

Supplement B. The Conditional First Moment Function When $(1 - q)^k = \rho$

This supplement derives a general first moment function derivation with $(1 - q)^k = \rho \in (0, 1)$.

We consider the derivation of a conditional first moment function from Section 5 with the generalization of

$$(1 - q)^k = \rho \in (0, 1).$$

As this is a heuristic derivation, less emphasis will be placed on providing exact error terms, and we will often replace values such as M, p, q with their large n counterparts.

For large n , and therefore large k , it suffices to consider $q = (1 + o(1)) \log(1/\rho)/k$, meaning $\nu = \log(1/\rho)$. In words, we assume that the probability of a given test being negative is ρ and each individual participates in a group test with probability (approximately) $\log(1/\rho)/k$. In our previous analysis we assumed $\rho = 1/2$.

First, we generalize the large n estimate of M and p , the number of positive tests and possible infected respectively. A modification of Lemma 6.1 reveals that

$$M = (1 + o(1))(1 - \rho) \log_2 \binom{n}{k} = (1 + o(1)) C \frac{(1 - \alpha)(1 - \rho)}{\log(2)} k \log(n).$$

Additionally, by conditioning on this value of M , a further modification of Lemma 6.2 gives

$$\begin{aligned} p &= (1 + o(1))(p - k) = (1 + o(1))(p - k)(1 - q)^{N-M} \\ &= (1 + o(1)) n^{1 + (1 + o(1))(\alpha - 1) \frac{C \rho \log(1/\rho)}{\log(2)}}. \end{aligned}$$

Second, we adjust the conditioning event $\mathcal{A}$ to account for a different choice of $q = \log(1/\rho)/k$. Recall that

$$\mathcal{A} = \{\deg(i) \leq 2aqM : \forall i \in \sigma^*\}.$$

By a union bound, $\mathbb{P}(\mathcal{A}^c) \leq k\mathbb{P}(B > 2aqM)$, where $B \sim \text{Binomial}(M, 2q)$ [13, Section 9.2.1 (ArXiv version)]. A classic Chernoff bound [6] implies that

$$\mathbb{P}(\mathcal{A}^c) \leq k \exp(-MD(2aq||2q)).$$

Then, using [13, Lemma A.4] we have $D(2aq||2q) \geq 2q(a \log(a) - a + 1)$, and thus

$$\mathbb{P}(\mathcal{A}^c) \leq k \exp(-M2q(a \log(a) - a + 1)).$$

Using rough estimates $k = n^\alpha$, $q = \log(1/\rho)/k$, $M = C(1 - \rho)(1 - \alpha)k \log(n)/\log(2)$, we have

(105)

$$\mathbb{P}(\mathcal{A}) = 1 - o(1), \text{ when } a \geq 1 \text{ satisfies } \frac{2C(1 - \rho) \log(1/\rho)}{\log(2)} (a \log(a) - a + 1) > \frac{\alpha}{1 - \alpha}.$$

Next, we derive the general conditional first moment function, specifically conditioning on event $\mathcal{A}$ with a valid choice of a from above. To do so, we adopt the notation from Section 5. Recall the events $E^{\leq s}(\sigma)$, $E^s(\sigma)$, $E_b(\sigma)$, these represent a candidate set σ missing at most s positive tests, exactly s tests, or any given test b . We similarly consider the random variable $Z_{t,\ell}^\sigma$ and null measure $\mathbb{Q}$ —now with the choice of $(1 - q)^k = \rho$.

After observing that $\mathbb{Q}(E^0(\sigma^*)) = (1 - \rho)^M$, the identical planting argument resulting in (21) for measure $\mathbb{Q}$ gives

$$\mathbb{P}(E) = (1 - \rho)^{-M} \mathbb{Q}(E \cap E^0(\sigma^*)).$$

For any k -set σ with $|\sigma \cap \sigma^*| = \ell$, recall $\alpha_s = \mathbb{Q}(E^0(\sigma^*) \cap E^s(\sigma))$ and $\beta_s = \mathbb{Q}(\mathcal{A} | E^0(\sigma^*) \cap E^s(\sigma))$ from Section 5. The sequence of calculations (22)-(25) then produce

$$(106) \quad \mathbb{E}[Z_{t,\ell} | \mathcal{A}] = (1 + o(1)) \binom{k}{\ell} \binom{p - k}{k - \ell} (1 - \rho)^{-M} \sum_{s=0}^t \alpha_s \beta_s.$$

Note, $\alpha_s = \binom{M}{s} \mathbb{Q}((E_b(\sigma))^C \cap E_b(\sigma^*))^s \mathbb{Q}(E_b(\sigma) \cap E_b(\sigma^*))^{M-s}$. The first factor is calculated with

$$(107) \quad \mathbb{Q}((E_b(\sigma))^C \cap E_b(\sigma^*)) = (1-q)^k (1 - (1-q)^{k-\ell}) = \rho(1 - \rho^{1-x}),$$

where $x = \ell/k$. The second factor uses $\mathbb{Q}(E_b(\sigma \cap \sigma^*)) = 1 - (1-q)^\ell = 1 - \rho^x$ and takes value,

$$(108) \quad \begin{aligned} \mathbb{Q}(E_b(\sigma) \cap E_b(\sigma^*)) &= (1 - \rho^x) + \rho^x(1 - \rho^{1-x})^2 \\ &= 1 - 2\rho + \rho^{2-x}. \end{aligned}$$

Equalities (107) and (108) then give,

$$\alpha_s = \binom{M}{s} (\rho(1 - \rho^{1-x}))^s (1 - 2\rho + \rho^{2-x})^{M-s}.$$

Event $\mathcal{A}$ implies that each individual in σ^* participates in at most $2aqM$ tests. Therefore, those individuals in $\sigma \cap \sigma^*$ must have degree $|\sigma \cap \sigma^*| \leq \ell \cdot 2aqM$. Clearly, under $\{E^0(\sigma^*) \cap E^s(\sigma)\}$, $\sigma \cap \sigma^*$ covers at most $M - s$ tests. Using [13, Section 9.2.2 (ArXiv version)] (as in Section 5) with the differing assumption $(1-q)^k = \rho$, we have

$$\beta_s \leq \mathbb{P}(B'_s \leq \ell \cdot 2aqM),$$

where $B'_s \sim \text{Binomial}(M - s, r(x))$ with $r(\ell/k) = r(x) = \frac{\mathbb{Q}(E_b(\sigma \cap \sigma^*))}{\mathbb{Q}(E_b(\sigma) \cap E_b(\sigma^*))} = \frac{1 - \rho^x}{1 - 2\rho + \rho^{2-x}}$.

Continuing, notice that

$$\mathbb{Q}((E_b(\sigma))^C \cap E_b(\sigma^*)) + \mathbb{Q}(E_b(\sigma) \cap E_b(\sigma^*)) = \rho(1 - \rho^{1-x}) + \rho(1/\rho - 2 + \rho^{1-x}) = 1 - \rho,$$

therefore we re-normalize the probabilities in α_s by prefactor $(1 - \rho)^{-M}$ (which, from (106), we already have at our disposal). Thus, with two random variables $B_1 \sim \text{Binomial}(M - t, r(x))$ and $B_2 \sim \text{Binomial}(M, s(x))$, with $s(x) = \frac{\rho}{1 - \rho}(1 - \rho^{1-x})$, we argue identically to (28)-(29), applying our estimates of α_s and β_s , to give

$$\mathbb{E}[Z_{t,\ell} | \mathcal{A}] \leq (1 + o(1)) \binom{k}{\ell} \binom{p-k}{k-\ell} \exp \left(-(M-t)D \left(\frac{\ell \cdot 2aqM}{M-t} \middle| \middle| r(x) \right) - MD \left(t/M \middle| \middle| s(x) \right) \right).$$

Now, the general conditional first moment function is constructed as the solution to

$$1 = \binom{k}{\ell} \binom{p-k}{k-\ell} \exp \left(-(M-t)D \left(\frac{\ell \cdot 2aqM}{M-t} \middle| \middle| r(\ell/k) \right) - MD(t/M \middle| \middle| s(\ell/k)) \right),$$

of variable t with fixed ℓ (and thus x). Taking the logarithm of both sides and substituting $\ell = xk$, $x \in [0, 1]$, $t = yM$, $y \in [0, 1]$, then gives

$$0 = \log \left(\binom{k}{\lfloor xk \rfloor} \binom{p-k}{\lfloor (1-x)k \rfloor} \right) - M(1-y)D \left(\frac{xk \cdot 2aqM}{M(1-y)} \middle| \middle| r(x) \right) - MD(y \middle| \middle| s(x)).$$

Recalling the value of $q = \log(1/\rho)/k$ and rearranging, our resulting general conditional first moment function is

$$(109) \quad \frac{1}{M} \log \left(\binom{k}{\lfloor xk \rfloor} \binom{p-k}{\lfloor (1-x)k \rfloor} \right) = (1-y)D \left(\frac{2a \log(1/\rho)}{1-y} x \middle| \middle| r(x) \right) + D(y \middle| \middle| s(x)).$$

Additionally, we can drop the first term on the right-hand side which come from the conditioning event $\mathcal{A}$ to get a generalized vanilla first moment function (derived under $\rho = 1/2$ in [26]) with

$$(110) \quad \frac{1}{M} \log \left(\binom{k}{\lfloor xk \rfloor} \binom{p-k}{\lfloor (1-x)k \rfloor} \right) = D(y \middle| \middle| s(x)).$$

For $x = 0$, we can easily invert equation (109), being,

$$\frac{1}{M} \log \binom{p-k}{k} = D(y||\rho).$$

Then, using the rough estimates of M, p and k , we can simplify the above left hand side to

$$\frac{k \log \left(\frac{n^{1-(1-\alpha)\frac{C\rho \log(1/\rho)}{\log(2)}}}{n^\alpha} \right)}{\frac{C(1-\alpha)(1-\rho)}{\log(2)} k \log(n)} = \frac{\log(2)(1-\alpha) \left(1 - \frac{C\rho \log(1/\rho)}{\log(2)} \right)}{C(1-\alpha)(1-\rho)} = \frac{\log(2) - C\rho \log(1/\rho)}{C(1-\rho)}.$$

Which then gives the value of $y(0) = y$, up to $1 + o(1)$ multiplicative error, as the solution to

$$(111) \quad \frac{\log(2) - C\rho \log(1/\rho)}{C(1-\rho)} = D(y||\rho).$$

A careful tracking of the errors using Lemma 6.1 and 6.2 would allow this first moment function to be a lower bound on $\phi(\ell)$ given the assumption $(1-q)^k = \rho$. We believe that a proof of the upper bound on $\phi(0)$ using the Paley-Zygmund inequality and finding the derivative of $y(x)$ from (109) at $x = 0$ will suffice to prove b -OGP results similar to the $\rho = 1/2$ case. Below we consider differing value of ρ and visualize the changes the the loss landscape. Figure 9 plots (111) as a function of C and outputting y (note we plot $1 - y$ to represent the proportion of covered tests). Figure 10 plots (109) and (110) as a function of x outputting y .

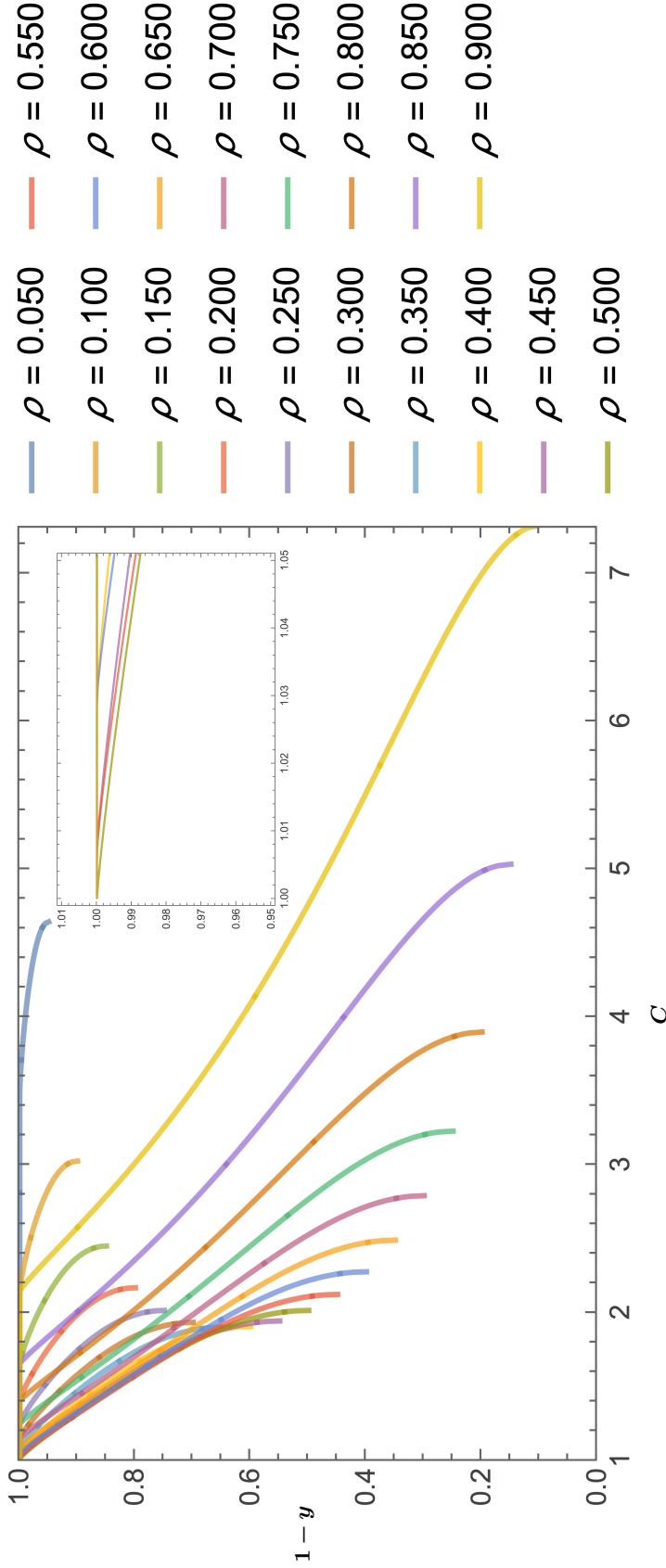


FIG 9. Plots of equation (111) of $1-y$ with respect to C (on the x-axis) and ρ (as the line color) using *Mathematica*. Each of these curves are equivalent to Φ_k from (2) under the assumption that any of the p participants covers any of the M sets with probability $(1 + o(1)) \log(1/\rho)/k$. Under this assumption we have the scaling $p = n^{\frac{1+(\alpha-1)C\rho\log(1/\rho)}{C\rho\log(2)}}$ and $M = C^{\frac{(1-\alpha)(1-\rho)}{\log(2)}} k \log(n)$. See Remark 4.5 for commentary on this scaling for $\rho = 1/2$. Two observations are of note concerning good choices of ρ . First, the value of ρ with the minimal C where the maximum proportion of covering is equal to its average value is not $\rho = 1/2$ but slightly smaller (with $\rho = 1/e$, i.e. $\nu = 1$, being a likely candidate). This suggests that the maximal proportion of covered sets for candidate k -sets σ with zero overlap (or perhaps even small overlap) behave like their expected value. This could suggest that the problem of finding the planted solution may become easier for moderate values of $C \in (1, 2)$ if ρ is decreased from $1/2$. Second, the above choice of M and p scaling induces an interesting symmetry for the minimal value of C where $1-y < 1$ (i.e. the point at which there does not exist a zero overlap k -set covering all M tests) for values of $\rho = \frac{1}{2} \pm \tau$ with $\tau \in (0, 1/2)$. The inlet of this graphic zooms into the plot at location $C = 1$ and $1-y = 1$ to further analyze this threshold. The first departure from the $1-y = 1$ line is at $C = 1$ for $\rho = 1/2$ (in green), then followed with slightly large C by the symmetric pair $\rho \in \{0.45, 0.55\}$ (in purple and red). As the lack of a zero overlap k -set is a necessary condition for σ^* to be a unique maximum of $H(\sigma)$, this suggests that $\nu = \log(2)$, equivalently $\rho = 1/2$, may be the optimal choice for finding a planted signal with C sufficiently close to one.

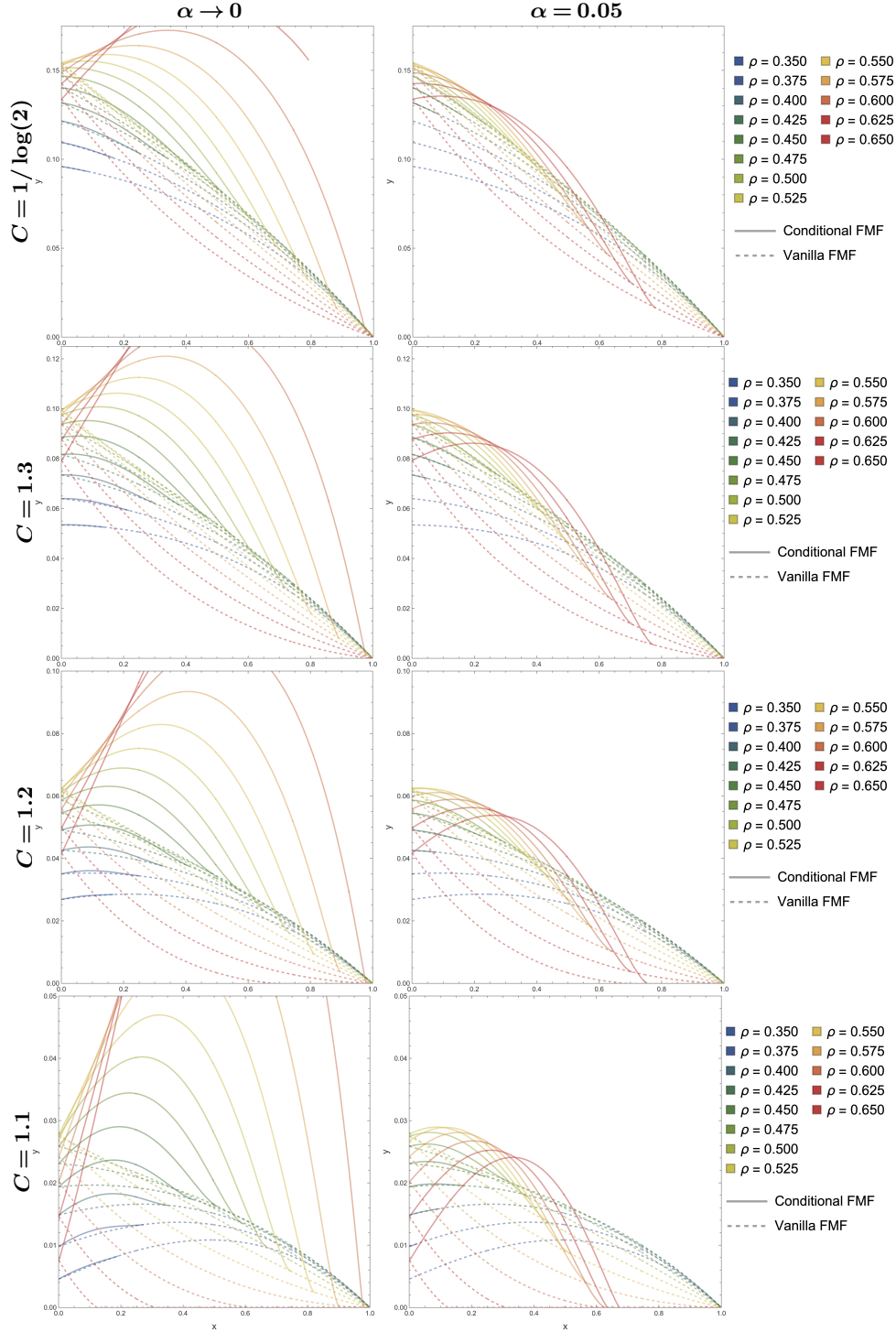


FIG 10. Plots of equations (109) and (110) of $y = t/M$ as a function of $x = \ell/k$ for $(C, \alpha) \in \{1.1, 1.2, 1.3, 1/\log(2)\} \times \{0, 0.05\}$ using Mathematica ($1/\log(2)$ was chosen as [13] found a low-degree lower bound on C for $\rho = 1/2$ in the $n \rightarrow +\infty$ limit). Each plot varies ρ and chooses α as the minimal value for (105) to hold. Solid lines are the conditional solution to (109) and dashed lines are the vanilla solutions to (110). Note, when these lines intersect, the conditional solution no longer exists (additionally, note the upper left plot has further issues of existence). Under the assumption that the conditional solution accurately describe the typical minimum number of missing tests given overlap $\ell = xk$ with σ^* , we make two observations. For smaller values of $C \in \{1.1, 1.2\}$, there is a tradeoff in the monotonicity of the vanilla solution and the conditional solution. With ρ small, the vanilla solution is non-monotonic, then as ρ increases, this vanilla solution regains monotonicity but not before the presence of the conditional solution—which is non-monotonic. Second, for moderate $C \in \{1.3, 1/\log(2)\}$ a slight reduction of ρ induces monotonicity in both the conditional and vanilla solutions. This indicates that a smaller value of ρ , perhaps $\rho = 1/e$ (i.e. $\nu = 1$), can lead to possible gains for our MCMC algorithm.