturing the fact that each test is positive if and only if it contains at least one of the items in  $\mathcal{K}$ .

As in [4] we write  $q_0 = (1 - p)^k$ , noting that each test is positive independently with probability  $1 - q_0$  (since it is negative, if and only if it contains none of the  $k$  defectives).

Again, as in [4,10] and other papers, we will often study what is referred to in [5] as the *sparse regime*. In this setting, the number of items  $n$  tends to infinity and the number of defectives  $k = k(n) = n^\theta$  for some explicit parameter  $\theta \in (0, 1)$ . In this context, it is natural to consider an asymptotic regime where the number of tests  $T = (c/q_0)k \log(n)$  for some constant  $c$ , noting that if  $p = 1/k$ , then asymptotically  $q_0$  converges to  $e^{-1}$ . Here and throughout our work, we write  $\log$  for the natural logarithm.

Another asymptotic setting of interest (see, e.g., [1] and [5] Sect. 5.5) is referred to as the *linear regime*. Here, again  $n$  tends to infinity, and the number of defectives  $k = k(n) = \beta n$  for some explicit parameter  $\beta \in (0, 1)$ . In this context, we consider an asymptotic regime where the number of tests  $T = cn$  for some constant  $c$ . Although, in this regime, Aldridge [1] proved that no nonadaptive algorithm can outperform individual testing, there is still interest in understanding the performance of two-stage or adaptive algorithms, and analysis of  $G$  of the kind presented in this paper can help with this.

However, in many practical group testing contexts, we are also interested in what we refer to as the *finite blocklength setting*, following terminology popularized for example by the work of Polyanskiy

et al. [27]. In this context, we wish to understand the performance of algorithms in solving concrete problems such as  $n = 500, k = 10$  (see [4]), where asymptotic bounds may not necessarily give the best guide to actual performance. Interest in finite blocklength problems was particularly prompted by the COVID pandemic, where for example the use of 96-well PCR plates [16] means that the number of tests may be bounded by (a multiple of) 96. This setting has typically been less well explored than asymptotic settings such as the sparse or linear regimes described above, however we provide some bounds in this context.

At various points, in our analysis, we will find it useful to work in terms of the falling moments of various random variables, and we will write  $M_{(s)}(Y) := \mathbb{E}(Y)_{(s)} = \mathbb{E}Y(Y-1) \dots (Y-s+1) = \mathbb{E}Y!/(Y-s)!$  for the  $s$ th falling moment of random variable  $Y$ . We will also write  $w(\mathbf{u})$  for the Hamming weight of the binary vector  $\mathbf{u}$ .

### 2.2. COMP algorithm

The COMP algorithm uses the test outputs  $Y$  and matrix  $X$  to produce an estimate of the defective set  $\mathcal{K}$  which we will write as  $\widehat{\mathcal{K}}_{\text{COMP}}$ . In fact, it is easier to consider the complement of  $\widehat{\mathcal{K}}_{\text{COMP}}$ : an item will appear in this complement if it appears in some negative test. Formally speaking

$$\widehat{\mathcal{K}}_{\text{COMP}}^c = \{i : Y_s = 0 \text{ for some } s \text{ with } X_{si} = 1\}. \tag{2}$$

Notice that if item  $j$  really is defective (i.e.,  $j \in \mathcal{K}$ ), then for every test  $s$  with  $X_{sj} = 1$  then  $Y_s = 1$  (by the definition of the group testing action in (1)), so that (2) implies that  $j$  is not in  $\widehat{\mathcal{K}}_{\text{COMP}}$ . In other words,  $\mathcal{K} \subseteq \widehat{\mathcal{K}}_{\text{COMP}}$  (see [5] Lem. 2.3).

COMP is an attractive algorithm because it is simple to perform and interpret, and because of this performance guarantee in one direction. However, in practice if we do not perform enough tests, then  $\widehat{\mathcal{K}}_{\text{COMP}}$  can be significantly larger than  $\mathcal{K}$ , meaning that many nondefective items would be misclassified, potentially causing problems in healthcare-related situations where unnecessary quarantine could result.

For this reason, Aldridge [2] proposed what he refers to as a conservative two-stage algorithm. Here, given a total budget of  $T$  tests, we should use  $T_1$  of them to perform the COMP algorithm in the usual way, and then the remaining  $T_2 := T - T_1$  tests to perform individual testing of each of the items in  $\widehat{\mathcal{K}}_{\text{COMP}}$ , which have not been classified as nondefective. While this algorithm may be inferior in performance to a two-stage algorithm which uses the  $T_1$  tests to perform the DD algorithm [4] followed by individual testing, it has the advantage of being transparent to perform for healthcare professionals without a mathematical background.

Clearly, the conservative two-stage algorithm of Aldridge [2] will succeed in finding all the defective items if the number of second stage tests is greater than or equal to the number of items to be tested. That is, it will succeed when  $T_2 = T - T_1 \geq |\widehat{\mathcal{K}}_{\text{COMP}}|$ , meaning that we would like to find the size of  $\widehat{\mathcal{K}}_{\text{COMP}}$ . Furthermore, for a given budget of  $T$  tests, since larger values of  $T_1$  give smaller  $\widehat{\mathcal{K}}_{\text{COMP}}$  (more stage one tests allow more nondefective items to be screened out) but leave fewer tests available in the second stage, we would like to find a sensible choice of  $T_1$  that manages this tradeoff.

### 2.3. Basic properties of intruding items

We now define the key property we will study in this paper:

**Definition 2.1.** We define

1. a nondefective item as “intruding” if it only appears in positive tests.
2. the binary random variables

$$G_i = \mathbb{I}(\text{item } i \text{ is nondefective and intruding}),$$

$\mathbf{G} = (G_1, \dots, G_n)$  the binary vector with these components, and  $G = \sum_i G_i = w(\mathbf{G})$ .

More formally, if nondefective item  $\ell$  is intruding, then  $Y_s = 1$  for every  $s$  with  $X_{s\ell} = 1$ , so that (by (2)) we know  $\ell \notin \widehat{\mathcal{K}}_{\text{COMP}}^c$ , so  $\ell \in \widehat{\mathcal{K}}_{\text{COMP}}$ . In other words, if a nondefective item is intruding, then COMP will mistakenly declare it to be defective. Since nonintruding items are declared to be nondefective by COMP, we know that the size of  $\widehat{\mathcal{K}}_{\text{COMP}}$  (which determines the success of the two-stage algorithm of Aldridge [2]) is exactly  $|\widehat{\mathcal{K}}_{\text{COMP}}| = k + G$ .

For a given nondefective item  $i$ , we can work out the marginal distribution of  $G_i$  relatively easily:

**Lemma 2.2.** *Under a Bernoulli test design with parameter  $p$ , for each nondefective item  $i$ , the marginal distribution of  $G_i$  is Bernoulli with parameter  $(1 - pq_0)^T$ , where we recall that we write  $q_0 = (1 - p)^k$ .*

*Proof.* Item  $i$  is intruding if no test contains item  $i$  and no defective item. The probability of the event that test  $t$  contains item  $i$  and no defective item is  $p(1 - p)^k = pq_0$ , so since successive tests are independent, the chance that we avoid this event for each test is  $(1 - pq_0)^T$ . □

### 3. Main results

#### 3.1. Moments and associated random variables

If the  $G_i$  were independent, then the analysis of the distribution of  $G$  would be easy: using Lemma 2.2, then  $G$  would be binomial with parameters  $n - k$  and  $(1 - pq_0)^T$ . However, there is a dependence between the  $G_i$ . In fact, if we learn that a given item is intruding, that suggests there might be more positive tests than average, which would imply that other items are more likely to be intruding.

We formalize this intuition by showing that the  $G_i$  are more likely to be equal to 1 together than independence would imply. That means that if COMP fails, it is more likely to fail badly (with a large total  $G$ ) than a naive analysis based on Lemma 2.2 might suggest. We prove this in two ways, first by showing in Corollary 3.3 that the  $G_i$  are pairwise positively correlated, and second by proving a stronger result (Proposition 3.5) which shows that the  $G_i$  have the property of association (see Definition 3.4), which is stronger than positive correlation (see the discussion in [17] for example).

In fact, we will deduce the positive correlation result (Corollary 3.3) from an expression for the falling moments of  $G$  (Proposition 3.1), which may be of independent interest, extending the result for EG implicit in [2] Thm. 1). We describe the distribution of  $G$  as a binomial mixture of binomial distributions as follows. As in [4], if we let  $M_0$  be the number of negative tests, then we know that:

$$M_0 \sim \text{Bin}(T, (1 - p)^k), \tag{3}$$

$$G \mid M_0 = m \sim \text{Bin}(n - k, (1 - p)^m). \tag{4}$$

The first result follows because a test is negative if and only if it contains no defective items, and for Bernoulli testing this occurs independently across tests with probability  $q_0 = (1 - p)^k$ . Moreover, the second result follows because a nondefective item is intruding if and only if it does not appear in any of the  $M_0$  negative tests, and each nondefective item is present in a given test independently with probability  $p$ . We can use this to prove the following result:

**Proposition 3.1.** *Under a Bernoulli test design with parameter  $p$ , the falling moments of  $G$  are given by*

$$M_{(s)}(G) = \mathbb{E}G(G - 1) \dots (G - s + 1) = \binom{n - k}{s} s! (1 - q_0 (1 - (1 - p)^s))^T, \tag{5}$$

for any integer  $s \geq 0$ .

We prove this result using the following intermediate lemma:

**Lemma 3.2** (Falling Moments of Binomial Distribution). *Suppose  $X \sim \text{Bin}(L, t)$ . Then, the  $s$ th falling moment of  $X$  is given by:*

$$M_{(s)}(X) = \binom{L}{s} s! \cdot t^s. \tag{6}$$

Proposition 3.1 follows using Lemma 3.2 when we recall the distributions of  $M_0$  and  $G \mid M_0 = m$  from (3) and (4), and apply the law of iterated expectation to express  $M_{(s)}(G) = \mathbb{E}[\mathbb{E}((G)_s \mid M_0)]$ . We omit the details for brevity.

Note that we can give an alternative proof of Proposition 3.1, using [20] Lem. 2.2, which gives a multinomial-type expansion for falling factorials based on the Vandermonde identity. This expansion simplifies in the case of binary random variables to give the fact that the falling moment can be expressed as a sum over sets:

$$M_{(s)}(G) = s! \sum_{S:|S|=s} \mathbb{E} \left( \prod_{i \in S} G_i \right). \tag{7}$$

The summation over sets contributes  $s! \binom{n-k}{s}$  equal expectation terms, each one of which equals the probability that all elements of a specified set are intruding, which corresponds to the event that none of them ever appear in a negative test, so the result follows by independence of all test items.

Using Proposition 3.1, we can deduce the following result that shows that the  $G_i$  have positive pairwise correlation:

**Corollary 3.3.** *For nondefective items  $i \neq j$ :*

$$\text{Cov}(G_i, G_j) = (1 - q_0(2p - p^2))^T - (1 - q_0p)^{2T} \geq 0. \tag{8}$$

*Proof.* We consider the variance of  $G$  in two different ways:

$$\text{Var}(G) = M_{(2)}(G) + \mathbb{E}G - (\mathbb{E}G)^2, \tag{9}$$

and (using the symmetry between pairs of  $G_i$  implied by the Bernoulli matrix design)

$$\begin{aligned} \text{Var}(G) &= \sum_i \text{Var}(G_i) + \sum_{i \neq j} \text{Cov}(G_i, G_j) \\ &= (n - k)(\mathbb{E}G_i - (\mathbb{E}G_i)^2) + (n - k)(n - k - 1)\text{Cov}(G_i, G_j) \\ &= \mathbb{E}G - \frac{(\mathbb{E}G)^2}{n - k} + (n - k)(n - k - 1)\text{Cov}(G_i, G_j), \end{aligned} \tag{10}$$

using the fact that for a binary random variable  $Y$  we have  $\text{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \mathbb{E}Y - (\mathbb{E}Y)^2$  and that  $(n - k)\mathbb{E}G_i = \mathbb{E}G$ . Now, equating (9) and (10), we obtain that

$$\begin{aligned} (n - k)(n - k - 1)\text{Cov}(G_i, G_j) &= M_{(2)}(G) - (\mathbb{E}G)^2 \left( 1 - \frac{1}{n - k} \right) \\ &= (n - k)(n - k - 1)[(1 - q_0(1 - (1 - p)^2))^T - (1 - q_0(1 - (1 - p)))^{2T}], \end{aligned}$$

using the expressions for  $M_{(2)}(G)$  and  $M_{(1)}(G)$  from Proposition 3.1, and the result follows on cancellation. □

However, we can prove a stronger property than just positive correlation. Recall the following definition:

**Definition 3.4** [17]. *Random variables  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  are (positively) associated if, for all increasing functions  $f$  and  $g$ ,*

$$\mathbb{E}(f(\mathbf{X})g(\mathbf{X})) \geq \mathbb{E}(f(\mathbf{X}))\mathbb{E}(g(\mathbf{X})). \tag{11}$$

We will prove the following proposition:

**Proposition 3.5.** *Under a Bernoulli test design, the random variables  $\mathbf{G} = (G_1, G_2, \dots, G_n)$  are associated.*

*Proof.* See [Appendix A](#). □

Combining Proposition 3.5 with Theorem 3.1 of [12], we find that  $G$  is larger, in a convex sense, than a binomial random variable  $H$  with parameters  $n - k$  and  $(1 - pq_0)^T$ . That is, we have that  $\mathbb{E}g(G) \geq \mathbb{E}g(H)$  for all real-valued functions  $g$  with  $g(x + 1) - 2g(x) + g(x - 1) \geq 0$  for all positive integers  $x$ , and for which the expectations exist (where we note from Property 3.4 of [13] that convex ordering on the integers is equivalent to convex ordering on the real line). This formalizes the notion that  $G$  is more variable than it would be if the  $G_i$  were independent.

The function  $g(x) = x!/(x - s)!$  satisfies this convexity condition, since direct calculation gives that  $g(x + 1) - 2g(x) + g(x - 1) = s(s - 1)(x - 1) \dots (x - s + 2)$  in this case. We thus deduce that (see Lemma 3.2 for the value of  $M_{(s)}(H)$ )

$$M_{(s)}(G) \geq M_{(s)}(H) = \binom{n - k}{s} s!(1 - pq_0)^{sT}, \tag{12}$$

that is, the falling moments of  $H$  act as lower bounds for those of  $G$ . Indeed, direct calculation gives that the ratio

$$\frac{M_{(s)}(G)}{M_{(s)}(H)} = \left( \frac{1 - q_0(1 - (1 - p)^s)}{(1 - pq_0)^s} \right)^T =: R(s)^T, \tag{13}$$

where  $R(s + 1) - R(s) = (pq_0(1 - q_0))(1 - (1 - p)^s)(1 - pq_0)^{-s-1} \geq 0$ , so that the ratio between successive falling moments of  $G$  and  $H$  is increasing in  $s$ .

Some numerical illustration of this is given in [Table 1](#), along with comparison of falling moments of  $G$  with those of other distributions which are more suitable than  $H$  as approximations of  $G$  and which we now discuss in more detail.

### 3.2. Negative binomial approximation

In this subsection, we start to show that the distribution of  $G$  can be approximated by  $Z$ , where  $Z \sim \text{NB}(r, q)$  follows the negative binomial distribution. For concreteness, we use the parameterization where the probability mass function for the negative binomial distribution is

$$\mathbb{P}(Z = z) = f(z; r, q) := \frac{\Gamma(z + r)}{\Gamma(r)z!} q^r (1 - q)^z \quad \text{for } z = 0, 1, 2, \dots \tag{14}$$

Note that in the case of integer  $r$ , the normalization constant  $\Gamma(z + r)/(\Gamma(r)z!) = \binom{z+r-1}{z}$ , and  $Z$  can be interpreted in terms of the number of failures to see the  $r$ th success in a sequence of Bernoulli trials. However, we do not require  $r$  to be an integer here.

The parameter  $r$  is sometimes referred to as the dispersion. In this sense, it is worth noting that the case  $r = 1$  corresponds to a geometric random variable (as mentioned above) and the limiting regime



**Table 1.** First four falling moments of approximating distributions in the case  $n = 500, k = 10, p = 0.1, T = 100$ .

$s$	$M_{(s)}(G)$ (true)	$M_{(s)}(Z)$ (negative binomial)	$M_{(s)}(Y)$ (Poisson)	$M_{(s)}(X)$ (geometric)	$M_{(s)}(H)$ (binomial)
1	14.088	14.088	14.088	14.088	14.088
2	252.71	252.71	198.49	397.0	198.09
3	5,716.9	5,505.1	2,796.6	16,779.4	2,779.5
4	161,487	141,110	39,400	945,605	38,919

True falling moments  $M_{(s)}(G)$  are given by (5). Negative binomial falling moments  $M_{(s)}(Z)$  are given by (15) with parameters are given by (18). Poisson falling moments are given by  $M_{(s)}(Y) = \lambda^s$ , where  $\lambda = M_{(1)}(G)$  is chosen to match the first moment of  $G$ . Geometric falling moments are given by  $M_{(s)}(X) = s!(1/\alpha - 1)^s$ , where  $\alpha = 1/(1 + M_{(1)}(G))$  is chosen to match the first moment of  $G$ . The binomial random variable  $H$  and its falling moments are as given in the discussion following Proposition 3.5.

$r \rightarrow \infty$  and  $q = r/(r + \lambda)$  (which corresponds to a mean-preserving limit—see Lemma 3.6) gives the mass function of a Poisson random variable with mean  $\lambda$ . However, in the finite blocklength setting, we will see that  $G$  is often well approximated by a distribution with  $1 \ll r \ll \infty$ , meaning that neither the geometric nor Poisson approximation are valuable.

One natural question is that of which negative binomial distribution (which choice of parameters) to use. We argue that one natural choice is based on a standard moment matching argument—since we need two parameters, we need to set  $\mathbb{E}G = \mathbb{E}Z$  and  $\mathbb{E}G^2 = \mathbb{E}Z^2$ . In fact, equivalently, since Proposition 3.1 and Lemma 3.6 give closed form expressions for the falling moments of  $G$  and  $Z$ , it is actually easier to solve  $M_{(s)}(G) = M_{(s)}(Z)$  for  $s = 1, 2$ . It is a straightforward exercise involving the Gamma function to prove the following:

**Lemma 3.6** (Falling moments of negative binomial distribution). *The falling moments of  $Z \sim NB(r, q)$  are given by:*

$$M_{(s)}(Z) = \frac{\Gamma(s+r)}{\Gamma(r)} \left(\frac{1-q}{q}\right)^s. \tag{15}$$

Hence, combining Proposition 3.1 and Lemma 3.6, we have successfully matched the moments if

$$\frac{r(1-q)}{q} = (n-k)(1-q_0p)^T = M_{(1)}(G), \tag{16}$$

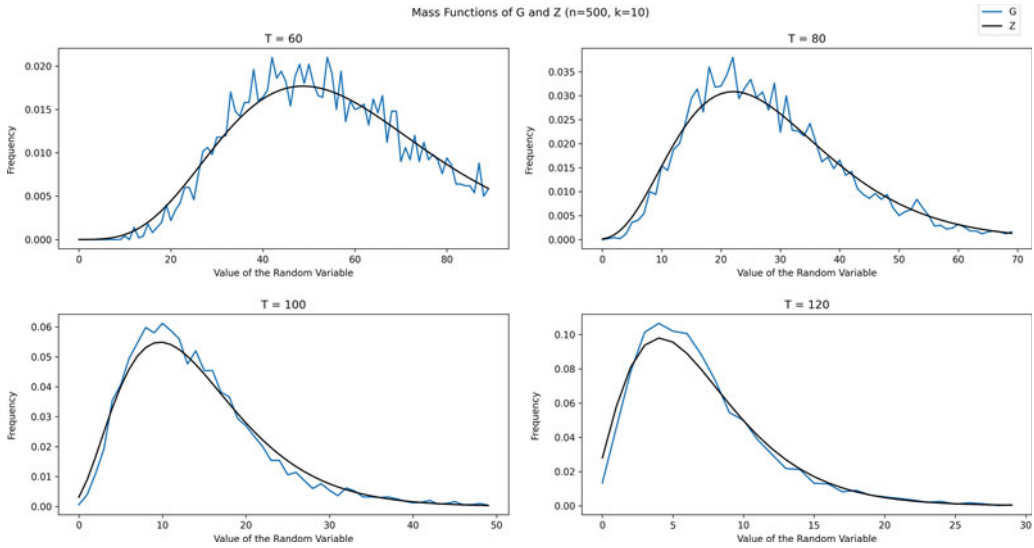
$$\frac{r(r+1)(1-q)^2}{q^2} = (n-k)(n-k-1)(1-q_0(2p-p^2))^T = M_{(2)}(G), \tag{17}$$

or, equivalently, if

$$r = \frac{M_{(1)}(G)^2}{M_{(2)}(G) - M_{(1)}(G)^2} \quad \text{and} \quad q = \frac{M_{(1)}(G)}{M_{(2)}(G) + M_{(1)}(G) - M_{(1)}(G)^2}. \tag{18}$$

In Figure 1, we illustrate the quality of this negative binomial approximation for  $G$  in the case  $n = 500, k = 10$ , and  $p = 0.1$ . We plot the mass function of the negative binomial random variable  $Z$  (with parameter choices as in (18)) and the corresponding estimated mass function for  $G$  obtained by simulation, for several values of the number of tests  $T$  in a nonadaptive algorithm. These experiments





**Figure 1.** Probability mass functions of  $G$  (blue, obtained by simulation) and approximating negative binomial random variable  $Z$  (black). In this case, the group testing parameters are  $n = 500$ ,  $k = 10$ , and  $p = 0.1$ . The four plots correspond to  $T = 60, 80, 100, 120$ .

show that this negative binomial distribution gives a good approximation to the distribution of  $G$  for various different values of  $T$ .

We also find that for many finite blocklength examples the extra flexibility offered by a two parameter approximation means that the negative binomial distribution with parameters given by (18) approximates the low-order falling moments of  $G$  (and hence the variance, skewness, and kurtosis) better than either the Poisson or geometric distributions with a single parameter chosen by matching means. This is illustrated in Table 1, again in the case  $n = 500$ ,  $k = 10$ ,  $p = 0.1$ ,  $T = 100$ . In this setting, the dispersion parameter  $r$  of  $Z$  is 3.66, which is well separated from both  $r = \infty$ , corresponding to the Poisson, and  $r = 1$ , corresponding to the geometric.

In general, direct calculation shows that for any  $n, p, q, T$  such that  $r \geq 1$ , then writing  $X, Z, Y$ , and  $H$  for the geometric, negative binomial, Poisson, and binomial, respectively, defined in Table 1, we have

$$M_{(s)}(X) \geq M_{(s)}(Z) \geq M_{(s)}(Y) \geq M_{(s)}(H) \quad \text{for all } s \geq 1. \tag{19}$$

This follows by rewriting the expressions in (19) in the form

$$s! \left( \frac{r(1-q)}{q} \right)^s \geq \frac{\Gamma(s+r)}{\Gamma(r)} \left( \frac{(1-q)}{q} \right)^s \geq \left( \frac{r(1-q)}{q} \right)^s \geq \left( \frac{r(1-q)}{q} \right)^s \frac{\binom{n-k}{s} s!}{(n-k)^s}$$

which simplifies to

$$s! r^s \geq (s+r-1)(s+r-2) \dots r \geq r^s \geq r^s \frac{(n-k) \dots (n-k-s+1)}{(n-k)^s}.$$

Recall from (12) above that  $M_{(s)}(G) \geq M_{(s)}(H)$  for all  $s$ . However, it is not the case that  $M_{(s)}(G) \geq M_{(s)}(Y)$  for all  $s$ , since  $M_{(s)}(G) = 0$  for  $s > n - k$ , whereas  $M_{(s)}(Y) > 0$  for all  $s$ , although from Table 1 it appears that  $M_{(s)}(G) \geq M_{(s)}(Y)$  for small  $s$  at least.

In the sparse regime described above (in which  $k = n^\theta$ ,  $p = 1/k$  and  $T = (c/q_0)k \log(n)$ ), we note that the dispersion parameter  $r$  of (18) tends to infinity as  $n \rightarrow \infty$ . This is equivalent to the fact that

$M_{(2)}(G)/M_{(1)}(G)^2$  tends to 1. We can deduce that this holds from (18) by writing

$$\begin{aligned} \frac{M_{(2)}(G)}{M_{(1)}(G)^2} &= \frac{(n-k)(n-k-1)(1-q_0(2p-p^2))^T}{(n-k)^2(1-q_0p)^{2T}} \\ &\simeq \left(1 + \frac{p^2q_0(1-q_0)}{(1-q_0p)^2}\right)^T \\ &\simeq \exp\left(\frac{p^2Tq_0(1-q_0)}{(1-q_0p)^2}\right) \\ &\simeq \exp\left(\frac{c(1-q_0)\log(n)}{k}\right) \rightarrow 1, \end{aligned} \tag{20}$$

since  $q_0p^2T = c \log(n)/k$ .

Similarly, in the linear regime ( $k = \beta n, T = cn$ ) using (20), we obtain

$$\frac{M_{(2)}(G)}{M_{(1)}(G)^2} \simeq \exp\left(\frac{cq_0(1-q_0)}{\beta^2n}\right) \rightarrow 1,$$

since  $p^2T = c/(\beta^2n)$ . A Poisson approximation to the distribution of  $G$  may thus be appropriate in the large- $n$  limit for the sparse and linear regimes, but the numerical results of this section make it clear that a negative binomial approximation is a more natural choice for finite blocklength applications.

### 3.3. Convergence of falling moments

Next, we show that, in this framework, matching the first two moments of  $G$  and  $Z$  ensures that all the falling moments converge.

We can see this informally by simplifying (5) and (15), respectively. The former gives (to leading order)

$$\begin{aligned} M_{(s)}(G) &= \binom{n-k}{s} s! (1 - q_0(1 - (1-p)^s))^T \\ &\simeq (n-k)^s (1 - q_0(ps))^T \\ &\simeq (n-k)^s \exp(-q_0psT) = ((n-k) \exp(-q_0pT))^s, \end{aligned} \tag{21}$$

using the fact that  $1 - (1-p)^s = 1 - (1-ps + O(p^2)) = ps + O(p^2)$ . The latter gives

$$M_{(s)}(Z) = \frac{\Gamma(s+r)}{\Gamma(r)} \left(\frac{1-q}{q}\right)^s \simeq \left(\frac{r(1-q)}{q}\right)^s. \tag{22}$$

Note that the moment matching condition of (16) gives that  $r(1-q)/q = (n-k)(1-q_0p)^T \simeq (n-k) \exp(-q_0pT)$ , meaning that (21) and (22) agree. A more formal comparison of falling moments is given in the following theorem, where we recall that with the usual choice  $p = 1/k$  we have  $q_0 = (1-p)^k \approx e^{-1}$ :

**Theorem 3.7.** Consider the number of intruding defectives  $G$  and negative binomial  $Z$  with parameters given by moment matching, satisfying (18). Under a Bernoulli test design, for any integer  $s \geq 1$ , if we write  $C = q_0(1-q_0)/(1-q_0p)^2$ , then the moment ratio satisfies

$$\frac{M_{(s)}(G)}{M_{(s)}(Z)} \geq \left(\frac{n-k-s}{(n-k)(1+(s-1)/(2r))}\right)^s \left(1 + \frac{1}{2}s(s-1)Cp^2 \left(1 - \frac{(s-2)(1-2q_0p)}{3(1-q_0p)}\right)\right)^T, \tag{23}$$

and

$$\frac{M_{(s)}(G)}{M_{(s)}(Z)} \leq \exp(s(s-1)Cp^2T(1-q_0p)^{2-s}). \tag{24}$$

*Proof.* See Appendix B. □

Hence, for any fixed  $s$ , the ratio  $M_{(s)}(G)/M_{(s)}(Z) \rightarrow 1$ , for both

1. the sparse regime with  $k = n^\theta$  for some  $\theta \in (0, 1)$  and  $T = cek \log n$ , and
2. the linear regime with  $k = \beta n$  and  $T = cn$ .

Here, we control the upper bound in (24) using the fact that (see Section 3.2) the  $p^2T$  is  $c \log n / (q_0k)$  or  $c / (\beta^2n)$ , respectively. Similarly, we control the lower bound (23) using the fact that in both regimes the  $n - k$  and dispersion parameter  $r$  tend to infinity (again see Section 3.2).

Additionally, the bounds (23) and (24) allow us to control the ratio  $M_{(s)}(G)/M_{(s)}(Z)$  in the finite blocklength regime, deducing bounds that can be compared with the concrete values given in Table 1 for example.

### 3.4. Stein–Chen method

Having seen in Sections 3.2 and 3.3 that a negative binomial distribution seems to be a reasonable approximation for the distribution of  $G$ , in this section, we adapt the Stein–Chen method to prove explicit error bounds in the approximation of  $G$  by a negative binomial distribution. We emphasize that these bounds apply for any finite blocklength application. The error in our approximation of  $G$  by  $Z$  will be measured in total variation distance, defined by

$$d_{TV}(G, Z) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{P}(G \in A) - \mathbb{P}(Z \in A)| = \inf_{(G, Z)} \mathbb{P}(G \neq Z), \tag{25}$$

where  $\mathbb{Z}^+ = \{0, 1, \dots\}$  and the infimum is taken over all couplings of  $G$  and  $Z$ .

First, we briefly review the Stein–Chen method in the context of negative binomial approximation. For a more detailed introduction to this technique more generally, we refer the reader to [28]. Recall from [8] that  $Z \sim \text{NB}(r, q)$  if and only if

$$\mathbb{E}[(1-q)(r+Z)g(Z+1) - Zg(Z)] = 0, \tag{26}$$

for all test functions  $g : \mathbb{Z}^+ \rightarrow \mathbb{R}$  for which the expectation exists. This follows as a consequence of the fact that the negative binomial probability mass function of (14) satisfies  $z\mathbb{P}(Z = z) = (1-q)(r+z-1)\mathbb{P}(Z = z-1)$ .

The key to the analysis is that for each set  $A \subseteq \mathbb{Z}^+$  we can define a function  $f_A : \mathbb{Z}^+ \rightarrow \mathbb{R}$  which satisfies  $f_A(0) = 0$  and the Stein–Chen equation

$$(1-q)(r+z)f_A(z+1) - zf_A(z) = \mathbb{I}(z \in A) - \mathbb{P}(Z \in A), \tag{27}$$

for all  $z \in \mathbb{Z}^+$ . Then, for any random variable  $Y$ , we can take the expectation of (27) over  $Y$  to obtain

$$\mathbb{E}((1-q)(r+Y)f_A(Y+1) - Yf_A(Y)) = \mathbb{P}(Y \in A) - \mathbb{P}(Z \in A). \tag{28}$$

Note that (as expected) if  $Y$  were negative binomial, then both RHS and LHS of (28) would be zero (the latter due to the characterization in (26)). However, we can deduce that if the LHS of (28) is small, then so is the RHS. Indeed, if the LHS of (28) is small uniformly over choices of  $A$ , then we can deduce a bound in total variation distance since, combining (25) and (27), we have

$$d_{TV}(Y, W) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{E}((1-q)(r+Y)f_A(Y+1) - Yf_A(Y))|. \tag{29}$$

Having reviewed the Stein–Chen method in general, we will now describe how we bound (28) in this specific case. For our approximating negative binomial distribution for  $G$ , we will make the following choices of the parameters  $q$  and  $r$ :

$$q = \frac{\mu}{\sigma^2} \quad \text{and} \quad r = \frac{\mu^2}{\sigma^2 - \mu} = \frac{1}{e^{Tp^2q_0} - 1}, \tag{30}$$

where

$$\mu = (n - k)e^{-Tpq_0} \quad \text{and} \quad \sigma^2 = (n - k)^2(e^{-Tp(2-p)q_0} - e^{-2Tpq_0}) + (n - k)e^{-Tpq_0},$$

and where, as before, we write  $q_0 = (1 - p)^k$ .

We remark that these parameter choices *do not* match the first two moments of  $Z$  with those of  $G$ , unlike those of (18). Instead, the parameters  $q$  and  $r$  are chosen to match the first two moments of  $Z$  with those of  $G''$ , to be defined precisely below, in which the binomial mixture which defines  $G$  is replaced by a particular Poisson mixture. This may seem a little unnatural at first, but makes sense in the setting of the proof in Appendix C, and the ultimate effect should be negligible since the distributions of  $G$  and  $G''$  are close, as our proof demonstrates.

Our main result is the following.

**Theorem 3.8.** *Let  $G$  be as above, and let  $Z$  have a negative binomial distribution with parameters  $q$  and  $r$  given by (30). Then, defining  $K = e^{Tpq_0}$ , we have*

$$d_{TV}(G, Z) \leq 2 \min \left\{ \frac{q_0}{4\sqrt{1 - q_0}}, \frac{1}{\sqrt{T}}\alpha(q_0) + \frac{1}{\sqrt{2\pi e}} \log \left( \frac{1}{\sqrt{1 - q_0}} \right) \right\} + \frac{1}{K} + \frac{(2 - q)(n - k)}{1 - q} \left( e^{r+1} K^r \exp(-Kr) + \int_0^1 \left| \widehat{\Gamma} \left( \left\lfloor \frac{\log(x)}{\log(1 - p)} \right\rfloor, Tq_0 \right) - \widehat{\Gamma}(r, Krx) \right| dx \right), \tag{31}$$

where

$$\alpha(q_0) = \frac{0.4748[\sqrt{1 - q_0}(1 + 2q_0^2e^{-q_0}) + q_0^2 + (1 - q_0)^2]}{\sqrt{q_0(1 - q_0)}},$$

and  $\widehat{\Gamma}(\cdot, \cdot)$  is the normalized upper incomplete gamma function, defined by

$$\widehat{\Gamma}(s, y) = \frac{1}{\Gamma(s)} \int_y^\infty t^{-s-1} e^{-t} dt,$$

where  $\Gamma(\cdot)$  is the gamma function.

*Proof.* See Appendix C. □

We conclude this section with some discussion and numerical illustration of the bound of Theorem 3.8. We will discuss further aspects of the convergence of this bound in Remark 3.9, but first use numerical illustrations to gain some initial understanding of its behavior. Firstly, we note that although our bound applies for any finite blocklength, there are examples in which it is worse than the trivial bound  $d_{TV}(G, Z) \leq 1$ , or performs poorly compared with the simulation results we observed in Section 3.2. For example, with  $n = 500$ ,  $k = 10$ ,  $p = 0.1$ , and  $T = 100$ , the bound of Theorem 3.8 gives  $d_{TV}(G, Z) \leq 1.80$ , which is clearly uninformative. We give some further examples of our upper bound in Table 2, all in the case  $n = 2500$  and for various values of  $k$ ,  $p$ , and  $T$ , from which it is clear that there are some cases in which the bound of Theorem 3.8 performs very well, and demonstrates proximity of the distribution of  $G$  to negative binomial.

**Table 2.** The upper bound of Theorem 3.8 in the case  $n = 2500$  for various values of  $k$ ,  $p$ , and  $T$ .

	$T = 500$			$T = 1000$		
	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.05$	$p = 0.1$	$p = 0.2$
$k = 5$	0.501	0.337	0.120	0.460	—	—
$k = 10$	0.342	0.216	0.066	0.307	0.198	0.057
$k = 20$	0.234	—	—	0.200	0.065	—

The symbol “—” indicates that the upper bound is larger than 1, and therefore uninformative.

It is interesting to note that in many cases the largest share of the error estimate in Theorem 3.8 comes from the first term of the upper bound, which arises from the approximation of the binomial random variable  $M_0$  by a Poisson random variable of the same mean (see the proof in Appendix C for details). Nevertheless, the bound we would obtain without making this approximation generally performs worse than our Theorem 3.8. We conjecture that this is because the integral in the final term of the upper bound is made smaller by this approximation of  $M_0$  (compared with the corresponding term without this approximation), resulting in a smaller upper bound overall because of the relatively large factors multiplying this integral.

**Remark 3.9.** While the incomplete gamma functions in (31) make the integral in that bound not straightforward to interpret directly, we can provide an upper bound on this quantity using concentration of measure inequalities.

We split the region of integration in three parts, with breaks at  $(1 \pm \epsilon)/K$ . On the region  $((1 - \epsilon)/K, (1 + \epsilon)/K)$ , we simply bound the integrand by 1, to give  $2\epsilon/K$ . Using the fact that for  $0 \leq u, v \leq 1$  we can bound  $|u - v| \leq \max(u, v)$  and  $|u - v| \leq \max(1 - u, 1 - v)$ , we can hence bound the integral in (31) by

$$\begin{aligned} & \frac{2\epsilon}{K} + \frac{1}{K} \max \left( \mathbb{P} \left( \xi' < \frac{(1 - \epsilon)}{K} \right), \mathbb{P} \left( \eta' < \frac{(1 - \epsilon)}{K} \right) \right) \\ & + \max \left( \mathbb{P} \left( \xi' > \frac{(1 + \epsilon)}{K} \right), \mathbb{P} \left( \eta' > \frac{(1 + \epsilon)}{K} \right) \right), \end{aligned}$$

where  $\eta' \sim \Gamma(r, rK)$  and  $\xi' = (1 - p)^{M'}$ , with  $M' \sim \text{Po}(Tq_0)$ ; see also (C.9) in the proof of Theorem 3.8.

The inequalities (C.5) and (C.6) below give us that we may upper bound both  $\mathbb{P}(\eta' > z)$  and  $\mathbb{P}(\eta' < z)$  by

$$(Kze)^r \exp(-rKz) = \exp(r(1 - Kz + \log(Kz))).$$

Hence, for example, we know by writing  $m(s) = s - \log(1 + s) \simeq s^2/2$  that

$$\mathbb{P} \left( \eta' > \frac{(1 + \epsilon)}{K} \right) \leq \exp(-rm(\epsilon)) \quad \text{and} \quad \mathbb{P} \left( \eta' < \frac{(1 - \epsilon)}{K} \right) \leq \exp(-rm(-\epsilon)).$$

A similar standard Chernoff bounding argument, as used for (C.5) and (C.6), gives that, for  $Y \sim \text{Po}(\lambda)$ , both  $\mathbb{P}(Y > y)$  and  $\mathbb{P}(Y < y)$  are bounded above by  $\exp(-\lambda h(y/\lambda - 1))$ , where  $h(s) = (1 + s) \log(1 + s) - s \simeq s^2/2$ . Hence, for example, we can bound

$$\begin{aligned} \mathbb{P}(\xi' > x) &= \mathbb{P} \left( M' < \frac{-\log x}{-\log(1 - p)} \right) \leq \mathbb{P} \left( M' < \frac{-\log x}{p} \right) \\ &\leq \exp \left( -Tq_0 h \left( \frac{-\log x}{Tq_0 p} - 1 \right) \right), \end{aligned}$$

if  $-\log x/p \leq \mathbb{E}M' = Tq_0$ , since  $M' \sim \text{Po}(Tq_0)$ . Hence, taking  $x = (1+\epsilon)/K = (1+\epsilon)e^{-Tpq_0}$ , we obtain

$$\mathbb{P}\left(\xi' > \frac{(1+\epsilon)}{K}\right) \leq \exp\left(-Tq_0h\left(-\frac{\log(1+\epsilon)}{Tq_0p}\right)\right).$$

Similarly, we can bound

$$\mathbb{P}(\xi' < x) = \mathbb{P}\left(M' > \frac{-\log x}{-\log(1-p)}\right) \leq \exp\left(-Tq_0h\left(\frac{-\log x}{-Tq_0\log(1-p)} - 1\right)\right),$$

if  $-\log x/-\log(1-p) \leq \mathbb{E}M' = Tq_0$ , since  $M' \sim \text{Po}(Tq_0)$ . Hence, taking  $x = (1-\epsilon)/K = (1-\epsilon)e^{-Tpq_0}$  and writing  $m(-p) = -p - \log(1-p) \geq 0$  as above, we obtain that

$$\mathbb{P}\left(\xi' < \frac{(1-\epsilon)}{K}\right) \leq \exp\left(-Tq_0h\left(-\frac{-\log(1-\epsilon) - Tq_0m(-p)}{-Tq_0\log(1-p)}\right)\right),$$

assuming that the numerator is positive, which holds, for example, if  $\epsilon > Tq_0m(-p)$ . This condition is satisfied, for example, for large  $n$  in the linear regime where  $T = cn$ ,  $k = \beta n$ , and  $p = 1/k$ , since  $m(p) \simeq p^2/2$  so  $Tq_0m(-p) \sim \text{const.}/n$  in this regime.

#### 4. Implications for group testing algorithms

As discussed previously, understanding the distribution of  $G$  allows us to control the performance of the conservative two-stage algorithm of Aldridge [2]. Specifically, we consider the scenario where  $T_1$  tests are performed in the first stage according to a Bernoulli testing design with parameter  $1/k$ , and then  $T_2$  individual tests are performed afterwards to resolve the status of items we are unsure about.

Given a fixed total budget of  $T$  tests, we can regard the standard Bernoulli test design as corresponding to  $T_1 = T$  and  $T_2 = 0$  (no second stage), and individual testing as corresponding to  $T_1 = 0$  and  $T_2 = T$  (no first stage). However, it is natural to consider strategies intermediate to these, to see if better performance can be obtained by choosing some  $T$  satisfying  $0 < T_1 \leq T$ .

Given an overall budget of  $T$  tests, this leaves us  $T_2 = T - T_1$  individual tests to find the status of  $k + G$  items, and we will succeed if  $T_2 \geq k + G$  or, equivalently, if  $G + T_1 + k \leq T$ . Equivalently, the algorithm will fail if  $G > T - T_1 - k$ .

Aldridge [2] Thm. 1 considers this failure event in terms of expected values. That is, we may wish to choose  $T_1$  to minimize

$$\begin{aligned} \mathbb{E}(T_1 + k + G) &= T_1 + k + M_{(1)}(G) = T_1 + k + (n - k)(1 - q_0p)^{T_1} \\ &\simeq T_1 + k + (n - k) \exp\left(-\frac{1}{ek}T_1\right), \end{aligned} \tag{32}$$

and as in [2] direct calculation gives that the optimal choice of  $T_1$  to control this expectation is

$$T_1^* = ke \log\left(\frac{n - k}{ke}\right). \tag{33}$$

Interestingly, since  $p = 1/k$  and  $q_0 \simeq 1/e$ , this choice makes the expected value

$$M_{(1)}(G) = (n - k)(1 - q_0p)^{T_1} \simeq (n - k) \exp\left(-\frac{T_1}{ek}\right) = k,$$

meaning that the average number of intruding nondefectives approximately equals the number of true defectives, so a randomly chosen individual test is positive with probability close to  $1/2$ , maximizing the information that we gain from it.

However, instead of simply finding the expected value we can use the moment values calculated in this paper to bound the error probability under this kind of two-stage strategy. For example, Chebyshev's inequality gives an upper bound on the error probability:

**Lemma 4.1.** *If we use  $T_1 = kc_1$  tests in the first stage and  $T_2 = kc_2$  tests in the second stage*

$$\mathbb{P}(\text{err}) \leq \min \left( 1, \frac{M_{(2)}(G)/M_{(1)}(G)^2 - 1 + 1/M_{(1)}(G)}{(n\beta(c_2 - 1)/M_{(1)}(G) - 1)^2} \right). \tag{34}$$

*Proof.* A standard argument gives

$$\begin{aligned} \mathbb{P}(\text{err}) &= \mathbb{P}(G > T_2 - k) = \mathbb{P}(G - \mathbb{E}G > T_2 - k - \mathbb{E}G) \\ &\leq \frac{\text{Var}(G)}{(T_2 - k - \mathbb{E}G)^2} = \frac{M_{(2)}(G) - M_{(1)}(G)^2 + M_{(1)}(G)}{(n\beta(c_2 - 1) - M_{(1)}(G))^2}, \end{aligned}$$

and the result follows. □

Note that the expression (34) becomes

$$\simeq \frac{\exp(c_1q_0p) - 1 + \exp(c_1q_0)/n}{(\beta(c_2 - 1) \exp(c_1q_0)/(1 - \beta) - 1)^2} \tag{35}$$

in the linear regime ( $k = \beta n$ ), since  $M_{(1)}(G) = (n - k)(1 - q_0p)^{T_1} \simeq n(1 - \beta) \exp(-c_1q_0)$  and  $M_{(2)}(G) = (n - k)(n - k - 1)(1 - q_0(2p - p^2))^{T_2} \simeq n^2(1 - \beta)^2 \exp(-c_1q_0(2 - p))$ .

Hence, if we take  $c_1 = e \log((1 - \beta)/(\beta e))$  as suggested by (33), so that  $\exp(c_1q_0) \simeq (1 - \beta)/\beta$ , then (35) becomes

$$\mathbb{P}(\text{err}) \leq \frac{\exp(2c_1q_0p) - 1 + (1 - \beta)/(n\beta)}{(c_2 - 2)^2},$$

so for any fixed  $c_2 > 2$  the error probability tends to zero at rate  $1/n$  as  $n \rightarrow \infty$ .

Furthermore, using the negative binomial approximation of this paper, we can find large deviations bounds on the success probability of the two-stage algorithm. Figure 2 shows that, in this case, this negative binomial approximation provides accurate bounds on the total number of tests needed.

That is, if we write  $Z$  for the negative binomial approximation with parameters given by (18), we know that for any  $g$  the tail probability  $\mathbb{P}(G > g) \simeq \mathbb{P}(Z > g)$ . Using a standard large deviations argument, we know:

**Proposition 4.2.** *If  $Z$  is negative binomial with parameters  $r$  and  $q$ , then for any  $g > \mathbb{E}Z$ :*

$$\mathbb{P}(Z \geq g) \leq \exp \left( -(g + r) D_{\text{KL}} \left( \frac{g}{g + r} \parallel 1 - q \right) \right), \tag{36}$$

where we write  $D_{\text{KL}}(v \parallel w) = v \log_e(v/w) + (1 - v) \log_e((1 - v)/(1 - w))$  for the Kullback–Leibler divergence from a Bernoulli( $v$ ) random variable to a Bernoulli( $w$ ).

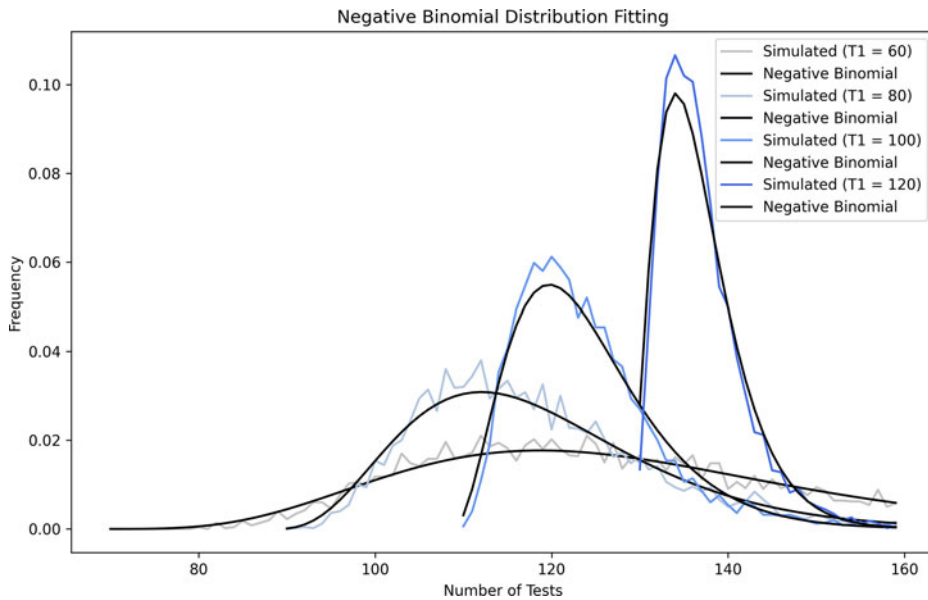
*Proof.* See Appendix D. □

Hence, again in the linear scenario ( $k = \beta n$ ), taking  $T_1 = k\beta e \log((1 - \beta)/\beta e)$  tests in the first stage (as suggested by (33)) and  $T_2 = kc_2$  in the second, then the probability of failure will be

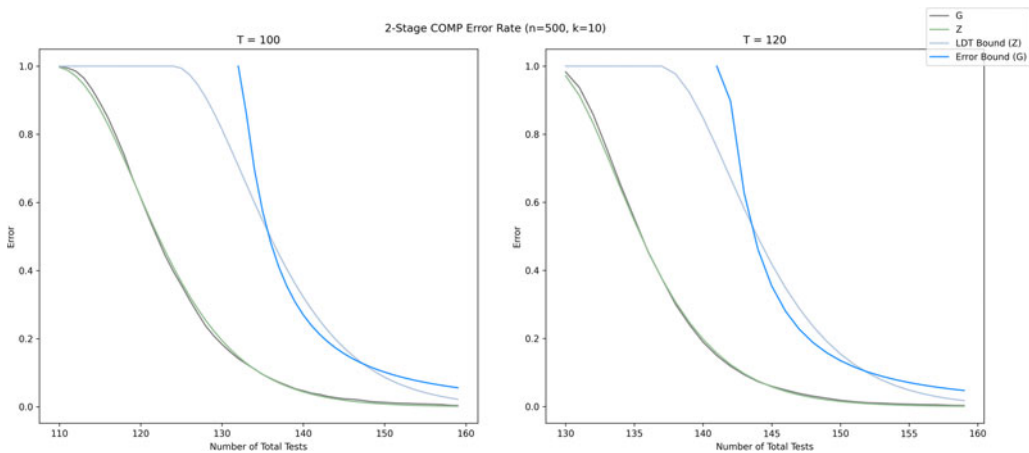
$$\mathbb{P}(G > T_2 - k) = \mathbb{P}(G > k(c_2 - 1)) \simeq \mathbb{P}(Z > k(c_2 - 1)), \tag{37}$$

which we can bound using Proposition 4.2. Using the explicit bounds above, we obtain an upper bound decaying exponentially in  $k$ .





**Figure 2.** Comparison of probability mass function of  $G$  (via simulation) and approximating negative binomial distribution  $Z$  (via calculation) for 2-stage COMP with a variety of values of  $T_1$  ( $n = 500$ ,  $k = 10$ ,  $p = 0.1$ ).



**Figure 3.** Comparison of error probability for 2-stage COMP - via simulation for  $G$  and direct calculation of negative binomial approximation  $Z$ . Upper bounds implied by (34) and (36) are provided for comparison ( $n = 500$ ,  $k = 10$ ,  $p = 0.1$ ).

In Figure 3, we see that the negative binomial approximation  $Z$  well approximates the distribution of  $G$  and that the upper bounds implied by (34) and (36) are somewhat tight.

This analysis could be extended to cover a two-stage version of the DD algorithm of [4]. In Stage 1 of this algorithm, there would potentially be some items which could be confirmed as “definitely defective” (because they appear in some positive test only otherwise containing items which are guaranteed by other tests to be nondefective). Such items would not need individual testing in Stage 2, which could potentially reduce the number of tests required by up to  $k$ ; this could be significant in the linear regime

at least. However, we leave this question as further work, for reasons of space and due to the complexity of the analysis of nonadaptive DD in [4].

## 5. Conclusion

We have identified the key role played by  $G$ , the number of intruding nondefective items, in group testing algorithms. Under the standard Bernoulli testing strategy, we have identified the distribution of  $G$ , given explicit expressions for its falling moments and shown how it can be well approximated by a negative binomial distribution with given parameters. This allows us to deduce results concerning the performance of the COMP and DD group testing algorithms.

**Acknowledgments.** The collaboration between Letian Yu and Oliver Johnson was supported by a Charles Kao bursary from the Chinese University of Hong Kong in the summer of 2021. We thank the Associate Editor and a referee for their helpful comments and suggestions.

**Competing interests.** The authors declare no conflict of interest.

## References

- [1] Aldridge, M.P. (2019). Individual testing is optimal for nonadaptive group testing in the linear regime. *IEEE Transactions on Information Theory* 65(4): 2058–2061.
- [2] Aldridge, M.P. (2020). Conservative two-stage group testing. arXiv:2005.06617.
- [3] Aldridge, M.P. & Ellis, D. (2022). Pooled testing and its applications in the COVID-19 pandemic. In M. del Carmen Boado-Penas, J. Eisenberg, & S. Sahin (eds), *Pandemics: Insurance and social protection*. Cham: Springer, pp. 217–249.
- [4] Aldridge, M.P., Baldassini, L., & Johnson, O.T. (2014). Group testing algorithms: Bounds and simulations. *IEEE Transactions on Information Theory* 60(6): 3671–3687.
- [5] Aldridge, M.P., Johnson, O.T., & Scarlett, J.M. (2019). Group testing: An information theory perspective. *Foundations and Trends in Information Theory* 15(3–4): 196–392.
- [6] Barbour, A.D., Holst, L., & Janson, S. (1992). *Poisson approximation*. Oxford: Clarendon Press.
- [7] Barbour, A.D., Gan, H.L., & Xia, A. (2015). Stein factors for negative binomial approximation in Wasserstein distance. *Bernoulli* 21(2): 1002–1013.
- [8] Brown, T.C. & Phillips, M.J. (1999). Negative binomial approximation with Stein's method. *Methodology and Computing in Applied Probability* 1(4): 407–421.
- [9] Chan, C.L., Che, P.H., Jaggi, S., & Saligrama, V. (2011). Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing, September*, pp. 1832–1839.
- [10] Coja-Oghlan, A., Gebhard, O., Hahn-Klimroth, M., & Loick, P. (2020). Optimal group testing. *Proceedings of 33rd Conference on Learning Theory (COLT'20)*, pp. 1374–1388.
- [11] Cormode, G. & Muthukrishnan, S. (2005). What's hot and what's not: Tracking most frequent items dynamically. *ACM Transactions on Database Systems (TODS)* 30(1): 249–278.
- [12] Denuit, M., Dhaene, J., & Ribas, C. (2001). Does positive dependence between individual risks increase stop-loss premiums? *Insurance: Mathematics and Economics* 28: 305–308.
- [13] Denuit, M., Lefèvre, C., & Utev, S. (2002). Measuring the impact of dependence between claims occurrences. *Insurance: Mathematics and Economics* 30: 1–19.
- [14] Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics* 14(4): 436–440.
- [15] Du, D. & Hwang, F. (1993). *Combinatorial group testing and its applications*. Series on Applied Mathematics. Singapore: World Scientific.
- [16] Erlich, Y., Gilbert, A., Ngo, H., Rudra, A., Thierry-Mieg, N., Wootters, M., Zielinski, D., & Zuk, O. (2015). Biological screens from linear codes: Theory and tools. *bioRxiv*, p. 035352.
- [17] Esary, J.D., Proschan, F., & Walkup, D.W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics* 38: 1466–1474.
- [18] Fortuin, C.M., Kasteleyn, P.W., & Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics* 22(2): 89–103.
- [19] Gaunt, R.E., Pickett, A.M., & Reinert, G. (2017). Chi-square approximation by Stein's method with application to Pearson's statistic. *The Annals of Applied Probability* 27(2): 720–756.
- [20] Harremoës, P., Johnson, O.T., & Kontoyiannis, I. (2010). Thinning, entropy and the law of thin numbers. *IEEE Transactions on Information Theory* 56(9): 4228–4244.
- [21] Hong, E.S. & Ladner, R.E. (2002). Group testing for image compression. *IEEE Transactions on Image Processing* 11(8): 901–911.

[22] Hwang, F.K. (1972). A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association* 67(339): 605–608.

[23] Johnson, O.T., Aldridge, M.P., & Scarlett, J. (2019). Performance of group testing algorithms with near-constant tests-per-item. *IEEE Transactions on Information Theory* 65(2): 707–723.

[24] Kautz, W.H. & Singleton, R.C. (1964). Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory* 10(4): 363–377.

[25] Luk, H.M. (1994). Stein’s method for the gamma distribution and related statistical applications. PhD thesis, University of Southern California.

[26] Mutesa, L., Ndishimye, P., Butera, Y., Souopgui, J., Uwineza, A., Rutayisire, R., Musoni, E., Rujeni, N., Nyatanyi, T., Ntagwabira, E., Semakula, M., Musanabaganwa, C., Nyamwasa, D., Ndashimye, M., Ujeneza, E., Mwikarago, I., Muvunyi, C., Mazarati, J., Nsanzimana, S., Turok, N., & Ndifon, W. (2021). A strategy for finding people infected with SARS-CoV-2: Optimizing pooled testing at low prevalence. *Nature* 589: 276–280. doi:10.1038/s41586-020-2885-5

[27] Polyanskiy, Y., Poor, H.V., & Verdú, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory* 56(5): 2307–2359.

[28] Ross, N. (2011). Fundamentals of Stein’s method. *Probability Surveys* 8: 210–293.

[29] Ross, N. (2013). Power laws in preferential attachment graphs and Stein’s method for the negative binomial distribution. *Advances in Applied Probability* 45(3): 876–893.

[30] Shevtsova, I. (2011). On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. arXiv:1111.6554.

[31] Weba, M. (1999). Bounds for the total variation distance between the binomial and the Poisson distribution in case of medium-sized success probabilities. *Journal of Applied Probability* 36(1): 97–104.

[32] Wolf, J.K. (1985). Born again group testing: Multiaccess communications. *IEEE Transactions on Information Theory* 31(2): 185–191.

[33] Yelin, I., Aharony, N., Tamar, E.S., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli, A., Gandali, N., Shkedi, O., Hashimshony, T., Mandel-Gutfreund, Y., Halberthal, M., Geffen, Y., Szwarcwort-Cohen, M., & Kishony, R. (2020). Evaluation of COVID-19 RT-qPCR test in multi sample pools. *Clinical Infectious Diseases* 71(16): 2073–2078.

**Appendix A. Proof of Proposition 3.5**

*Proof.* We use Proposition 1 of [18], which shows that it is enough to verify that the  $G_i$  satisfy the so-called FKG condition:

$$\mathbb{P}(\mathbf{G} = \mathbf{x} \vee \mathbf{y})\mathbb{P}(\mathbf{G} = \mathbf{x} \wedge \mathbf{y}) \geq \mathbb{P}(\mathbf{G} = \mathbf{x})\mathbb{P}(\mathbf{G} = \mathbf{y}), \tag{A.1}$$

where  $\vee$  and  $\wedge$  represent the maximum and minimum, respectively.

By independence, we can describe the distribution of  $\mathbf{G}$  conditional on  $M_0$ , the number of negative tests. Recall that  $M_0$  has a binomial distribution with parameters  $T$  and  $q_0$ . For any  $\mathbf{g}$ , we can write

$$\begin{aligned} \mathbb{P}(\mathbf{G} = \mathbf{g}) &= \sum_m \mathbb{P}(M_0 = m) P_m^{w(\mathbf{g})} (1 - P_m)^{n-k-w(\mathbf{g})} \\ &= \sum_m \mathbb{P}(M_0 = m) (1 - P_m)^{n-k} R_m^{w(\mathbf{g})}, \end{aligned}$$

where  $P_m = (1 - p)^m$  and  $R_m = P_m / (1 - P_m)$ . This follows since we can check which tests the defective items appear in first, and then each nondefective item is independently intruding with probability  $P_m$ , since it must not appear in the  $m$  negative tests.

Using this expression, and writing  $\mathbb{P}(\mathbf{G} = \mathbf{x} \vee \mathbf{y})\mathbb{P}(\mathbf{G} = \mathbf{x} \wedge \mathbf{y})$  in the form

$$\left( \sum_m \mathbb{P}(M_0 = m) \mathbb{P}(\mathbf{G} = \mathbf{x} \vee \mathbf{y} \mid M_0 = m) \right) \left( \sum_\ell \mathbb{P}(M_0 = \ell) \mathbb{P}(\mathbf{G} = \mathbf{x} \wedge \mathbf{y} \mid M_0 = \ell) \right)$$

we can verify the FKG condition (A.1) by writing

$$\begin{aligned} & \mathbb{P}(\mathbf{G} = \mathbf{x} \vee \mathbf{y})\mathbb{P}(\mathbf{G} = \mathbf{x} \wedge \mathbf{y}) - \mathbb{P}(\mathbf{G} = \mathbf{x})\mathbb{P}(\mathbf{G} = \mathbf{y}) \\ &= \sum_{m,\ell} \mathbb{P}(M_0 = m)\mathbb{P}(M_0 = \ell)(1 - P_m)^{n-k}(1 - P_\ell)^{n-k} \{R_m^{w(\mathbf{x} \vee \mathbf{y})} R_\ell^{w(\mathbf{x} \wedge \mathbf{y})} - R_m^{w(\mathbf{x})} R_\ell^{w(\mathbf{y})}\} \\ &=: \sum_{m,\ell} \mathbb{P}(M_0 = m)\mathbb{P}(M_0 = \ell)(1 - P_m)^{n-k}(1 - P_\ell)^{n-k} \alpha_{m,\ell}(\mathbf{x}, \mathbf{y}). \end{aligned}$$

We can pair up these  $\alpha_{m,\ell}$  terms, noticing the fact that  $w(\mathbf{x} \vee \mathbf{y}) + w(\mathbf{x} \wedge \mathbf{y}) = w(\mathbf{x}) + w(\mathbf{y})$  means that  $\alpha_{m,m}$  vanishes, and that, in general, we can write

$$\alpha_{m,\ell} + \alpha_{\ell,m} = \frac{1}{R_m^{w_+} R_\ell^{w_+}} (R_m^{w_+} R_\ell^{w(\mathbf{x})} - R_\ell^{w_+} R_m^{w(\mathbf{x})})(R_m^{w_+} R_\ell^{w(\mathbf{y})} - R_\ell^{w_+} R_m^{w(\mathbf{y})}),$$

where for brevity we write  $w_+ = w(\mathbf{x} \vee \mathbf{y})$  and  $w(\mathbf{x} \wedge \mathbf{y}) = w(\mathbf{x}) + w(\mathbf{y}) - w_+$ . Observe that since  $w_+ \geq w(\mathbf{x})$  and  $w_+ \geq w(\mathbf{y})$  both these bracketed terms have the same sign, so  $\alpha_{m,\ell} + \alpha_{\ell,m} \geq 0$  and the FKG condition is satisfied.  $\square$

### Appendix B. Proof of Theorem 3.7

*Proof.* Write  $L = (n - k)$  for the number of nondefective items. We know that

$$\begin{aligned} M_{(s)}(G) &= \frac{L!}{(L - s)!} (1 - q_0(1 - (1 - p)^s))^T, \\ M_{(s)}(Z) &= \frac{\Gamma(r + s)}{\Gamma(r)} \left(\frac{1 - q}{q}\right)^s, \end{aligned}$$

where we write  $q_0 = (1 - p)^k$  for brevity. By moment matching, we know that  $L(1 - q_0p)^T = r(1 - q)/q$ , so the ratio

$$\begin{aligned} \frac{M_{(s)}(G)}{M_{(s)}(Z)} &= \left(\frac{L!}{(L - s)!L^s} \frac{\Gamma(r)r^s}{\Gamma(r + s)}\right) \left(\frac{1 - q_0(1 - (1 - p)^s)}{(1 - q_0p)^s}\right)^T \\ &= \left(\frac{L!}{(L - s)!L^s} \frac{\Gamma(r)r^s}{\Gamma(r + s)}\right) \left((1 - q_0) \left(\frac{1}{1 - q_0p}\right)^s + q_0 \left(\frac{1 - p}{1 - q_0p}\right)^s\right)^T. \end{aligned} \tag{B.1}$$

We will treat the two bracketed (falling factorial and  $T$ th power) terms of (B.1) separately.

#### 1. Falling factorial term.

Note that

$$\frac{L!}{(L - s)!L^s} \frac{\Gamma(r)r^s}{\Gamma(r + s)} = \frac{L(L - 1) \dots (L - s + 1)}{L^s} \frac{r^s}{r(r + 1) \dots (r + s - 1)} \leq 1, \tag{B.2}$$

by a termwise comparison. Similarly, we can bound (B.2) from below using the arithmetic mean-geometric mean inequality as

$$\begin{aligned} \frac{L!}{(L - s)!L^s} \frac{\Gamma(r)r^s}{\Gamma(r + s)} &\geq \left(\frac{L - s}{L}\right)^s \frac{1}{\prod_{i=0}^{s-1} (1 + i/r)} \\ &\geq \left(\frac{L - s}{L}\right)^s \frac{1}{\left(\frac{1}{s} \sum_{i=0}^{s-1} (1 + i/r)\right)^s} = \left(\frac{L - s}{L(1 + (s - 1)/(2r))}\right)^s. \end{aligned}$$

2. **Tth power term.** We can write the second term of (B.1) as  $(1 - R)^T$ , where we write  $\bar{q} = 1 - q_0p$ ,

$$R = (1 - q_0) \left( 1 - \left( \frac{1}{\bar{q}} \right)^s \right) + q_0 \left( 1 - \left( \frac{1-p}{\bar{q}} \right)^s \right). \tag{B.3}$$

We first provide an upper bound on  $R$  using the fact that  $\theta(x) := (1+x)^s - (1+xs + x^2s(s-1)/2 + x^3s(s-1)(s-2)/6) \geq 0$  (this result follows using the fact that  $\theta(0) = \theta'(0)$  and since  $\theta''(x) = s(s-1)((1+x)^{s-2} - x(s-2) - 1) \geq 0$  by Bernoulli's inequality). Equivalently, for any  $x$ , we can write  $1 - (1+x)^s \leq -sx - \frac{1}{2}s(s-1)(x^2 + x^3(s-2)/3)$ . Hence, taking  $x_1 = 1/\bar{q} - 1 = pq_0/\bar{q}$  and  $x_2 = (1-p)/\bar{q} - 1 = -p(1-q_0)/\bar{q}$ , respectively, in the two terms of (B.3) this gives

$$R \leq -\frac{1}{2}s(s-1)Cp^2 \left( 1 - \frac{(s-2)(1-2q_0)p}{3\bar{q}} \right),$$

since the linear terms cancel as  $(1 - q_0)x_1 + q_0x_2 = 0$ , and where we recall that we write  $C = q_0(1 - q_0)/\bar{q}^2$ .

We can give a complementary lower bound on  $R$  using a standard argument based on the mean value theorem with  $f(t) = t^s$  by rewriting (B.3) to obtain

$$\begin{aligned} R &= (1 - q_0) \left( f(1) - f \left( 1 + \frac{pq_0}{\bar{q}} \right) \right) + q_0 \left( f(1) - f \left( 1 - \frac{p(1-q_0)}{\bar{q}} \right) \right) \\ &= -(1 - q_0) \frac{pq_0}{\bar{q}} f'(\beta) + q_0 \frac{p(1-q_0)}{\bar{q}} f'(\alpha) \\ &= -\frac{pq_0(1-q_0)}{\bar{q}} (\beta - \alpha) f''(\gamma), \end{aligned} \tag{B.4}$$

for some  $\alpha \in (1 - p(1 - q_0)/\bar{q}, 1)$ ,  $\beta \in (1, 1 + pq_0/\bar{q})$ , and  $\gamma \in (\alpha, \beta)$ . Then, since we know  $(\beta - \alpha) \leq p/\bar{q}$  and  $\gamma \leq \beta \leq 1 + pq_0/\bar{q}$ , the expression (B.4) gives

$$R \geq -\frac{p^2q_0(1-q_0)}{\bar{q}^2} s(s-1) \left( 1 + \frac{pq_0}{\bar{q}} \right)^{s-2} = -s(s-1)Cp^2 \left( \frac{1}{\bar{q}} \right)^{s-2},$$

meaning that

$$(1 - R)^T \leq \exp(s(s-1)Cp^2Tq_0^{2-s}),$$

and the proof is complete. □

### Appendix C. Proof of Theorem 3.8

In proving the theorem, our strategy will be to replace  $G$  by the mixed Poisson version  $G''$  defined below (bounding the error in making this replacement). We then approximate  $G''$  by a negative binomial distribution by noting that a negative binomial can itself be written as a mixed Poisson with gamma mixing distribution. Lemma C.1 allows us to transfer our negative binomial approximation problem for a mixed Poisson distribution into a gamma approximation problem for the mixing distribution. We may then bound the appropriate distance from our mixing distribution to gamma to complete the proof.

Before proceeding with this programme, we first define a further metric we will need: the Wasserstein distance, denoted by  $d_W$ . For any non-negative, real-valued random variables  $X$  and  $Y$ , we define

$$d_W(X, Y) = \sup_{h \in \mathcal{H}_W} |\mathbb{E}h(X) - \mathbb{E}h(Y)| = \int_0^\infty |\mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x)| dx, \tag{C.1}$$

where  $\mathcal{H}_W$  is the set of absolutely continuous functions  $h : \mathbb{R}^+ \rightarrow \mathbb{R}$  with  $\|h'\| \leq 1$ , and  $\|\cdot\|$  is the supremum norm defined by  $\|g\| = \sup_x |g(x)|$  for any real-valued function  $g$ .

**Lemma C.1.** *Let  $Z$  have a negative binomial distribution with parameters  $r$  and  $q$ , and let  $H$  have a mixed Poisson distribution,  $H|\xi \sim \text{Po}(\xi)$  for some positive random variable  $\xi$ . Let  $\eta \sim \Gamma(r, \lambda)$  have a gamma distribution with density function  $(\lambda^r / \Gamma(r))x^{r-1}e^{-\lambda x}$ , for  $x > 0$ , where  $\lambda = q/(1 - q)$ . Then,*

$$d_{TV}(H, Z) \leq \frac{2 - q}{1 - q} d_W(\xi, \eta).$$

*Proof.* It can be easily checked by direct calculation that  $Z$  has the mixed Poisson distribution  $Z|\eta \sim \text{Po}(\eta)$ . Following Stein’s method for negative binomial approximation (see [7,8,29] and the review in Section 3.4), we let  $f = f_A$  satisfy  $f(0) = 0$  and (see (27))

$$(1 - q)(r + j)f(j + 1) - jf(j) = I(j \in A) - \mathbb{P}(Z \in A),$$

where  $A \subseteq \mathbb{Z}^+$ , so that we may write (see (29))

$$d_{TV}(H, Z) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{E}[(1 - q)(r + H)f(H + 1) - Hf(H)]|.$$

We note the following bounds on  $f$ , taken from Lemma 3 of [8] and Lemmas 2.2 and 2.3 of [29], respectively:

$$\sup_j |f(j)| \leq \frac{1}{1 - q}, \quad |\Delta f(j)| \leq \frac{1}{j}, \quad \sup_j |\Delta(D^{(r)}f)(j)| \leq \frac{2 - q}{(1 - q)r}, \tag{C.2}$$

where  $\Delta f(j) = f(j + 1) - f(j)$  and  $D^{(r)}f(j) = (j/r + 1)f(j + 1) - (j/r)f(j)$ , so that

$$\Delta(D^{(r)}f)(j) = \left(\frac{j + 1}{r} + 1\right) \Delta f(j + 1) - \frac{j}{r} \Delta f(j).$$

Now, we define  $g(x) = (1 - q)\mathbb{E}[f(H + 1) | \xi = x]$ . Using the fact that  $H$  has a mixed Poisson distribution, a direct calculation shows that  $g'(x) = (1 - q)\mathbb{E}[\Delta f(H + 1) | \xi = x]$ , and similarly for the second derivative of  $g$ . We also note that (see p. 12 of [6] for example), since  $H$  has a mixed Poisson distribution,

$$\xi \mathbb{E}[f(H + 1) | \xi] = \mathbb{E}[Hf(H) | \xi]. \tag{C.3}$$

This then allows us to write

$$\begin{aligned} & \mathbb{E}[(1 - q)(r + H)f(H + 1) - Hf(H)] \\ &= \mathbb{E}\mathbb{E}[(1 - q)(r + H)f(H + 1) - Hf(H) | \xi] \\ &= \mathbb{E}\mathbb{E}[(1 - q)r f(H + 1) + (1 - q)\xi f(H + 2) - \xi f(H + 1) | \xi] \\ &= \mathbb{E}[\xi g'(\xi) + (r - \lambda \xi)g(\xi)]. \end{aligned}$$

This latter expression is closely related to Stein’s method for gamma approximation, as developed by Luk [25]; see also [19] and references therein for more recent developments. In particular, it is known

that since  $\eta$  has a gamma distribution,  $\mathbb{E}[\eta g'(\eta) + (r - \lambda\eta)g(\eta)] = 0$ . Letting  $h(x) = xg'(x) + (r - \lambda x)g(x)$  (and noting that our earlier calculations show that  $h$  is differentiable), we may therefore write

$$d_{TV}(H, Z) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{E}h(\xi) - \mathbb{E}h(\eta)| \leq \sup_{A \subseteq \mathbb{Z}^+} \|h'\| d_W(\xi, \eta).$$

To complete the proof, it remains only to bound  $|h'(x)| \leq (2 - q)/(1 - q)$ . To that end, we note that

$$\begin{aligned} h'(x) &= xg''(x) + g'(x) + (r - \lambda x)g'(x) - \lambda g(x) \\ &= (1 - q)(x\mathbb{E}[\Delta^2 f(H + 1) | \xi = x] + (1 + r - \lambda x)\mathbb{E}[\Delta f(H + 1) | \xi = x] \\ &\quad - \lambda\mathbb{E}[f(H + 1) | \xi = x]) \\ &= (1 - q)(\mathbb{E}[H\Delta^2 f(H) | \xi = x] + (1 + r)\mathbb{E}[\Delta f(H + 1) | \xi = x] - \lambda\mathbb{E}[H\Delta f(H) | \xi = x] \\ &\quad - \lambda\mathbb{E}[f(H + 1) | \xi = x]) \\ &= (1 - q)r\mathbb{E}[\Delta(D^{(r)} f)(H) | \xi = x] \\ &\quad - (1 - q)\lambda(\mathbb{E}[H\Delta f(H) | \xi = x] + \mathbb{E}[f(H + 1) | \xi = x]), \end{aligned}$$

where the penultimate inequality again uses (C.3). Using the bounds (C.2), we therefore have

$$|h'(x)| \leq \frac{(1 - q)r(2 - q)}{(1 - q)r} + (1 - q)\lambda \left(1 + \frac{1}{1 - q}\right) = 2 - q + (1 - q)\lambda + \lambda = \frac{2 - q}{1 - q},$$

as required, since  $\lambda = q/(1 - q)$ . □

*Proof of Theorem 3.8.* We now use Lemma C.1 to establish Theorem 3.8. Recalling that  $M_0 \sim \text{Bin}(T, q_0)$ , we define  $M' \sim \text{Po}(Tq_0)$ . Similarly, recalling that  $G$  has the mixed binomial distribution  $G|M_0 \sim \text{Bin}(n - k, (1 - p)^{M_0})$ , we define  $G'$  and  $G''$  as follows:

$$\begin{aligned} G'|M' &\sim \text{Bin}(n - k, (1 - p)^{M'}), \\ G''|M' &\sim \text{Po}((n - k)(1 - p)^{M'}). \end{aligned}$$

We then write

$$d_{TV}(G, Z) \leq d_{TV}(G, G') + d_{TV}(G', G'') + d_{TV}(G'', Z),$$

and bound each of these three terms separately.

Firstly, we note that

$$d_{TV}(G, G') \leq d_{TV}(M_0, M') \leq 2 \min \left\{ \frac{q_0}{4\sqrt{1 - q_0}}, \frac{1}{\sqrt{T}}\alpha(q_0) + \frac{1}{\sqrt{2\pi e}} \log \left( \frac{1}{\sqrt{1 - q_0}} \right) \right\},$$

where the final inequality comes from the main result of Weba [31] combined with the sharpened value 0.4748 of the constant in the Berry–Esseen theorem due to Shevtsova [30].

Secondly, using the fact that  $d_{TV}(\text{Bin}(m, p'), \text{Po}(mp')) \leq p'$  for any parameters  $m$  and  $p'$  (see p. 8 of [6]), we have that

$$d_{TV}(G', G'') \leq \mathbb{E}[(1 - p)^{M'}] = e^{-T p q_0}.$$

To complete the proof, it remains only to bound  $d_{TV}(G'', Z)$ . To that end, we apply Lemma C.1, noting that the parameters  $q$  and  $r$  of  $Z$  are chosen such that the first two moments of  $Z$  match those of  $G''$ : straightforward calculations show that  $\mathbb{E}[G''] = \mu$  and  $\text{Var}(G'') = \sigma^2$ . Lemma C.1 gives

$$d_{TV}(G'', Z) \leq \frac{2 - q}{1 - q} d_W(\xi, \eta),$$



where  $\xi = (n-k)(1-p)^{M'}$  and  $\eta \sim \Gamma(r, \lambda)$  has a gamma distribution with rate parameter  $\lambda = (1-q)/q$ . Using scaling properties of the gamma distribution and the Wasserstein distance, we have that

$$d_w(\xi, \eta) = d_w((n-k)\xi', (n-k)\eta') = (n-k)d_w(\xi', \eta'),$$

where  $\xi' = (1-p)^{M'}$  and  $\eta' \sim \Gamma(r, Kr)$ , with  $K$  as in the statement of the theorem. We then write

$$d_w(\xi', \eta') = \int_0^1 |\mathbb{P}(\xi' > x) - \mathbb{P}(\eta' > x)| dx + \int_1^\infty \mathbb{P}(\eta' > x) dx, \tag{C.4}$$

and bound the two terms on the right-hand side of (C.4) separately. Beginning with the final term of (C.4), we note that, for  $Z \sim \Gamma(\alpha, \beta)$ , a standard Chernoff bounding argument gives us that, for any  $z > \alpha/\beta$  and  $t > 0$ ,

$$\mathbb{P}(Z > z) \leq \frac{\mathbb{E}e^{tZ}}{e^{tz}} = \frac{(1-t/\beta)^{-\alpha}}{e^{tz}} = \left(\frac{\beta e}{\alpha}\right)^\alpha \exp(-\beta z)z^\alpha, \tag{C.5}$$

where we take the optimal choice that  $t = \beta - \alpha/z$ . (Observe that the same argument applies to bound

$$\mathbb{P}(Z < z) \leq \left(\frac{\beta e}{\alpha}\right)^\alpha \exp(-\beta z)z^\alpha, \tag{C.6}$$

for  $z < \beta/\alpha$ , simply by again taking  $t = \beta - \alpha/z < 0$ ). Since  $\eta' \sim \Gamma(r, rK)$ , the expression (C.5) tells us that

$$\mathbb{P}(\eta' > x) \leq (Ke)^r \exp(-Krx)x^r.$$

This allows us to write the final term of (C.4) as

$$\int_1^\infty \mathbb{P}(\eta' > x) dx \leq \frac{e^r \Gamma(r+1)}{r^{r+1}K} \int_1^\infty \frac{(Kr)^{r+1}}{\Gamma(r+1)} x^r \exp(-Krx) dx \tag{C.7}$$

$$\leq \frac{e^r \Gamma(r+1)}{r^{r+1}K} \left(\frac{Kre}{r+1}\right)^{r+1} \exp(-Kr) \tag{C.1}$$

$$= \frac{e^{2r+1} \Gamma(r+1)Kr}{(r+1)^{r+1}} \exp(-Kr) \leq e^{r+1} K^r \exp(-Kr), \tag{C.8}$$

since we recognize the integrand in (C.7) as the density of a  $\Gamma(r+1, Kr)$  random variable and again apply (C.5). The expression (C.8) follows on observing that  $v(r) := e^r \Gamma(r+1)/(r+1)^{r+1} \leq 1$  for  $r \geq 0$ . We can see this, for example, since  $v(0) = 1$ , and  $v(r)$  is decreasing in  $r$  since  $(d/dr) \log v(r) = \psi(r+1) - \log(r+1) \leq 0$ , where  $\psi$  is the digamma function.

Finally, we write the first term of (C.4) as

$$\begin{aligned} \int_0^1 |\mathbb{P}(\xi' > x) - \mathbb{P}(\eta' > x)| dx &= \int_0^1 \left| \mathbb{P}\left(M' < \left\lceil \frac{\log(x)}{\log(1-p)} \right\rceil\right) - \mathbb{P}(\eta' > x) \right| dx \\ &= \int_0^1 \left| \widehat{\Gamma}\left(\left\lceil \frac{\log(x)}{\log(1-p)} \right\rceil, Tq_0\right) - \widehat{\Gamma}(r, Krx) \right| dx. \end{aligned} \tag{C.9}$$

This completes the proof of the theorem. □

**Appendix D. Proof of Proposition 4.2**

*Proof.* For any  $u > 0$  we may write, using Markov's inequality,

$$\begin{aligned} \mathbb{P}(Z \geq g) &\leq \frac{\mathbb{E}[e^{uZ}]}{e^{ug}} = \exp(\log M_Z(u) - ug) \\ &= \exp\left(r \log\left(\frac{q}{1 - (1 - q)e^u}\right) - ug\right), \end{aligned} \quad (\text{D.1})$$

where we use the fact that the moment generating function of the negative binomial distribution  $\text{NB}(r, q)$  is

$$\mathbb{E}[e^{uZ}] = M_Z(u) = \left(\frac{q}{1 - (1 - q)e^u}\right)^r.$$

Direct calculation then gives that the optimal value of  $u$  to substitute is

$$u^* = \log\left(\frac{g}{(g + r)(1 - q)}\right),$$

(note that the assumption  $g > \mathbb{E}Z = r(1 - q)/q$  ensures that  $g/[(g + r)(1 - q)] > 1$  so that  $u^* > 0$  as required in (D.1)). The result follows on substitution in (D.1), since this choice of  $u = u^*$  makes

$$\frac{q}{1 - (1 - q)e^u} = \frac{q(g + r)}{r}.$$

□