Group testing via zero-error channel capacity

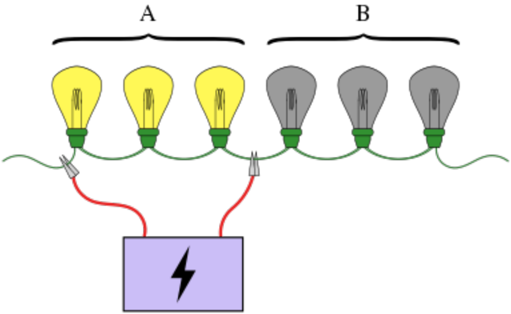
Sam Florin Matthew Ho Rahul Thomas
Mentor: Dr. Zilin Jiang

Greenwich High School, Palo Alto High School, Cherry Creek High School

October 17-18, 2020
MIT PRIMES Conference

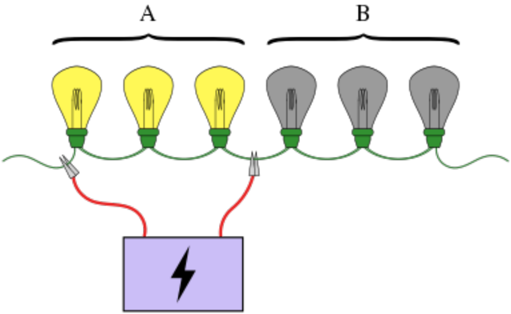
Introduction to Group Testing

- What is group testing? [1]



[1] This image is from Wikimedia Commons, by CheChe

- What is group testing? [1]



- Origin of group testing

[1] This image is from Wikimedia Commons, by CheChe

Single Item Group Testing

Natural question: n people we want to test for COVID-19, there is 1 sick person. How many tests needed?

Single Item Group Testing

Natural question: n people we want to test for COVID-19, there is 1 sick person. How many tests needed?

- Binary search gives us an algorithm that requires $\lceil \log_2(n) \rceil$ steps.

Single Item Group Testing

Natural question: n people we want to test for COVID-19, there is 1 sick person. How many tests needed?

- Binary search gives us an algorithm that requires $\lceil \log_2(n) \rceil$ steps.
- This turns out to be optimal.

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

- Idea: What if we just selected half of these n people? If both or none of the infected people are in this subset, we recurse. Otherwise, do binary search on each half. This gives an upper bound of about $2\log_2(n)$.

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

- Idea: What if we just selected half of these n people? If both or none of the infected people are in this subset, we recurse. Otherwise, do binary search on each half. This gives an upper bound of about $2\log_2(n)$.
- Is there some way to "parallelize" the search in the two halves?

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

- Idea: What if we just selected half of these n people? If both or none of the infected people are in this subset, we recurse. Otherwise, do binary search on each half. This gives an upper bound of about $2\log_2(n)$.
- Is there some way to "parallelize" the search in the two halves?

Our approach

- Connection to binary adder channel, and characterization of channel via an optimization problem;

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

- Idea: What if we just selected half of these n people? If both or none of the infected people are in this subset, we recurse. Otherwise, do binary search on each half. This gives an upper bound of about $2\log_2(n)$.
- Is there some way to "parallelize" the search in the two halves?

Our approach

- Connection to binary adder channel, and characterization of channel via an optimization problem;
- Simplify and numerically solve the optimization problem.

2 Item Group Testing

Next natural question: what happens if 2 of n people are sick?

- Idea: What if we just selected half of these n people? If both or none of the infected people are in this subset, we recurse. Otherwise, do binary search on each half. This gives an upper bound of about $2\log_2(n)$.
- Is there some way to "parallelize" the search in the two halves?

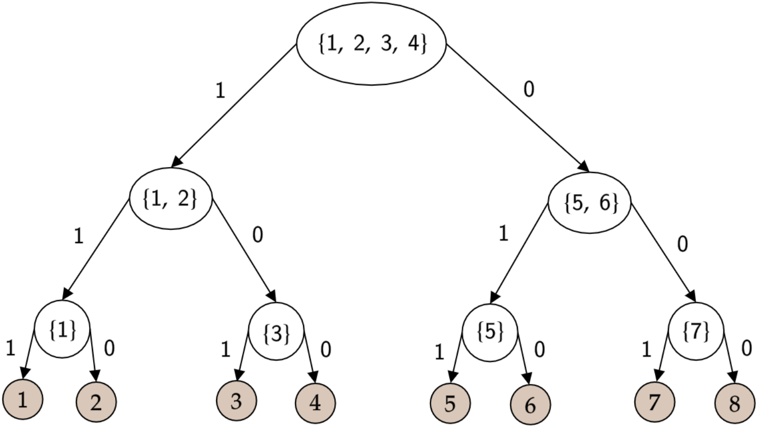
Our approach

- Connection to binary adder channel, and characterization of channel via an optimization problem;
- Simplify and numerically solve the optimization problem.

Main result The best algorithm gives about $1.266\log_2(n)$ tests.

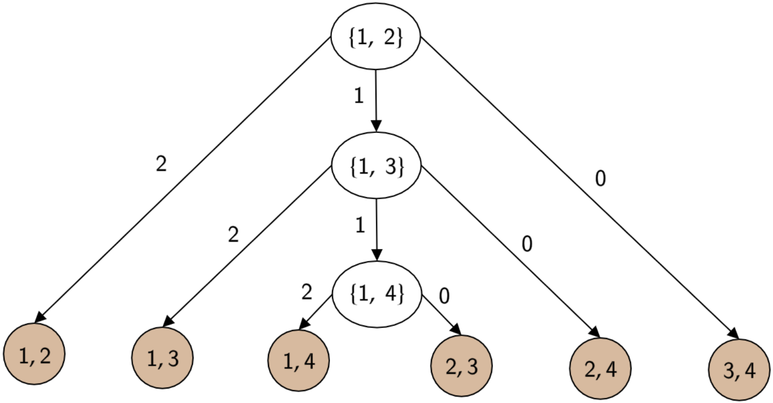
Binary Decision Tree for Single Item Testing

Testing procedure for $\{1, 2, \ldots, 8\}$

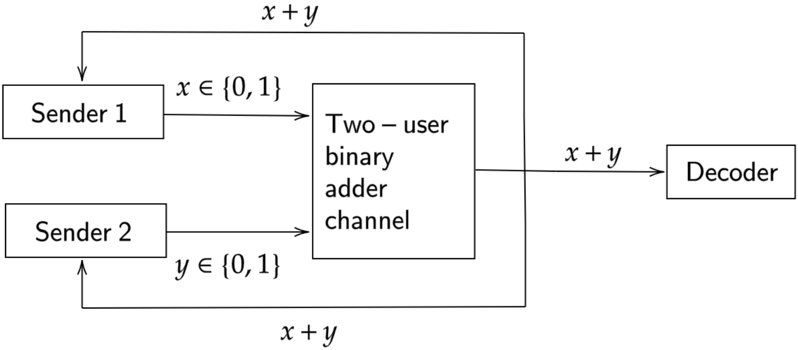


Ternary Decision Tree for Double Item Testing

Testing procedure for $\{1, 2, 3, 4\}$



Two-User Binary Adder Channel with Feedback



Efficiency of the Channel

Definition

The *channel capacity* c is the maximum rate at which bits of information can be sent across the channel.

Efficiency of the Channel

Definition

The *channel capacity* c is the maximum rate at which bits of information can be sent across the channel.

- The optimal number of pooled tests is $\sim 2 \log_2 n/c$.

Efficiency of the Channel

Definition

The *channel capacity* c is the maximum rate at which bits of information can be sent across the channel.

- The optimal number of pooled tests is $\sim 2\log_2 n/c$.
- Computing c is equivalent to an optimization problem involving the entropy function

$$H(x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} x_i \log_2 x_i$$

and the related function

$$L(x) = H\left(\frac{1-x}{2}, x, \frac{1-x}{2}\right).$$

Channel Optimization Problem

Using bounds on channel capacity by Dueck, the group testing problem is equivalent to the following optimization problem:

Channel Optimization Problem

Using bounds on channel capacity by Dueck, the group testing problem is equivalent to the following optimization problem:

Maximize: $\sum_{i=1}^{n} p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$ given

Channel Optimization Problem

Using bounds on channel capacity by Dueck, the group testing problem is equivalent to the following optimization problem:

Maximize: $\sum_{i=1}^{n} p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$ given

- $0 \leq a_i, b_i, p_i \leq 1$,

Channel Optimization Problem

Using bounds on channel capacity by Dueck, the group testing problem is equivalent to the following optimization problem:

Maximize: $\sum_{i=1}^{n} p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$ given

- $0 \leq a_i, b_i, p_i \leq 1$,
- $\sum_{i=1}^{n} p_i = 1$

Channel Optimization Problem

Using bounds on channel capacity by Dueck, the group testing problem is equivalent to the following optimization problem:

Maximize: $\sum_{i=1}^{n} p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$ given

- $0 \leq a_i, b_i, p_i \leq 1,$
- $\sum_{i=1}^{n} p_i = 1$
- $L\left(\sum_{i=1}^{n} p_i(a_i\bar{b}_i + \bar{a}_i b_i + 2c_i)\right) \geq \sum_{i=1}^{n} p_i H\left(a_i b_i - c_i, a_i\bar{b}_i + c_i, \bar{a}_i b_i + c_i, \bar{a}_i\bar{b}_i - c_i\right)$ for all points $\boldsymbol{c} = (c_1, ..., c_n)$ that make the terms inside $H(\cdot, \cdot, \cdot, \cdot)$ non-negative.

Bounding n

- Is there some n_0 such that, for $n > n_0$, the maximum achieved in the optimization problem doesn't increase?

Bounding n

- Is there some n_0 such that, for $n > n_0$, the maximum achieved in the optimization problem doesn't increase?
- Dueck's paper says that $n_0 = 6$ works. But can we do better?

Uniqueness Theorem

We wish to show that $\widetilde{\boldsymbol{c}}$, defined as the $\boldsymbol{c}$ that causes the inequality to be as tight as possible, is uniquely defined. Furthermore, we wish to show it can be defined by taking partial derivatives of the inequality bounding $\boldsymbol{c}$.

Reducing n_0 to 3

In order to reduce n_0 down to 3, we want to show that any possible $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ with $n \geq 4$ can be adjusted by changing $\boldsymbol{p}$ to $\boldsymbol{p}^*$ while preserving

- $\sum_{i=1}^{n} p_i^* = 1$
- $\sum_{i=1}^{n} p_i^*(a_i\bar{b}_i + \bar{a}_i b_i + 2\widetilde{c}_i)$ is fixed
- $\sum_{i=1}^{n} p_i^* H\left(a_i b_i - \widetilde{c}_i, a_i\bar{b}_i + \widetilde{c}_i, \bar{a}_i b_i + \widetilde{c}_i, \bar{a}_i\bar{b}_i - \widetilde{c}_i\right)$ is fixed
- $\sum_{i=1}^{n} p_i^*(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$ is non-decreasing

It can be shown that, if $n \geq 4$, then p^* can be created such that $\exists i$ with $p_i^* = 0$.

Conjectures

- We suspect $n = 1$ suffices and that the maximum occurs at $a_1 = b_1 = \frac{\log(2+\sqrt{3})-\log(2)}{2\log(2+\sqrt{3})} \approx 0.23684$ giving a maximum of ≈ 1.57948.

Conjectures

- We suspect $n = 1$ suffices and that the maximum occurs at $a_1 = b_1 = \frac{\log(2+\sqrt{3})-\log(2)}{2\log(2+\sqrt{3})} \approx 0.23684$ giving a maximum of ≈ 1.57948.
- This would mean 2 defects out of n could be found with $\left(\frac{2}{H(a_1,\bar{a}_1)+H(b_1,\bar{b}_1)} + O(1)\right)\log_2(n) \approx 1.266\log_2(n)$ tests.

Acknowledgements

We'd like to thank the following people/organizations/animals:

- Our mentor, Dr. Zilin Jiang
- The PRIMES program
- Parents
- Sam's cat

References I

M. Aigner.
Search problems on graphs.
Discrete Appl. Math., 14(3):215–230, 1986.

A.Ya. Belokopytov and V.N. Luzgin.
Block transmission of information in a summing multiple access channel with feedback.
Probl. Inf. Transm., 23(4):347–351, 1987.

G. Dueck.
The zero error feedback capacity region of a certain class of multiple-access channels.
Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 14(2):89–103, 1985.

References II

L. Gargano, V. Montouri, G. Setaro, and U. Vaccaro.
An improved algorithm for quantitative group testing.
Discrete Applied Mathematics, 36(3):299 – 306, 1992.

Zilin Jiang, Nikita Polyanskii, and Ilya Vorobyev.
On capacities of the two-user union channel with complete feedback.
IEEE Trans. Inform. Theory, 65(5):2774–2781, 2019.

Zhen Zhang, T. Berger, and J. Massey.
Some families of zero-error block codes for the two-user binary adder channel with feedback.
IEEE Transactions on Information Theory, 33(5):613–619, September 1987.