ECE 236A Project 2 (Group 10)

Sunay Bhat, Samuel Gessow, Steven Zhiying Li, Terri Tsai, Dominic Yang

1 Problem Statement

In this report, we will address various methods to solve the group testing problem.

Assume we are given a population of size n, with infection probability p. We are given a test matrix $X \in \mathbf{R}^{T \times n}$, meaning T tests were conducted for subgroups of the population. $x_{ij} = 1$ if test i includes person j and 0 otherwise. $y \in \mathbf{R}^T$ represents the binary test result.

We will consider three cases:

1. **No Community Structure**: We assume no knowledge of community structure and perfect results. (*Sec* 2)
2. **Community structure**: We assume some knowledge of community structure and perfect results. (*Sec* 3)
3. **Noisy decoding**: We assume test results are imperfect for each of the prior cases. (*Sec* 4)

2 No Community Structure

Let $\tilde{z} \in \mathbf{R}^n$ be a binary vector predicting which people are infected. We want to smallest satisfying set (SSS). We can arrive at an ILP (*omited*). The following is the associated relaxed LP:

$$\begin{aligned} \min_{\tilde{z}} \quad & \sum_{j=1}^{n} \tilde{z}_j \\ \text{subject to} \quad & \sum_{j=1}^{n} x_{ij}\tilde{z}_j \geq 1, \text{ if } y_i = 1 \\ & \sum_{j=1}^{n} x_{ij}\tilde{z}_j = 0, \text{ if } y_i = 0 \\ & \mathbf{0} \leq \tilde{z} \leq \mathbf{1} \end{aligned}$$

This formulation produces real solutions and so we must use a rounding scheme to determine infected members.

2.1 Rounding Heuristics

2.1.1 Thresholding

The most basic rounding method is rounding each $\tilde{z}_i$ to its closest integer with simple rounding. We can generalize this method by replacing 0.5 with some threshold τ.

$$z_i = \begin{cases} 1 & \text{if } \tilde{z}_i \geq \tau \\ 0 & \text{if } \tilde{z}_i < \tau \end{cases}$$

Note that as τ increases, fewer people will be marked as positive. So for high τ, we should expect more false negatives, and for low τ, we should expect more false positives.

However, thresholding is not guaranteed to give a feasible solution to the ILP, this motivated us to create a method that we call *dynamic rounding*.

2.1.2 Dynamic Rounding

To produce a threshold for which the rounded solution is feasible, we pick the highest threshold that will achieve a feasible solution. Denote the solution thresholded at τ by $round(\tilde{z}, \tau)$.

$$\tau^* = \sup\{\tau \in [0,1] : round(\tilde{z}, \tau) \text{ is feasible}\}$$

We compute this τ^* by ordering the relaxation values $\tilde{z}$ in descending order, and then iterating through each of them, setting the threshold at that relaxation value and checking if the rounded solution is feasible. Once we find a feasible solution, we stop, and the relaxation value that we are on is the computed threshold τ^*.

We can also introduce a notion of satisfied constraints. We state that a constraint corresponding to a positive test is already satisfied if we have identified one person in the test group as positive. Negative tests are automatically satisfied as we can immediately identify anyone in those tests as negative.

Then we may proceed as above with the ordering scheme, but if at any point we identify someone as appearing only in already satisfied constraints, we can mark them as negative and the solution obtained will remain feasible.

Through testing, we concluded that for the no noise cases, dynamic rounding is a good rounding method. When compared to thresholding, dynamic rounding gives us similar HD error but slightly better FN errors. (figure B.1)

2.2 Results and Comments

We perform tests using the above method (*relaxed LP* with *dynamic rounding*) to get the False Positive (FP), False Negative (FN), Hamming error rate when $n = 1000$, and $p = 0.1$ and 0.2 in figure A.1.

Comments are followed after each plots in the Appendix.

3 Community Structure

In this section, we take the community structure into account. We assume a partition of the population into F numbers of families $\{f_1, \ldots, f_m\}$, and we assume each family is infected with probability p. In addition, if a family is infected, then any individual in the family is infected the with probability $q \gg p$. This assumption is imposed due to the fact that people within the same family are more likely to infect each other.

3.1 Linear Programs: *Alternative LP*

We developed two methods to solve this problem. Method 1 creates a new matrix $X_F \in \mathbf{R^{n \times F}}$ based on X, and then uses a F-dim binary indicator vector to determine whether a family contain any infected individuals. This method was better than no consideration for community structure, but our subsequent method preformed better, so we did not apply it to our design in the end. (details of method 1 is described in Appendix C, performance of method 1 versus others is demonstrated in Figure. B.2).

We developed Method 2, *Alternative LP*, by combining method 1 with the LP described in section 2 into a single linear program. This was done by introducing binary variables for each family indicating if it contains an infected person or not. We then attached a regularizing constant R in the objective function to bias towards making as few families infected as possible when R is large.

$$
\begin{aligned}
\min_{z} \quad & \sum_{j=1}^{n} z_j + R \sum_{k=1}^{m} w_k \\
\text{subject to} \quad & \sum_{j=1}^{n} x_{ij} z_j \geq y_i, \quad \text{if } y_i = 1 \\
& \sum_{j=1}^{n} x_{ij} z_j = 0, \quad \text{if } y_i = 0 \\
& z_j \leq w_k, \quad \forall j \in F_k \\
& z \in \{0,1\}^n, \; w \in \{0,1\}^m
\end{aligned}
$$

Again, we relax this into a non-integer LP and then use *dynamic rounding* to determine which individuals are infected.

Testing this against the family structure method 1 and our no community base LP, we see a noticeable improvement. Method 2 is what we decided to use on family structures. We observed that the larger the constant R is, the better our estimation gets. This makes sense intuitively, as we are introducing a larger penalty for infecting a new family. Here, we set $R = 20$. (See figure B.2)

3.2 Results and Comments

We perform tests using method above (*alternative LP* and *dynamic rounding*) to get the False Positive (FP), False Negative (FN), Hamming error rate when $n = 1000$, and $p = 0.1, q = 0.1$ or $p = 0.6, q = 0.2$ for (case 1) figure A.2 and (case 2) figure A.3. Comments are followed after each plots in the Appendix.

4 Noisy Decoding

4.1 No Community Structure, with Noise

In realistic testing environments, we cannot expect that the test results are fully accurate. To account for this, we assume that every positive test is flipped to negative with probability p^- and that any negative test is flipped to positive with probability p^+ (For z-channel noise, $p^+ = 0$, for symmetric noise, $p^- = p^+$).

We adjusted our ILP to account for this possibility by introducing binary variables ξ^-, ξ^+ for each test which indicate whether the test was incorrect.

$$\begin{aligned}
\min \sum_{j=1}^{n} z_j + \zeta^- \sum_{i:y_i=0} \xi_i^- + \zeta^+ \sum_{i:y_i=1} \xi_i^+ \\
\text{subject to } \sum_{j=1}^{n} x_{ij} z_j \geq 1 - \xi_i^+, \quad \text{if } y_i = 1 \\
\sum_{j=1}^{n} x_{ij} z_j \geq \xi_i^-, \qquad \text{if } y_i = 0 \\
z_j, \xi_i^+, \xi_i^- \in \{0, 1\}
\end{aligned}$$

We then relaxed this ILP into a LP, and returned integer solutions via threshold rounding at 0.5 to determine which individuals are infected.

Note we have two different weights for positive and negative tests ζ^- and ζ^+, which must be chosen to correctly penalize the cost of an incorrect false positive or false negative test. Note that if only false positives or only false negatives are possible, we only need to include one of the sets of variables $\xi^+ or \xi^-$. In the z-channel-noise case, we only take into account the slack variables ξ^-.

Through testing, we concluded that the best results are $\zeta^- \geq 1$ and $\zeta^+ \geq 1$. We set ours to be $\zeta^- = 1$ and $\zeta^+ = 1$. (See figures B.3, B.4)

4.2 Community Structure, with Noise

To improve estimation results in the presence of noisy inputs, we can implement family structure and solve the following LP and use *thresholding* at 0.5 to determine infected individual.

$$\begin{aligned}
\min_z \sum_{j=1}^{n} z_j + R \sum_{k=1}^{m} w_k + \zeta^- \sum_{i:y_i=0} \xi_i^- + \zeta^+ \sum_{i:y_i=1} \xi_i^+ \\
\text{subject to} \sum_{j=1}^{n} x_{ij} z_j \geq y_i - \xi_i^+, \text{ if } y_i = 1 \\
\sum_{j=1}^{n} x_{ij} z_j \geq \xi_i^-, \qquad \text{if } y_i = 0 \\
z_j \leq w_k, \qquad \forall j \in F_k \\
z_j, w_k, \xi_i^+, \xi_i^- \in \{0, 1\}
\end{aligned}$$

This LP is derived from the method 2 (in section 3) with slack variables ζ^- and ζ^+ added to account for noise. Again, note that the above LP is for the symmetrical-noise case, and that we only take into account the slack variables ξ^- in the z-channel noise case.

Based on testing, we set our constants here to be $R = 30$, $\zeta^- = 1$ and $\zeta^+ = 100$ for symmetric and $R = 40$, $\zeta^- = 1$ for z-channel. We also found in the symmetrical-noise case that adding a constraint which lower bounds the number number of infected persons by the expected amount of infections finds improvement in each of the metrics. In this case, we added the constraint $\sum_j z_j \geq np^*$ where p^* is given by (1) in the appendix.

4.3 Results and Comments

We perform tests using the above methods in Section 4 to get the False Positive (FP), False Negative (FN), Hamming error rate when $n = 1000$, and $p_{\text{noisy}} = 0.1$ or 0.2 for (z-channel noise) figures A.4, A.5 and (basic symmetric channel noise) A.6, A.7. Comments are followed after each plots in the Appendix.

Appendix

A Simulation Results

A.1 No Community Structure

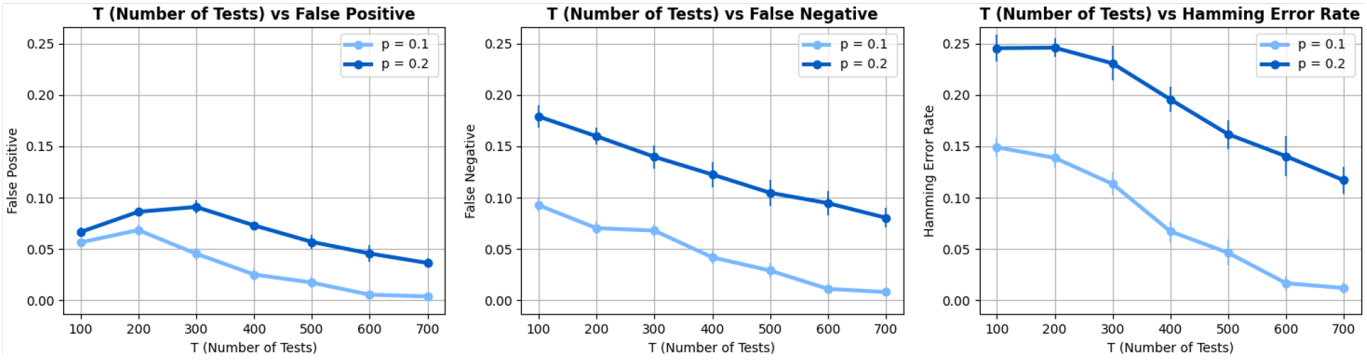

Figure A.1: FP, FN, Hamming error rate for group testing *without* community structure

Comments: This is the baseline LP results against two infection rates with no noise or community structure given. We found a significant improvement when using a dynamic rounding scheme described in the report above over *thresholding* at 0.5.

A.2 Community Structure

A.2.1 Case 1 ($F = 200$)

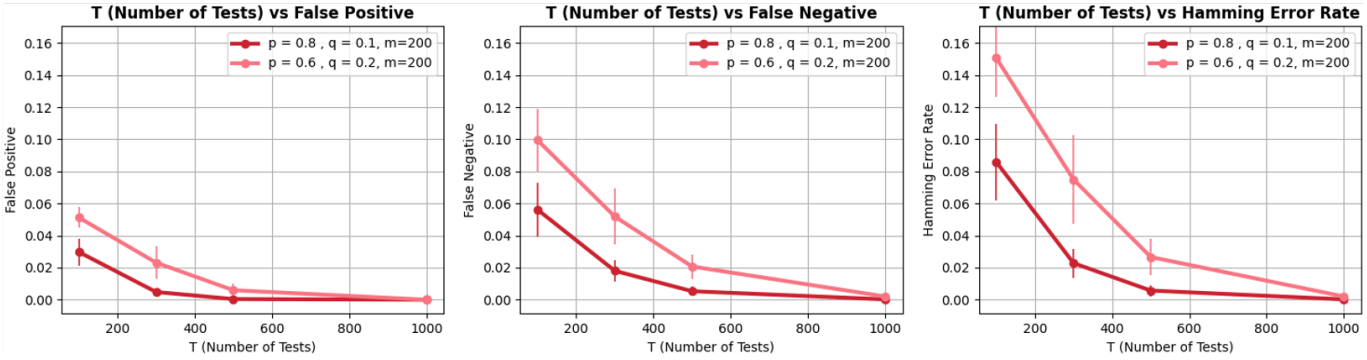

Figure A.2: FP, FN, Hamming error rate for group testing *with* community structure (Case 1).

Comments: As seen above considering the community structure decreases both the amount of false positives and negatives as well as the overall hamming rate. We also found a significant improvement by using dynamic rounding over the simple *thresholding* at 0.5.

A.2.2 Case 2 ($F = 20$)

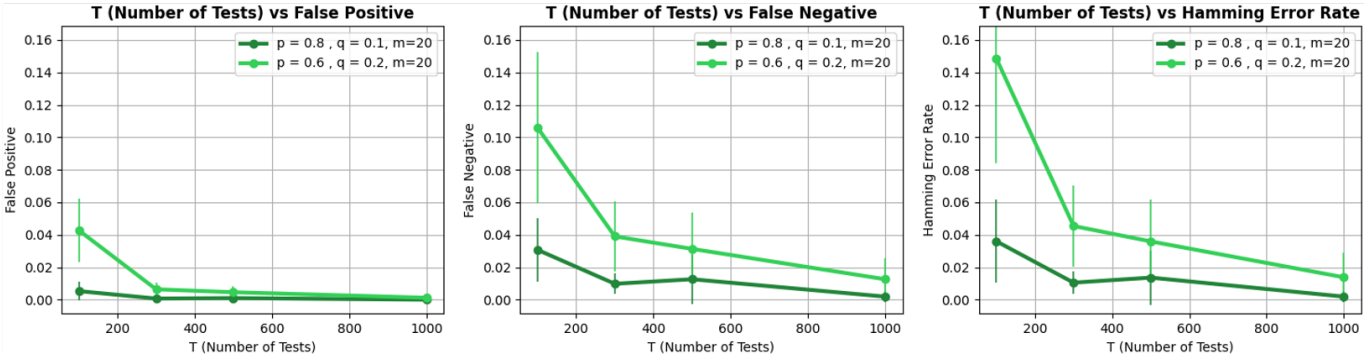

Figure A.3: FP, FN, Hamming error rate for group testing *with* community structure (Case 2)

Comments: Decreasing the number of families does not change the results significantly. However in the the very extreme cases where $F = 1$ or $F = n$ the performance would be identical to not considering community structure at all.

A.3 Noisy Decoding

A.3.1 Assuming z-channel noise

We experiment with two different flipping over probabilities $p_{\text{noisy}} = 0.1$ and $p_{\text{noisy}} = 0.2$.

($\mathbf{p_{\text{noisy}} = 0.1}$)

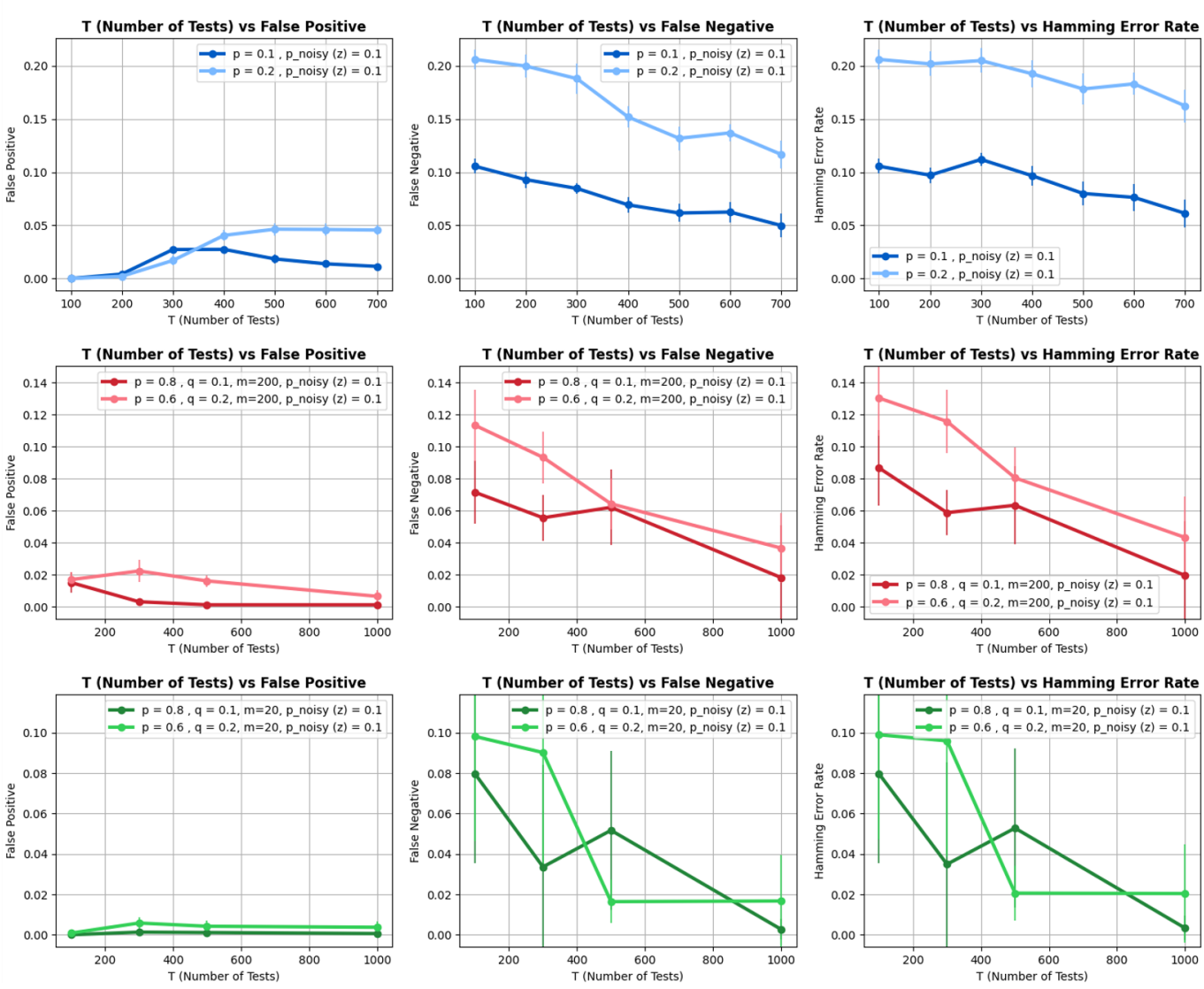

Figure A.4: FP, FN, Hamming error rate for group testing under z-channel noise with $p_{\text{noisy}} = 0.1$; **(blue)** no community structure, **(red)** community structure (case 1), **(green)** community structure (case 2)

Comments: The above plots show z-channel noisy decoding with a 0.1 flipping probability of results. The first set of plots is the standard LP with no knowledge of community structure. We see a significant improvement in both community structure hamming distances, false positives, and false negatives (all errors), when we use the *alternate LP* with z-channel noise slack variables introduced to allow for false negative flips.

($\mathbf{p}_{\text{noisy}} = \mathbf{0.2}$)

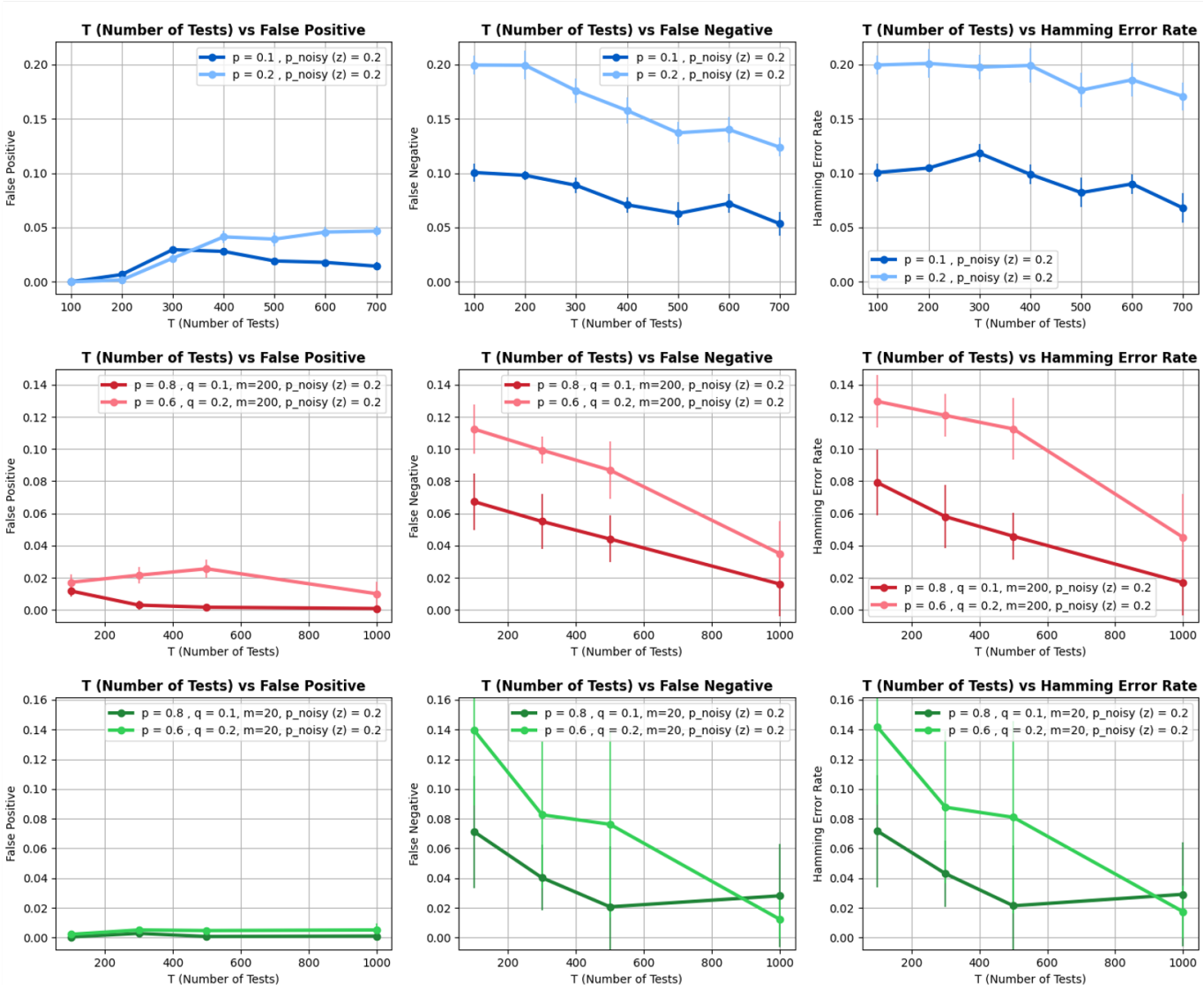

Figure A.5: FP, FN, Hamming error rate for group testing under z-channel noise with $p_{\text{noisy}} = 0.2$; **(blue)** no community structure, **(red)** community structure (case 1), **(green)** community structure (case 2)

Comments: The improvement in error reduction is consistent in the above plots at a noise corruption probability of 0.2 for the alternate LP that considers community structure.

A.3.2 Assuming binary symmetric channel noise

We experiment with two different flipping over probabilities $p_{\text{noisy}} = 0.1$ and $p_{\text{noisy}} = 0.2$.

($\mathbf{p_{\text{noisy}} = 0.1}$)

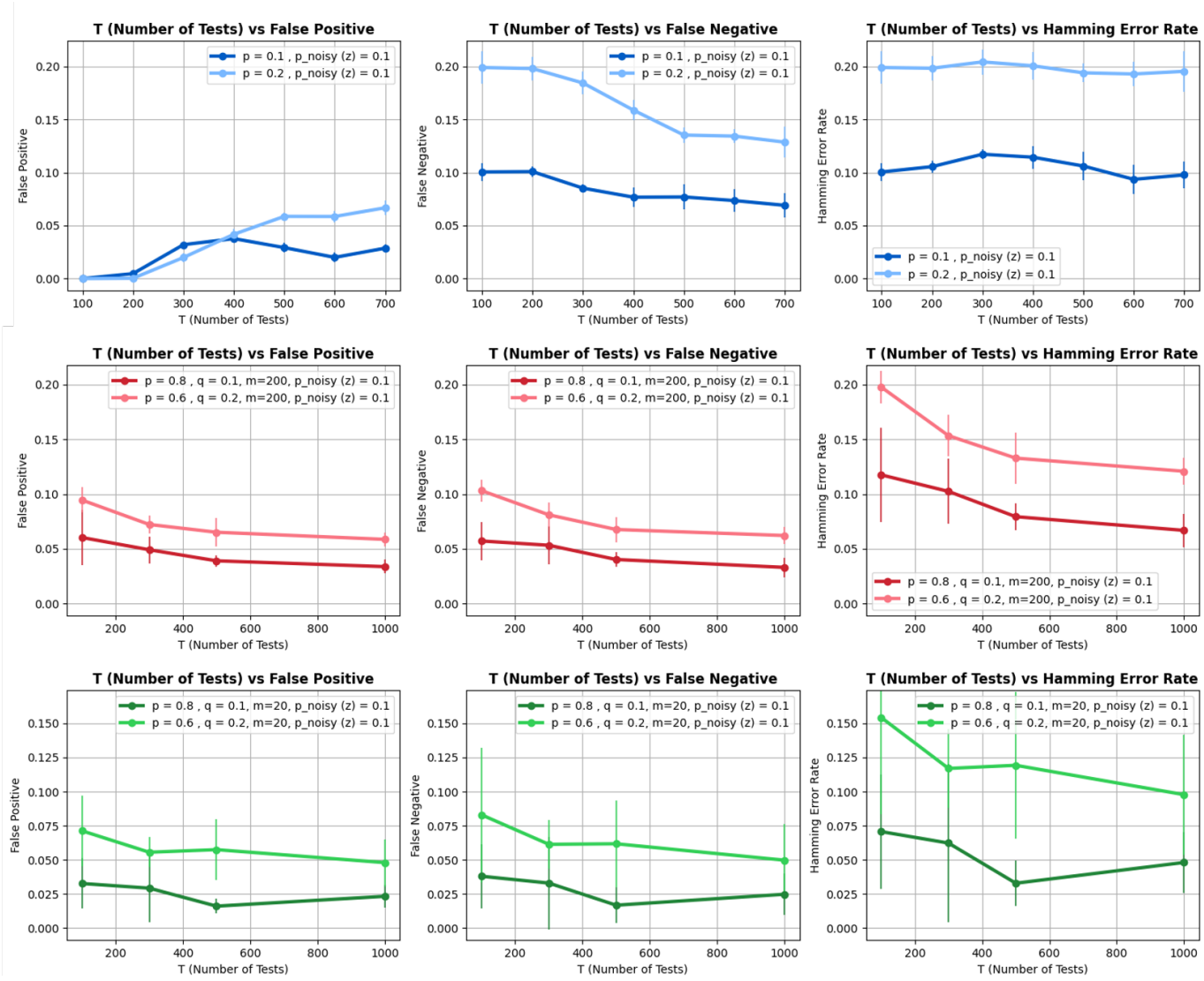

Figure A.6: FP, FN, Hamming error rate for group testing under binary symmetric channel noise with $p_{\text{noisy}} = 0.1$; **(blue)** no community structure, **(red)** community structure (case 1), **(green)** community structure (case 2)

Comments: The symmetric noise case was more nuanced in which the improvement for the alternate LP considering group structure was varied based on number of tests T, number of families M, and number of people n (not changed in project scope). In the above plots with the weights chosen, in particular of the $m = 20$ case ,there is noticeable improvement in hamming error rate and false negatives, and a marginal improvement in the $m = 200$ case in false negatives and in Hamming as T increases. All of this showed that trades could be made by conditioning weights on the input variables of number of tests, number of people, and number of people per family, but we were unable to explore these relationships any deeper for this project.

$(\mathbf{p}_{\text{noisy}} = \mathbf{0.2})$

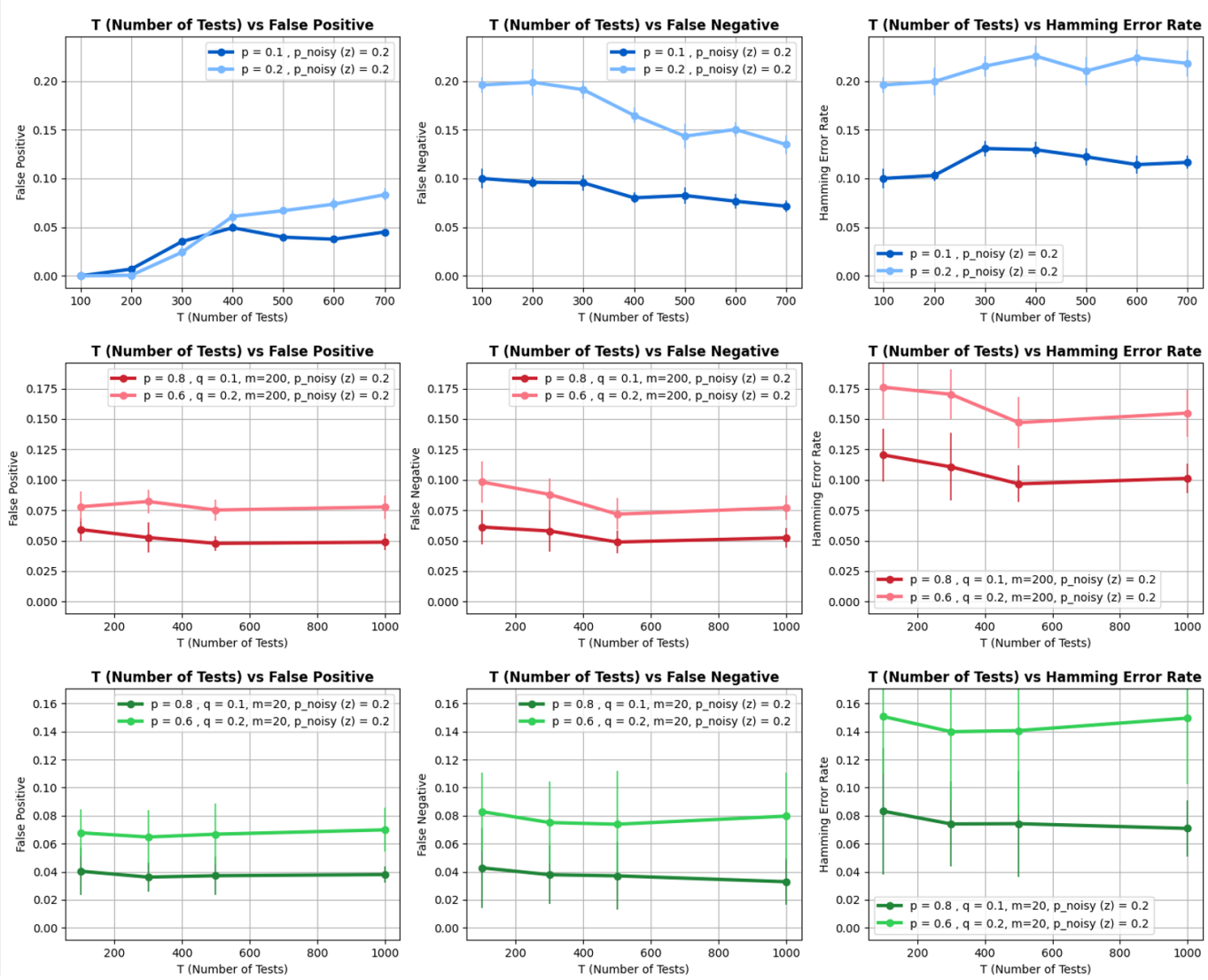

Figure A.7: FP, FN, Hamming error rate for group testing under binary symmetric channel noise with $p_{\text{noisy}} = 0.2$; **(blue)** no community structure, **(red)** community structure (case 1), **(green)** community structure (case 2)

Comments: For $p_{\text{noisy}} = 0.2$, the results are consistent with the above cases. We observe that the error rate in all cases increases, which is to be expected.

B Simulations to get best Hyper-parameters

B.1 Dynamic Rounding vs. Thresholding

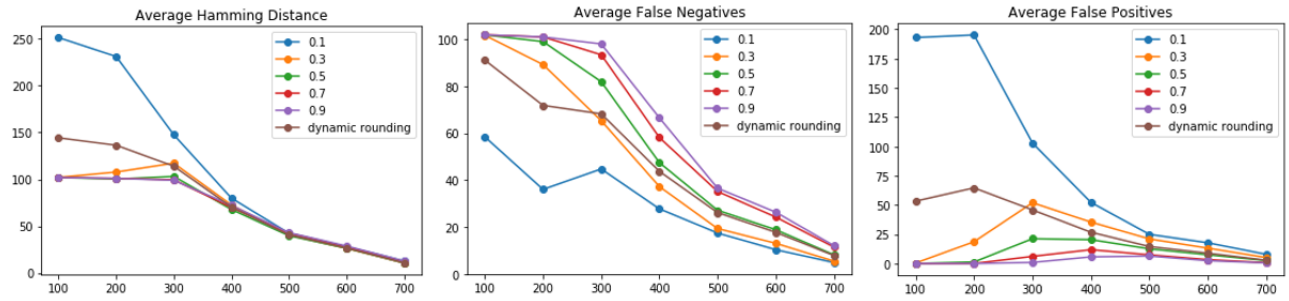

Figure B.1: Dynamic rounding performance vs thresholding performance, $n = 1000$ $p = 0.1$

B.2 Community Structure: Alternative LP vs. Others

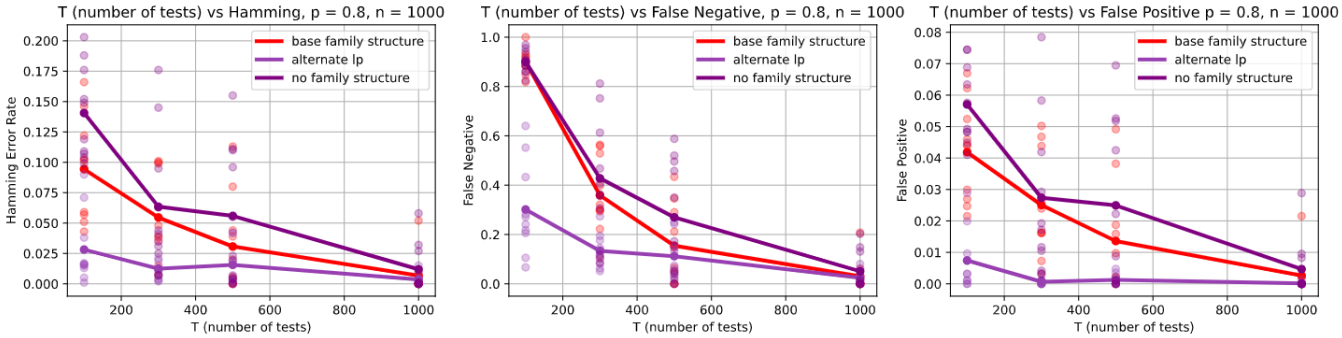

Figure B.2: Family Method 1 LP vs Family Method 2 LP vs Base LP for $n = 1000$, $m = 50$, $p = 0.8$, $q = 0.1$

B.3 Noisy Decoding: Choice of ξ^+ and ξ^-

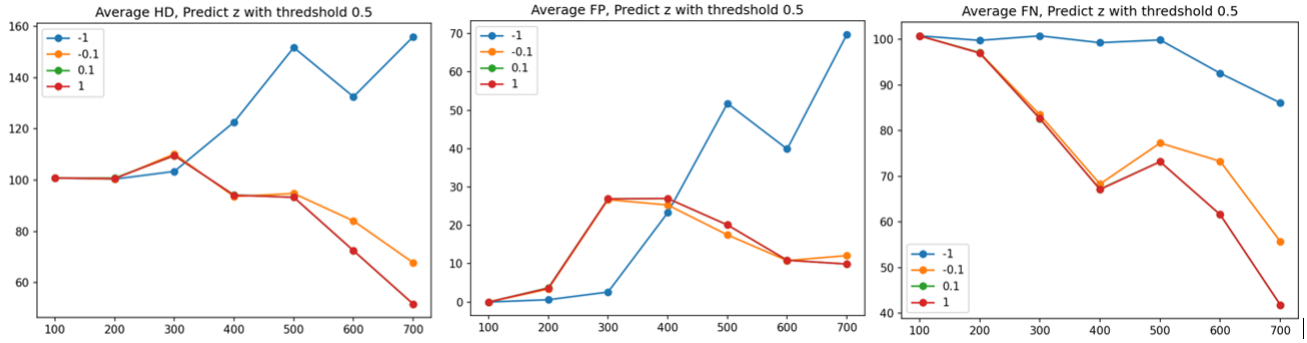

Figure B.3: Testing various ζ^- values for z-channel noise case, $n = 1000$ $p = 0.1$

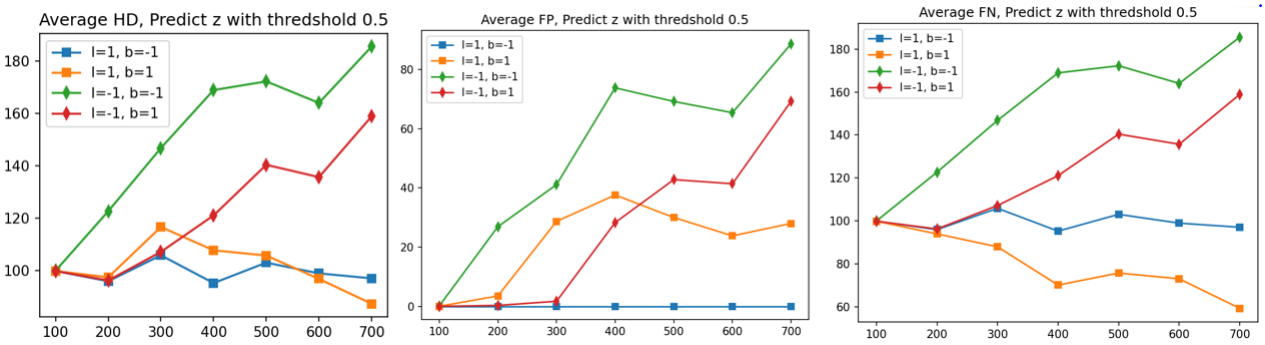

Figure B.4: Testing various ζ^- and ζ^+ values for symmetrical noise case, $n = 1000$ $p = 0.1$

C Community Structure: Method 1 of LP Formulation

Given a family structure, we can then produce an associated problem where instead of determining which persons are infected, we are interested in predicting which families contain an infected individual. In pursuit of this goal, we can construct a new test matrix X_F from the original test matrix X in the following manner:

$$(X_F)_{tf} = \begin{cases} 1 & \exists j \in f : X_{tj} = 1, j \text{ not DND} \\ 0 & \text{otherwise} \end{cases}$$

This test matrix indicates for each test which families could be responsible for the positive result.

We can then have the following integer program which operates instead on families:

$$\begin{aligned} &\min_{z_f} \sum_{j=1}^{n} z_{f,j} \\ &\text{subject to} \sum_{j=1}^{n} x_{F,ij} z_{f,j} \geq y_i, \qquad \text{if } y_i = 1 \\ &z_f \in \{0,1\}^n \end{aligned}$$

Note we remove any constraint corresponding to a negative test because we cannot guaranteed that if there is someone in the family who tested negative for a given test that the entire family is negative.

We then naturally consider the associated relaxation, and can use any rounding scheme from above to produce a prediction on which families contain an infected individual.

Given a prediction z_f, we can produce a prediction for infected individuals by setting $z_i = 0$ if the family f that i is in has $z_f = 0$. After setting these values, we can use again any rounding scheme from above to predict the remaining individuals. We chose to round with *dynamic rounding* in this method.

D Probabilistic Estimates

If we know the infection rate p, we can predict how many positive tests we should expect. Suppose that each person is tested K times in T tests. The probability that any given test is negative is equal to the probability that each person is either negative or does not appear in the test. This is equal to the 1 minus the probability that a person is positive and appears in the test, pK/T. Hence the the probability that the entire test is negative is

$$P(\text{Test } i \text{ is negative}) = \left(1 - \frac{pK}{T}\right)^n$$

Then the expected number of negative tests is given by

$$\mathbb{E}[N] = T\left(1 - p\frac{K}{T}\right)^n$$

and the expected number of positive tests is given by

$$\mathbb{E}[P] = T - T\left(1 - p\frac{K}{T}\right)^n$$

Using this, we can derive an estimate on the infection rate from a count of positive tests P by solving for p. If we do this, we get

$$p = \frac{T}{K}\left(1 - \left(1 - \frac{M}{T}\right)^{1/n}\right) \tag{1}$$

An attempt was made to estimate the noise using the expected value of false positives and negatives and lower bound the number of y flips using these values in both z and bi-directional noise.

Then the expected number of false negatives is given by $p_{noisy} \cdot \mathbb{E}[P]$ and the expected number of false positives is given by $p_{noisy} \cdot \mathbb{E}[N]$. However, without knowing the values of the 2-3 sources of noise, there was too much variation in the expected values of false negatives and positives and thus the bounding constraints on flips had little impact.