Group testing for two defectives and the zero-error channel capacity

Sam Florin* Matthew Ho† Rahul Thomas‡

January 14, 2021

Abstract

The issue of identifying defects in a set with as few tests as possible has many applications, including in maximum efficiency pool testing during the COVID-19 pandemic. This research aims to determine the rate of growth of the number of tests required relative to the logarithm of the size of the set. In particular, we focus on the case where there are exactly two defects in the set, which is equivalent to the problem of determining the zero-error capacity of a two-user binary adder channel with complete feedback. The channel capacity is given by a non-linear optimization problem involving entropy functions, whose optimal value remains unknown. In this paper, using the linear dependence technique, we are able to reduce the complexity of the optimization problem significantly. We also gather numerical evidence for the conjectured optimal value.

1 Introduction

A major issue governments face during the COVID-19 pandemic is the problem of how to test people efficiently in order to quickly isolate and treat the infected in order to prevent further spread. One promising approach is to test larger pools of people by combining batches of samples, commonly known as "pooled testing". When the proportion of infected patients is low, this approach is extremely efficient compared to the naive approach of individually testing every person–meaning that pooled testing can be very useful in situations where there are testing equipment shortages or where testing can be extremely costly. This makes testing a large number of people much more efficient.

Unfortunately, in the real world, COVID-19 tests do not have infinite accuracy or precision, so there is an upper bound to the number of samples that can be batched together. Also, while scientists can determine whether a batch tests positive, it is more difficult to tell how many samples in the batch are positive. In practice, this means that once a test comes back positive, each sample is tested one by one–which is inefficient if the proportion of positive samples is high. However, it is of interest to study the ideal case, where we can combine any combination of arbitrarily many samples at once and have perfect tests.

We can model the pooled testing as follows. Assume that we have n people who we wish to test. We can choose any subset of these n people and test their pooled samples, determining how many people inside this subset are infected. We wish to minimize the number of tests needed to identify with certainty which patients are infected. In this paper, we focus on the case that two out of n people are infected, and we study the minimum number, denoted by $w(n)$, of tests needed to conclusively identify these two people.

*Greenwich High School, Greenwich, CT, USA.
†Palo Alto High School, Palo Alto, CA, USA.
‡Cherry Creek High School, Greenwood Village, CO, USA.

Under the guise of "quantitative group testing", a lot of research, which dates back to the 80s, has been devoted to the determination of the asymptotic behavior of $w(n)$. See [Aig86, ZBM87, GMSV92, BL87]. The best bound on $w(n)$ was given by Gargano [GMSV92]. He introduced a recursive algorithm based on the related quantity $w(m,n)$, the minimum number of weighings to identify two defects in disjoint sets of size m, n, to raise the lower bound to $w(n) \geq 2.28 \log n$. (We let the base of log be 2 unless otherwise stated.)

In [JPV19], a connection between the group testing problem and the multiple-access channel with complete feedback was mentioned — $w(n)$ is asymptotic to $2(\log n)/c$ holds, where c is the maximum total zero-error capacity of two-user adder channel with complete feedback. This zero-error capacity is characterized by Dueck [Due85] as the following non-linear optimization problem involving entropy functions.

$$\begin{aligned} \text{Maximize:} \quad & H(X \mid U) + H(Y \mid U) \\ \text{Subject to:} \quad & I(U, Z^*) \geq H(X^*, Y^* \mid U, Z^*). \end{aligned}$$

Here, $H(\cdot \mid \cdot)$ is the conditional entropy, and $I(\cdot, \cdot)$ is the mutual information. U, X, Y are random variables on finite sets $\mathcal{U}, \mathcal{X}, \mathcal{Y}$, respectively, and $\mathcal{Z}$ is a finite set. Furthermore, $\mathcal{P}(U, X, Y)$ is the set of triples of random variables (X^*, Y^*, Z^*) on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, such that $P_{X|U} = P_{X^*|U}$, $P_{Y|U} = P_{Y^*|U}$, and $P(Z^* = z | Y^* = y, X^* = x) = 0$ if and only if $P(z|y,x) = 0$ in the multiple-access channel.

The optimal value of the above optimization problem is central in both group testing and information theory, and it is a long-standing problem to determine precisely the exact optimal value. In this paper, we make a significant progress on reducing the complexity of the optimization problem.

The rest of the paper is organized as follows. In Section 2, we transform the above non-linear optimization problem into a more concrete one involving undetermined number of variables. The number of variables is also called the cardinality of the optimization problem. In Section 3, we review a bound on the cardinality, and we show this bound can be further reduced. The proof relies crucially on a uniqueness result, which is shown in Section 4. In Section 5, we consider the case where $n = 1$ and work to determine the optimal rate. In Section 6, we discuss the results of numerical experiments performed to explore the conjectured optimal value.

2 Connection to two-user binary adder channel

For the entirety of the paper, we denote $1 - x$ by $\bar{x}$. Furthermore, the entropy function H, when used on a set of nonnegative reals rather than a random variable, is defined as

$$H(x_1, ..., x_n) := -\sum_{i=1}^{n} x_i \log x_i.$$

We reformulate Dueck's optimization in the following more down-to-earth form. Our proof follows Belokopytov's computation in [BL87].

Theorem 1. *Dueck's optimization problem is equivalent to the following optimization problem* OPT_n^**:*

$$\begin{aligned} \textit{Maximize:} \quad & \sum_{i=1}^{n} p_i (H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)) \\ \textit{Subject to:} \quad & \sum_{i=1}^{n} p_i = 1, p_i \geq 0, a_i, b_i \in [0,1] \textit{ for all } i \in [n], \end{aligned}$$

$$H\begin{pmatrix} \sum_{i=1}^n p_i(a_ib_i - c_i) \\ \sum_{i=1}^n p_i(a_i\bar{b}_i + \bar{a}_ib_i + 2c_i) \\ \sum_{i=1}^n p_i(\bar{a}_i\bar{b}_i - c_i) \end{pmatrix} \geq \sum_{i=1}^n p_i H\begin{pmatrix} a_ib_i - c_i & a_i\bar{b}_i + c_i \\ \bar{a}_ib_i + c_i & \bar{a}_i\bar{b}_i - c_i \end{pmatrix},$$

for all c_i that make the terms inside $H(\cdot,\cdot,\cdot,\cdot)$non-negative.

Proof. Define finite sets $\mathcal{U} = \{1, 2, ..., n\}, \mathcal{X} = \{0, 1\}, \mathcal{Y} = \{0, 1\}, \mathcal{Z} = \{0, 1, 2\}$. Let U, X, Y be random variables on $\mathcal{U}, \mathcal{X}, \mathcal{Y}$, respectively. To define U, for every $i \in \mathcal{U}$ set $p(U = i) = p_i \in [0, 1]$, where $\sum_{i=1}^n p_i = 1$. For convenience of notation, let $p_i(S)$ be the probability that event S occurs given that $U = i$. To define X and Y, set $p_i(X = 0) = a_i, p_i(X = 1) = \bar{a}_i, p_i(Y = 0) = b_i, p_i(Y = 1) = \bar{b}_i$, where $a_i, b_i \in [0, 1]$.

We define triples of random variables (X^*, Y^*, Z^*) on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, that belong to $\mathcal{P}(U, X, Y)$. First, the conditions $P_{X^*|U} = P_{X|U}, P_{Y^*|U} = P_{Y|U}$ are equivalent to $p_i(X^* = 0) = a_i, p_i(X^* = 1) = \bar{a}_i, p_i(Y^* = 0) = b_i, p_i(Y^* = 1) = \bar{b}_i$ for all $i \in U$. This already defines the random variables X^*, Y^*. The last condition, $p_i(Z^* = z|X^* = x, Y^* = y)$ is nonzero if and only if $x + y = z$, can be used to completely define Z^* once we set $p_i(Z^* = 0|X^* = 0, Y^* = 0) = a_ib_i - c_i$ for some c_i. First, $p_i(Z^* = 1|X^* = 1, Y^* = 0) = p_i(Y^* = 0) - p_i(Z^* = 0|X^* = 0, Y^* = 0) = b_i - (a_ib_i - c_i) = \bar{a}_ib_i + c_i$. Similarly, $p_i(Z^* = 1|X^* = 1, Y^* = 0) = \bar{b}_ia_i + c_i$ and $p_i(Z^* = 2|X^* = 1, Y^* = 1) = \bar{a}_i\bar{b}_i - c_i$. These four probabilities together define the random variable Z^*. Note that varying over all $(X^*, Y^*, Z^*) \in \mathcal{P}(U, X, Y)$ is the same as varying over all real c_i that make these probabilities nonnegative.

With the probability distributions of the random variables, we can rewrite the maximized quantity as:

$$\begin{aligned} H(X|U) + H(Y|U) &= \sum_{i\in U}\sum_{x\in X} p_iH(p_i(X = x)) + \sum_{i\in U}\sum_{y\in Y} p_iH(p_i(Y = y)) \\ &= \sum_{i=1}^n\sum_{x=0}^1 p_iH(p_i(X = x)) + \sum_{i=1}^n\sum_{y=0}^1 p_iH(p_i(Y = y)) \\ &= \sum_{i=1}^n (p_iH(a_i) + p_iH(\bar{a}_i)) + \sum_{i=1}^n (p_iH(b_i) + p_iH(\bar{b}_i)) \\ &= \sum_{i=1}^n p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)). \end{aligned}$$

Thus, maximizing $H(X|U) + H(Y|U)$ is equivalent to maximizing $\sum_{i=1}^n p_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i))$.

Now, we rewrite the inequality constraint, first simplifying the difference:

$$\begin{aligned} I(Z^*; U) - H(X^*, Y^*|U, Z^*) &= H(Z^*) - H(Z^*|U) - H(X^*, Y^*|U, Z^*) \\ &= H(Z^*) - (H(Z^*|U) + H(X^*, Y^*, U, Z^*) - H(U, Z^*)) \\ &= H(Z^*) - (H(X^*, Y^*, U, Z^*) - H(U)) \\ &= H(Z^*) - (H(X^*, Y^*, U) - H(U)) \\ &= H(Z^*) - H(X^*, Y^*|U), \end{aligned}$$

where the second to last line follows because the probability distribution of Z^* is completely dependent on those of X^*, Y^*, U. We form an equivalent expression for this difference by using our definitions of random variables Z^*, X^*, Y^*:

$$H(Z^*) - H(X^*, Y^*|U) = \sum_{z\in\mathcal{Z}} H(p(Z^* = z)) - \sum_{i\in\mathcal{U}}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_iH(p_i(X^* = x, Y^* = y))$$

$$
\begin{aligned}
&= \sum_{z=0}^{2} H\left(\sum_{i=1}^{n} p_i p_i(Z^* = z)\right) - \sum_{i=1}^{n}\sum_{x=0}^{1}\sum_{y=0}^{1} p_i H(p_i(X^* = x, Y^* = y)) \\
&= H\left(\sum_{i=1}^{n} p_i(a_i b_i - c_i)\right) + H\left(\sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i)\right) + H\left(\sum_{i=1}^{n} p_i(\bar{a}_i \bar{b}_i - c_i)\right) \\
&\quad - \sum_{i=1}^{n} p_i \left(H(a_i b_i - c_i) + H(a_i \bar{b}_i + c_i) + H(\bar{a}_i b_i + c_i) + H(\bar{a}_i \bar{b}_i - c_i)\right) \\
&= H\left(\sum_{i=1}^{n} p_i(a_i b_i - c_i), \sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i), \sum_{i=1}^{n} p_i(\bar{a}_i \bar{b}_i - c_i)\right) \\
&\quad - \sum_{i=1}^{n} p_i H(a_i b_i - c_i, a_i \bar{b}_i + c_i, \bar{a}_i b_i + c_i, \bar{a}_i \bar{b}_i - c_i).
\end{aligned}
$$

Therefore, the minimum of $I(Z^*;U) - H(X^*, Y^*|U, Z^*)$ over all $(X^*, Y^*, Z^*) \in \mathcal{P}(U, X, Y)$ being nonnegative is equivalent to:

$$
H\left(\sum_{i=1}^{n} p_i(a_i b_i - c_i), \sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i), \sum_{i=1}^{n} p_i(\bar{a}_i \bar{b}_i - c_i)\right) \geq \sum_{i=1}^{n} p_i H(a_i b_i - c_i, a_i \bar{b}_i + c_i, \bar{a}_i b_i + c_i, \bar{a}_i \bar{b}_i - c_i)
$$

for all c_i that make all parameters of H nonnegative. As the inequalities and maximized quantities in Dueck's optimization problem and OPT_n^* are equivalent, with necessary conditions $\sum_{i=1}^{n} p_i = 1$ and $p_i, a_i, b_i \in [0, 1]$ for all $i \in U = [n]$ included in the latter, Dueck's optimization problem is equivalent to OPT_n^*. □

We now transform this optimization problem to another optimization problem.

Theorem 2. *The optimization problem* OPT_n^* *is equivalent to the following optimization problem* OPT_n.

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

over all points $\boldsymbol{a} = (a_1, ..., a_n), \boldsymbol{b} = (b_1, ..., b_n), \boldsymbol{p} = (p_1, ..., p_n)$ *in* $[0, 1]^n$ *such that*

$$
\sum_{i=1}^{n} p_i = 1, \tag{1}
$$

$$
L\left(\sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i)\right) \geq \sum_{i=1}^{n} p_i H\left(a_i b_i - c_i, a_i \bar{b}_i + \bar{a}_i b_i + 2c_i, \bar{a}_i \bar{b}_i - c_i\right), \tag{2}
$$

for all points $\boldsymbol{c} = (c_1, ..., c_n)$ *that make the terms inside* $H(\cdot, \cdot, \cdot)$ *non-negative, where*

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

Proof. Let the maximum value of the objective function be M^* for OPT_n^*, achieved with points $\boldsymbol{a}_0, \boldsymbol{b}_0, \boldsymbol{p}_0$, and M for OPT_n, achieved with points $\boldsymbol{a}_1, \boldsymbol{b}_1, \boldsymbol{p}_1$. We wish to show $M^* = M$.

First, we show $M^* \geq M$. Because $x \mapsto -x \log x$ is convex, for any nonnegative a, b, c such that $a+b+c = 1$, $H(a, b, c) \geq H(b) + H\left(\frac{1-b}{2}\right) + H\left(\frac{1-b}{2}\right) = L(b)$. Hence,

$$
H\left(\sum_{i=1}^{n} p_i(a_i b_i - c_i), \sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i), \sum_{i=1}^{n} p_i(\bar{a}_i \bar{b}_i - c_i)\right) \geq L\left(\sum_{i=1}^{n} p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i)\right)
$$

so $\boldsymbol{a}_1, \boldsymbol{b}_1, \boldsymbol{p}_1$ immediately satisfy the inequality constraint of OPT_n^*. Since the objective function does not change between the two optimization problems, OPT_n^* can achieve a maximum of M for these points, which immediately implies $M^* \geq M$.

Now, we show $M \geq M^*$. It suffices to construct points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ from $\boldsymbol{a}_0, \boldsymbol{b}_0, \boldsymbol{p}_0$ that satisfy the conditions of OPT_n and yield a value of M^* in the objective function. Let $\boldsymbol{a}_0, \boldsymbol{b}_0, \boldsymbol{p}_0 = (a_1, ..., a_n), (b_1, ..., b_n), (p_1, ..., p_n)$, respectively, where $\sum_{i=1}^n p_i = 1$. Now define $\boldsymbol{a} = (a_1, ..., a_n, \bar{a}_1, ..., \bar{a}_n)$ and $\boldsymbol{b} = (b_1, ..., b_n, \bar{b}_1, ..., \bar{b}_n)$, which each have length $2n$. Furthermore, define $\boldsymbol{p} = \left(\frac{p_1}{2}, ..., \frac{p_n}{2}, \frac{p_1}{2}, ..., \frac{p_n}{2}\right)$, so that the elements still sum to one and the first condition still holds. Note that

$$F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}) = \sum_{i=1}^n \frac{p_i}{2}(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)) + \sum_{i=1}^n \frac{p_i}{2}(H(\bar{a}_i, a_i) + H(\bar{b}_i, b_i)) = \sum_{i=1}^n p_i((H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)).$$

Thus, $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ yield a value of $M^* = F(\boldsymbol{a}_0, \boldsymbol{b}_0, \boldsymbol{p}_0)$ in the objective function. All that remains is to check that they satisfy the conditions of OPT_n, which can be shown by noting that the H and L expressions are equal for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$:

$$\begin{aligned}
&H\left(\sum_{i=1}^n \frac{p_i}{2}(a_i b_i - c_i) + \sum_{i=1}^n \frac{p_i}{2}(\bar{a}_i \bar{b}_i - c_i), 2\sum_{i=1}^n \frac{p_i}{2}(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i), \sum_{i=1}^n p_i(\bar{a}_i \bar{b}_i - c_i) + \sum_{i=1}^n \frac{p_i}{2}(a_i b_i - c_i)\right) \\
&= H\left(\sum_{i=1}^n \frac{p_i}{2}(a_i b_i + \bar{a}_i \bar{b}_i - 2c_i), \sum_{i=1}^n p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i), \sum_{i=1}^n \frac{p_i}{2}(a_i b_i + \bar{a}_i \bar{b}_i - 2c_i)\right) \\
&= L\left(\sum_{i=1}^n p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i)\right).
\end{aligned}$$

Because $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ satisfy the conditions of OPT_n and yield a value of M^*, the maximum in this optimization problem is $M \geq M^*$.

Finally, we combine $M \geq M^*$ and $M^* \geq M$ to get $M^* = M$. Since optimization problems OPT_n^* and OPT_n yield the same maximum, they are equivalent. □

We now show that the inequality constraint (2) may be replaced with an equality constraint while preserving the optimal value, using a continuity argument. For convenience, we define

$$G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}) := L\left(\sum_{i=1}^n p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i)\right) - \sum_{i=1}^n p_i H\left(a_i b_i - c_i, a_i \bar{b}_i + c_i, \bar{a}_i b_i + c_i, \bar{a}_i \bar{b}_i - c_i\right).$$

Proposition 3. *The optimal value of* OPT_n *will not change if* (2) *is replaced by*

$$G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}) = 0. \tag{3}$$

Proof. Assume that for some points $\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}$ we have an optimal value of the objective function. Furthermore, assume that there is at least one coordinate of $\boldsymbol{a}$ or $\boldsymbol{b}$ which is not 0.5. If

$$G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}) > 0,$$

then we can perturb this coordinate by some sufficiently small positive ϵ in the direction of 0.5 while continuing to satisfy $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}) > 0$. (We can guarantee the constraint remains satisfied because $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c})$ is continuous.) This will increase our objective function, giving an even more optimal value of the objective function, which is a contradiction.

Now, it is straightforward to show that it is impossible for all coordinates of $\boldsymbol{a}$ and $\boldsymbol{b}$ to be 0.5, because (2) will not be satisfied. This implies that any optimal values of $\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}$ must occur at equality in (2), as desired. □

For the rest of the paper, we shall assume (3) as the constraint instead of (2).

3 Bounding cardinality of optimization problem

With the following uniqueness theorem, we can let $n = 3$ in OPT_n without loss. As in the previous section, we define

$$G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}) := L\left(\sum_{i=1}^{n} p_i(a_i\bar{b}_i + \bar{a}_i b_i + 2c_i)\right) - \sum_{i=1}^{n} p_i H\left(a_i b_i - c_i, a_i\bar{b}_i + c_i, \bar{a}_i b_i + c_i, \bar{a}_i\bar{b}_i - c_i\right). \tag{4}$$

Because (3) must be satisfied for all $\boldsymbol{c}$, we only have to check the point $\boldsymbol{c}$ which minimizes $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p},\boldsymbol{c}}$.

Definition 4. Fix points $\boldsymbol{p} = (p_1, ..., p_n)$, $\boldsymbol{a} = (a_1, ..., a_n)$, $\boldsymbol{b} = (b_1, ..., b_n)$. The *minimum point* $\boldsymbol{c}$ is defined to be the point $(c_1, ..., c_n)$ for which $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c})$ is minimized.

Now, we begin to characterize these minimum points. We postpone the proof of this characterization to Section 4.

Theorem 5 (Characterization of minimum points)**.** *Assume $\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}$ satisfying the conditions in* OPT_n *give an maximum value of the objective function. Then there is a unique minimum point $\boldsymbol{c}$. Furthermore, this point satisfies $\frac{\partial G}{\partial c_i} = 0$ for all $i \in [n]$.*

With the following theorem, as long as $n \geq 4$, we can always create a solution for $n-1$ that is at least as optimal any solution for n. Thus, we can eventually set $n = 3$ in our optimization problem.

Theorem 6. *Let $\boldsymbol{p} = (p_1, ..., p_n)$, $\boldsymbol{a} = (a_1, ..., a_n)$, $\boldsymbol{b} = (b_1, ..., b_n)$ be points satisfying the conditions of OPT_n, where $n \geq 4$, such that $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ result in the maximum possible value of the objective function F. There exist points $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_{n-1})$,$\hat{\boldsymbol{a}} = (\hat{a}_1, ..., \hat{a}_{n-1})$,$\hat{\boldsymbol{b}} = (\hat{b}_1, ..., \hat{b}_{n-1})$ that also satisfy these conditions and $F(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{p}}) \geq F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p})$.*

Proof. We will provide a construction of such $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{p}}$ starting from $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$. Let $\boldsymbol{c}^* = (c_1^*, ..., c_n^*)$ be the minimum point of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$, which is unique by Theorem 5. We shall define a nonzero point $\boldsymbol{v} := (v_1, ..., v_n)$ such that

$$\sum_{i=1}^{n} v_i = 0, \quad \sum_{i=1}^{n} v_i(a_i\bar{b}_i + \bar{a}_i b_i + 2c_i^*) = 0,$$

$$\sum_{i=1}^{n} v_i H(a_i b_i - c_i^*, a_i\bar{b}_i + c_i^*, \bar{a}_i b_i + c_i^*, \bar{a}_i\bar{b}_i - c_i^*) = 0, \quad \sum_{i=1}^{n} v_i(H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)) \geq 0.$$

Because there are $n \geq 4$ variables, we can find infinite solutions to the three linear equations in $v_1, ..., v_n$. Thus, one of these solutions has at least one nonzero v_i. If this solution satisfies the inequality, we have found a valid $\boldsymbol{v}$; otherwise, simply reverse the signs of all v_i, as the the first three equations will still hold and the inequality will become true.

Consider the point $\boldsymbol{p}_t = \boldsymbol{p} + \boldsymbol{v}t$ for any t. Note that $\sum_{i=1}^n (p_i + v_i t) = \sum_{i=1}^n p_i + t\sum_{i=1}^n v_i = 1$, so $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}_t$ satisfy (1) of OPT_n. Now let $\boldsymbol{c}_t = (c_{1t}, ..., c_{nt})$ be the minimum point of $\boldsymbol{p}_t, \boldsymbol{a}, \boldsymbol{b}$. Theorem 5 tells us that for all $i \in [n]$, $(\partial G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}/\partial c_i)|_{c_i = c_{it}} = 0$ as the optimal value of the objective function is conserved, which after simplification yields the system

$$\left(\frac{1-X}{2X}\right)^2 = \frac{(a_i b_i - c_{it})(\bar{a}_i \bar{b}_i - c_{it})}{(a_i \bar{b}_i + c_{it})(\bar{a}_i b_i + c_{it})} \quad \text{and} \quad X = \sum_{i=1}^n (p_i + v_i t)(a_i \bar{b}_i + \bar{a}_i b_i + 2c_{it}).$$

We claim that $\boldsymbol{c}_t = \boldsymbol{c}^*$ is the unique solution. It suffices to substitute $c_{it} = c_i^*$ and verify the equalities:

$$\left(\frac{1-X}{2X}\right)^2 = \frac{(a_i b_i - c_i^*)(\bar{a}_i \bar{b}_i - c_i^*)}{(a_i \bar{b}_i + c_i^*)(\bar{a}_i b_i + c_i^*)}$$

$$X = \sum_{i=1}^n p_i(a_i \bar{b}_i + b_i \bar{b}_i + 2c_i^*) + t\sum_{i=1}^n v_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i^*) = \sum_{i=1}^n p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i^*).$$

Note that the last step follows from the definition of v_i. These equalities are exactly the same as $\frac{\partial G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}}{\partial c_i}(c_i^*) = 0$ for all $i \in [n]$, which is immediately true by Theorem 5 on points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$. This solution is unique, so $\boldsymbol{c}_t = \boldsymbol{c}^*$.

Thus, $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}_t}(\boldsymbol{c}_t) = G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}_t}(\boldsymbol{c}^*)$. Also, from the second and third equations in the definition of v_i,

$$\begin{aligned} G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}_t}(\boldsymbol{c}^*) &= L\left(\sum_{i=1}^n (p_i + v_i t)(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i^*)\right) - \sum_{i=1}^n (p_i + v_i t) H\left(a_i b_i - c_i^*, a_i \bar{b}_i + c_i^*, \bar{a}_i b_i + c_i^*, \bar{a}_i \bar{b}_i - c_i^*\right) \\ &= L\left(\sum_{i=1}^n p_i(a_i \bar{b}_i + \bar{a}_i b_i + 2c_i^*)\right) - \sum_{i=1}^n p_i H\left(a_i b_i - c_i^*, a_i \bar{b}_i + c_i^*, \bar{a}_i b_i + c_i^*, \bar{a}_i \bar{b}_i - c_i^*\right) = G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}^*). \end{aligned}$$

Combining these two equations with $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(\boldsymbol{c}^*) = 0$, which follows from (3) of OPT_n, we obtain $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}_t}(\boldsymbol{c}_t) = 0$. As $\boldsymbol{c}_t$ is the minimum point of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}_t$, this immediately implies $G_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}_t}(\boldsymbol{c}) = 0$ for valid $\boldsymbol{c}$, i.e. $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}_t$ satisfy (3) of OPT_n.

Finally, linearly increase t until some element of $\boldsymbol{p}_t$ becomes zero, at $t = t_0$. Say the ith element is zero. Let $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{p}}$ be the point formed when the ith elements of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ are removed, with length $n-1$. These points still satisfy (1), (3) of OPT_n, and $F(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{p}}) = F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}_t) \geq F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p})$, so our construction is complete. □

4 Uniqueness of minimum point in the constraint

The entirety of this section is devoted to proving Theorem 5. We first follow the steps shown in Section 4 of [BL87] to show that the minimum point given optimal $\boldsymbol{p}, \boldsymbol{a}$, and $\boldsymbol{b}$ cannot lie on the boundary and they must satisfy $\frac{\partial G}{\partial c_i} = 0$ for each i. We then demonstrate the uniqueness of the minimum point for $n = 1$ and use a key lemma to prove uniqueness for $n \geq 2$. At the end of the section, we combine these results to prove Theorem 5.

Proposition 7. *The optimal values for c_i, assuming $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ give an optimal value, are where $\frac{\partial G}{\partial c_i} = 0$, and are not on the boundary of the region.*

Remark. The condition for $\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}$ being optimal is necessary, because there are cases where a_i, b_i are close to 0 or 1, causing the value for c_i given by the derivative being zero to fall outside of the possible range. For example, letting $\boldsymbol{p} = (0.3, 0.3, 0.4), \boldsymbol{a} = (0.22, 0.98, 0.11), \boldsymbol{b} = (0.28, 0.96, 0.33)$ causes the partial derivative formula for $\boldsymbol{c}$ to output $\boldsymbol{c} = (-0.0113, -2.02, -0.011)$, with c_2 clearly being outside the possible range.

Proof. We begin by noting the following for all i:

$$a_ib_i - c_i \geq 0 \implies c_i \leq a_ib_i, \quad a_i\bar{b}_i + c_i \geq 0 \implies c_i \geq a_ib_i - a_i,$$
$$\bar{a}_ib_i - c_i \geq 0 \implies c_i \geq a_ib_i - b_i, \quad \bar{a}_i\bar{b}_i - c_i \geq 0 \implies c_i \leq (1-a_i)(1-b_i).$$

Therefore, we have two cases where a value of c_i is on the boundary. Either $c_i = \max(a_ib_i - a_i, a_ib_i - b_i)$ or $c_i = \min(a_ib_i, (1-a_i)(1-b_i))$.

We assume without loss of generality that $p_i \neq 0$ for all i. Define

$$Y := \sum_i p_i(a_i\bar{b} + \bar{a}b_i + 2c_i), \quad J_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{c}, i) := \frac{\partial G_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{c})}{\partial c_i} = p_i\left(2\log\frac{1-Y}{2Y} - \log\frac{(a_ib_i - c_i)(\bar{a}_i\bar{b}_i - c_i)}{(a_i\bar{b}_i + c_i)(\bar{a}_ib_i + c_i)}\right).$$

We shall assume for the sake of contradiction that for some i we have c_i is on the boundary, that this choice of $\boldsymbol{c}$ minimizes G, and $G_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{c}) = 0$.

First, assume that $c_i = \max(a_ib_i - a_i, a_ib_i - b_i)$ for some fixed i. We create a new point $\boldsymbol{c'}$ such that $c'_i = c_i + \epsilon$ for an arbitrarily small value of $\epsilon > 0$, with all other indices $j \neq i$ satisfying $c'_j = c_j$. Then because $\boldsymbol{c}$ results in the global minimum of the function G, we must have $\frac{\partial G}{\partial c'_i} = J_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c'}}(i) > 0$. But we have that

$$\lim_{\epsilon\to 0}\log\frac{(a_ib_i - c'_i)(\bar{a}_i\bar{b}_i - c'_i)}{(a_i\bar{b}_i + c'_i)(\bar{a}_ib_i + c'_i)} = \infty.$$

The combination of these facts implies that we must have

$$\lim_{\epsilon\to 0} Y = 0.$$

Because we assumed $p_j \neq 0$ for all j, we have that for all j, $2c_j = 2a_jb_j - a_j - b_j$. But we have for all c_j that $2c_j \geq 2\max(a_jb_j - a_j, a_jb_j - b_j) \geq 2a_jb_j - a_j - b_j$, with equality if and only if $a_j = b_j$ for all j.

We assumed that $\boldsymbol{c}$ results in G having a global minimum, so we then have the following:

$$\begin{aligned} L(0) - \sum_i p_iH(a_i, 1-a_i) &\leq L\left(\sum_i p_i\left(2a_i - 2a_i^2\right)\right) - \sum_i p_iH\left(a_i^2, a_i(1-a_i), a_i(1-a_i), (1-a_i)^2\right) \\ &= L\left(\sum_i p_i\left(2a_i - 2a_i^2\right)\right) - 2\sum_i p_iH(a_i, 1-a_i). \end{aligned}$$

This results in

$$R_1 = R_2 = \sum_i p_iH(a_i, 1-a_i) \leq L\left(\sum_i p_i\left(2a_i - 2a_i^2\right)\right) - L(0) \leq \log 3 - 1 = 0.5850.$$

This is worse than the construction $a = \langle 0.23684\rangle, b = \langle 0.23684\rangle, c = \langle 0.04071\rangle$ resulting in $R_1 = R_2 = 0.78974$.

Now, assume that $c_i = \min(a_ib_i, 1 - a_i - b_i - a_ib_i)$ for some fixed i. We redefine $\boldsymbol{c'}$ such that $c'_i = c_i - \epsilon$ for an arbitrarily small $\epsilon > 0$ with all $j \neq i$ satisfying $c'_j = c_j$. Here, we have that $\frac{\partial G}{\partial c'_i} = J_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c'}}(i) < 0$, along with:

$$\lim_{\epsilon\to 0}\log\frac{(a_ib_i - c'_i)(\bar{a}_i\bar{b}_i - c'_i)}{(a_i\bar{b}_i + c'_i)(\bar{a}_ib_i + c'_i)} = -\infty.$$

This then implies $\lim_{\epsilon\to 0} Y = 1$. Because $G = 0$ and $L(Y) = 0$, we must have for all j

$$H(a_jb_j - c_j, a_j\bar{b}_j + c_j, \bar{a}_jb_j + c_j, \bar{a}_j\bar{b}_j) = 0.$$

This implies all four of $a_jb_j - c_j, a_j\bar{b}_j + c_j, \bar{a}_jb_j + c_j, \bar{a}_j\bar{b}_j - c_j$ are integers. Namely, $a_jb_j - c_j + a_j\bar{b}_j + c_j = a_j$ and $a_jb_j - c_j + \bar{a}_jb_j + c_j = b_j$ must also be integers, so we have $a_j, b_j \in \{0, 1\}$ for all j. This gives $R_1 = R_2 = 0$. □

Now, we prove that the minimum point $\boldsymbol{c}$ is unique when the cardinality is 1.

Proposition 8. *For all $a, b \in [0,1]$, there is a unique $x \in (0,1)$ which is a zero of the function*

$$f(x) = a\bar{b} + \bar{a}b + 2c(x) - x$$

where $c(x)$ is defined as the unique solution to

$$\frac{(ab - c(x))(\bar{a}\bar{b} - c(x))}{(a\bar{b} + c(x))(\bar{a}b + c(x))} = \left(\frac{1-x}{2x}\right)^2$$

satisfying $\max(-a\bar{b}, -\bar{a}b) \leq c(x) \leq \min(ab, \bar{a}\bar{b})$.

Proof. Solving for $c(x)$, we find

$$c(x) = \frac{x}{2} - \frac{a}{2} - \frac{b}{2} + ab.$$

Substituting, we find that we want to show $h(x) := (2x)^2(a+b-x)(2-a-b-x) - (1-x)^2(a-b+x)(-a+b+x)$ has a unique root. It suffices to show that when $h'(x) = 0$ we always have $h(x) > 0$. Now, $2h(x) - xh'(x) = 2((a-b)^2 + 3x^2)(1-x)$ is positive for $x \in (0,1)$ as desired. □

We now proceed to the case where the cardinality is 2. First, we will present a lemma regarding the behavior of $\frac{f'}{f}$.

Lemma 9. *When $(a,b) \neq (0,0)$,*

$$\frac{f'(x)}{f(x)} \neq \frac{1}{x} \text{ for } x \in (0,1).$$

Proof. We first have:

$$f(x) = \left\{ x < \frac{1}{3} : -x - k_0(x) + \sqrt{s_0(x)},\ x \geq \frac{1}{3} : -x - k_0(x) - \sqrt{s_0(x)} \right\},$$

where

$$\begin{aligned}
k_0(x) &= -\frac{4x^2}{(3x-1)(x+1)} \\
k_1(x) &= -\frac{8x(x-1)}{(x+1)^2(3x-1)^2} \\
s_0(x) &= (a-b)^2 + k_0(x)^2 + 2k_0(x)(a+b-2ab) \\
s_1(x) &= 2k_0(x)k_1(x) + 2k_1(x)(a+b-2ab).
\end{aligned}$$

We wish to show there are no (non-corner) solutions to the equation

$$f(x) - xf'(x) = 0.$$

We split into two cases.

First, we examine the case where $0 < x < \frac{1}{3}$. Here,

$$f(x) = -x - k_0(x) + \sqrt{s_0(x)}.$$

Then

$$f'(x) = -1 - k_1(x) + \frac{s_1(x)}{2\sqrt{s_0(x)}}.$$

Again, we want to show that

$$\begin{aligned} 0 &= -x - k_0(x) + \sqrt{s_0(x)} + x + xk_1(x) - x \cdot \frac{s_1(x)}{2\sqrt{s_0(x)}} \\ &= -k_0(x) + \sqrt{s_0(x)} + xk_1(x) - x \cdot \frac{s_1(x)}{2\sqrt{s_0(x)}} \end{aligned}$$

has no solutions.

Rearranging gives us that we want to show

$$2s_0(x) - x \cdot s_1(x) = 2(k_0(x) - xk_1(x)) \cdot \sqrt{s_0(x)}.$$

Cross multiplying then gives

$$8(3x^4 + x^2)\sqrt{s_0(x)} = (x+1)^2(3x-1)^2(xs_1(x) - 2s_0(x)).$$

Further expansion then gives:

$$\begin{aligned} &\frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 - 2(a-b)^2(x+1)^2(1-3x)^2 + 16x^3(3x+1)(a+b-2ab) \\ &\qquad = 8x^2\left(3x^2+1\right)\sqrt{s_0(x)}, \end{aligned}$$

where $s_0(x) = (a-b)^2 + (k_0(x))^2 + 2k_0(x)(a+b-2ab)$ and $k_0(x) = \frac{4x^2}{(1-3x)(x+1)}$. Since

$$\begin{aligned} &\frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 - 2(a-b)^2(x+1)^2(1-3x)^2 + 16x^3(3x+1)(a+b-2ab) \\ &\qquad \leq \frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 + 16x^3(3x+1)(a+b-2ab), \end{aligned}$$

we can also try to show

$$\frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 + 16x^3(3x+1)(a+b-2ab) \leq 8x^2\left(3x^2+1\right)\sqrt{s_0(x)}.$$

Dividing through by $8x^2$ gives

$$\frac{8x^2(1-x)}{(x+1)(1-3x)} - 4x^2 + 2x(3x+1)(a+b-2ab) \leq \left(3x^2+1\right)\sqrt{s_0(x)}.$$

This simplifies to

$$\frac{4x^2(3x^2+1)}{(x+1)(1-3x)} + 2x(3x+1)(a+b-2ab) \leq \left(3x^2+1\right)\sqrt{s_0(x)}.$$

Since both sides of the inequality are non-negative, squaring them preserves the inequality, giving

$$\begin{aligned} &\left(\frac{4x^2(3x^2+1)}{(x+1)(1-3x)}\right)^2 + 2\left(\frac{4x^2(3x^2+1)}{(x+1)(1-3x)}\right)(2x(3x+1)(a+b-2ab)) + (2x(3x+1)(a+b-2ab))^2 \\ &\qquad \leq (3x^2+1)^2 s_0(x). \end{aligned}$$

Substituting in the values of $s_0(x)$ gives

$$\left(\frac{4x^2(3x^2+1)}{(x+1)(1-3x)}\right)^2 + 2\left(\frac{4x^2(3x^2+1)}{(x+1)(1-3x)}\right)(2x(3x+1)(a+b-2ab)) + (2x(3x+1)(a+b-2ab))^2$$
$$\leq (3x^2+1)^2((a-b)^2 + \left(-\frac{4x^2}{(3x-1)(x+1)}\right)^2 + 2\left(-\frac{4x^2}{(3x-1)(x+1)}\right)(a+b-2ab)).$$

This simplifies to

$$2\left(\frac{4x^2(3x^2+1)}{(x+1)(1-3x)}\right)(2x(3x+1)(a+b-2ab)) + (2x(3x+1)(a+b-2ab))^2$$
$$\leq (3x^2+1)^2((a-b)^2 + 2\left(-\frac{4x^2}{(3x-1)(x+1)}\right)(a+b-2ab)).$$

Clearing denominators gives

$$16x^3(3x^2+1)(3x+1)(a+b-2ab) + 4x^2(3x+1)^2(a+b-2ab)^2(x+1)(1-3x)$$
$$\leq (3x^2+1)^2(x+1)(1-3x)(a-b)^2 + 8x^2(3x^2+1)^2(a+b-2ab).$$

This simplifies to

$$(x+1)(1-3x)((3x^2+1)^2(a-b)^2 - 4x^2(3x+1)^2(a+b-2ab)^2)$$
$$\geq 8x^2(3x^2+1)(a+b-2ab)(2x(3x+1)-3x^2) = 8x^2(3x^2+1)(a+b-2ab)(-(x+1)(1-3x))$$

Dividing through by $(x+1)(1-3x)$ and simplifying gives

$$(3x^2+1)^2(a-b)^2 - 4x^2(3x+1)^2(a+b-2ab)^2 + 8x^2(3x^2+1)(a+b-2ab) \geq 0.$$

Calling this function $A(x)$, the goal is to show that $A(x) \geq 0$ for $x \in \left[0, \frac{1}{3}\right]$. Note that

$$A(x) = (3x^2+1)^2(a-b)^2 - 4x^2(3x+1)^2(a+b-2ab)^2 + 8x^2(3x^2+1)(a+b-2ab)$$
$$\geq (a-b)^2 - 4x^2(3x+1)^2(a+b-2ab)^2 + 8x^2(a+b-2ab) := B(x).$$

We wish to show that $B(x) \geq 0$ for $x \in \left[0, \frac{1}{3}\right]$. Letting $Y = a+b-2ab$, note that

$$B'(x) = -8xY(18x^2Y + 9xY + Y - 2)$$

$$B''(x) = -8Y(54x^2Y + 18xY - 2)$$

By Vieta's formula, the sum of the roots of $B''(x)$, if it does have 2 roots, is $-\frac{1}{3}$. This means $B''(x)$ has at most one root in $[0, \frac{1}{3}]$. And since $B'(0) = 0$ and $B''(0) = -8Y(Y-2) \geq 0$ as $0 \leq Y \leq 1$.

If $Y = 0$, then $B(x)$ is a constant function and therefore will not dip below 0. (Notice that we have equality here if and only if $a = b = 0$.) Otherwise, $B''(x) > 0$.

If $B''(x)$ has no roots in $[0, \frac{1}{3}]$, then B is always convex up so is always non-decreasing and therefore is always at most 0.

If $B''(x)$ has one root in $[0, \frac{1}{3}]$ then B goes from convex up to convex down. Then $B'(x)$ can have at most 1 root in $[0, \frac{1}{3}]$ meaning B goes from increasing to decreasing. This means the only possible critical point of B in $[0, \frac{1}{3}]$ would be a local maximum.

Thus, the minimum of B must occur at either $x = 0$ or $x = \frac{1}{3}$. Since $B(0) = (a-b)^2 \geq 0$ and $B(\frac{1}{3}) = (a-b)^2 - \frac{4}{9}(a+b-2ab)(4(a+b-2ab)-2)$ which has a minimum of 0 at $a = \frac{1}{2}, b = \frac{1}{2}$, $B(x) \geq 0$ for all $x \in [0, \frac{1}{3}]$.

Now, we proceed to the case where $\frac{1}{3} \leq x \leq 1$. Here,

$$f(x) = -x - k_0(x) - \sqrt{s_0(x)}.$$

Then

$$f'(x) = -1 - k_1(x) - \frac{s_1(x)}{2\sqrt{s_0(x)}}.$$

$$\begin{aligned} 0 &= -x - k_0(x) - \sqrt{s_0(x)} + x + xk_1(x) + x \cdot \frac{s_1(x)}{2\sqrt{s_0(x)}} \\ &= -k_0(x) - \sqrt{s_0(x)} + xk_1(x) + x \cdot \frac{s_1(x)}{2\sqrt{s_0(x)}}. \end{aligned}$$

Rearranging gives us

$$2\sqrt{s_0(x)}\left(k_0(x) - xk_1(x)\right) = x \cdot s_1(x) - 2s_0(x).$$

Cross multiplying then gives

$$-8(3x^4 + x^2)\sqrt{s_0(x)} = (x+1)^2(3x-1)^2(xs_1(x) - 2s_0(x)).$$

Further expansion then gives:

$$\frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 - 2(a-b)^2(x+1)^2(1-3x)^2 + 16x^3(3x+1)(a+b-2ab) = -8x^2\left(3x^2+1\right)\sqrt{s_0(x)},$$

where, again, $s_0(x) = (a-b)^2 + (k_0(x))^2 + 2k_0(x)(a+b-2ab)$ and $k_0(x) = \frac{4x^2}{(1-3x)(x+1)}$.

Clearly, the right hand side of this equation is at least 0. Assume for the sake of contradiction that the equation holds for some a, b, x not satisfying $a = 1 - b, x = 1$. We can simplify the left hand side:

$$\begin{aligned} &\frac{64x^4(1-x)}{(x+1)(1-3x)} - 32x^4 - 2(a-b)^2(x+1)^2(1-3x)^2 + 16x^3(3x+1)(a+b-2ab) \\ &\qquad = \frac{32x^4\left(3x^2+1\right)}{(x+1)(1-3x)} - 2(a-b)^2(x+1)^2(1-3x)^2 + 16x^3(3x+1)(a+b-2ab). \end{aligned}$$

Squaring both sides, we find that

$$\begin{aligned} &\frac{1024x^8\left(3x^2+1\right)^2}{(x+1)^2(3x-1)^2} + 256x^6(3x+1)^2(a+b-2ab)^2 - \frac{1024x^7(3x+1)\left(3x^2+1\right)(a+b-2ab)}{(x+1)(3x-1)} \\ &+2(a-b)^2(x+1)^2(3x-1)^2\left[2\cdot\frac{32x^4\left(3x^2+1\right)}{(x+1)(3x-1)} + 2(a-b)^2(x+1)^2(3x-1)^2 - 2\cdot 16x^3(3x+1)(a+b-2ab)\right] \\ &\qquad = 64x^4\left(3x^2+1\right)^2(a-b)^2 + \frac{1024x^8\left(3x^2+1\right)^2}{(x+1)^2(3x-1)^2} - \frac{512x^6\left(3x^2+1\right)^2}{(3x-1)(x+1)}(a+b-2ab). \end{aligned}$$

After some cancellation, expansion of terms, and dividing both sides by 4, we find that this is equivalent to:

$$\begin{aligned} &16(a-b)^2x^4\left(3x^2+1\right)^2 - (a+b-2ab)\frac{128x^6\left(3x^2+1\right)^2}{(3x-1)(x+1)} - 64(a+b-2ab)^2x^6(3x+1)^2 \\ &\qquad + (a+b-2ab)\frac{256x^7(3x+1)\left(3x^2+1\right)}{(x+1)(3x-1)} = 32(a-b)^2x^4(x+1)(3x-1)\left(3x^2+1\right) \end{aligned}$$

$$+ (a-b)^4 (x+1)^4 (3x-1)^4 - 16 (a+b-2ab)(a-b)^2 x^3 (3x+1)(x+1)^2 (3x-1)^2 .$$

Combining terms, we find

$$\begin{aligned} -64 (a+b-2ab)^2 x^6 (3x+1)^2 + 128 (a+b-2ab) x^6 \left(3x^2+1\right) &= 16 (a-b)^2 x^4 \left(3x^2+1\right)\left(3x^2+4x-3\right) \\ + (a-b)^4 (x+1)^4 (3x-1)^4 - 16 (a+b-2ab)(a-b)^2 x^3 (3x+1)(x+1)^2 (3x-1)^2 &. \end{aligned}$$

By rearranging terms, we realize that this is equivalent to showing

$$\begin{aligned} 16x^4 \left(3x^2+1\right) &\left(8x^2 (a+b-2ab) - \left(3x^2+4x-3\right)(a-b)^2\right) \\ &= \left(8x^3 (3x+1)(a+b-2ab) - (x+1)^2 (3x-1)^2 (a-b)^2\right)^2 . \end{aligned}$$

To show that this is impossible, we will show the left hand side is greater than the right hand side, except when $a = 1-b, x = 1$. For ease of notation, define

$$\begin{aligned} m(x) &:= 16x^4 \left(3x^2+1\right)\left(8x^2 (a+b-2ab) - \left(3x^2+4x-3\right)(a-b)^2\right), \\ n(x) &:= \left(8x^3 (3x+1)(a+b-2ab) - (x+1)^2 (3x-1)^2 (a-b)^2\right)^2 . \end{aligned}$$

We claim that $m(x) > n(x)$. Set $s = a+b-1$ and $t = a-b$. Without loss of generality, we may assume that $t \geq 0$. Under this assumption, give $t \geq 0$, because $a, b \in [0,1]$, one can show that $|s| \leq 1-t$. Using the fact that

$$a+b-2ab = \tfrac{1}{2}(1-s^2+t^2),$$

we can rewrite $m(x)$ and $n(x)$ as

$$\begin{aligned} m(x) &:= 16x^4 \left(3x^2+1\right)\left(4x^2 \left(1-s^2+t^2\right) - \left(3x^2+4x-3\right) t^2\right), \\ n(x) &:= \left(4x^3 (3x+1)\left(1-s^2+t^2\right) - (x+1)^2 (3x-1)^2 (a-b)^2\right)^2 . \end{aligned}$$

The function $m(x) - n(x)$ can be seen as a quadratic polynomial of s^2 with a negative leading coefficient. Therefore it suffices to check $m(x) \geq n(x)$ for $s^2 = 0$ and $s^2 = (1-t)^2$.

Case 1: $s^2 = 0$**.** In this case, $m(x) - n(x)$ equals

$$(1-x)^2 \left[48x^6 + 8x^3(1+7x-9x^2-3x^3)t^2 - (1-3x-5x^2+3x^3)^2 t^4\right].$$

The expression in the square bracket can be seen as a quadratic polynomial of t^2. with a negative leading coefficient. To prove $m(x) \geq n(x)$ in the $s^2 = 0$ case, it suffices to check

$$48x^6 + 8x^3(1+7x-9x^2-3x^3)t^2 - (1-3x-5x^2+3x^3)^2 t^4$$

is non-negative at the endpoints $t = 0$ and $t = 1$. When $t = 0$, the above equals $48x^6 \geq 0$. When $t = 1$, the above equals

$$(1-x)^2(-1+4x+10x^2-12x^3+15x^4),$$

which can be easily shown to be non-negative.

Case 2: $s^2 = (1-t)^2$**.** In this case, $m(x) - n(x)$ equals

$$t(-t + 2tx - 4x^2 + 3tx^2)^2 \left[8x^2 + 24x^4 - (-1 + 2x + 3x^2)^2 t\right].$$

As $t \geq 0$, it suffices to check the expression in the square bracket is non-negative at $t = 1$. When $t = 1$, that expression equals

$$-1 + 4x + 10x^2 - 12x^3 + 15x^4$$

which again is non-negative. (Notice that we have equality here if and only if $a = b = 0$, which makes $t = 0$.)

This proves our claim that $\frac{f'(x)}{f(x)} \neq \frac{1}{x}$. □

Now, assume we have p_1, p_2 such that $p_1 + p_2 = 1$ and $0 \leq p_1, p_2 \leq 1$. For the sake of notation, define $f_{a,b}(x) := a\bar{b} + \bar{a}b + 2c(x) - x$. We now show that when the cardinality is 2 we have a unique minimum point when $\boldsymbol{p}, \boldsymbol{a}$, and $\boldsymbol{b}$ are optimal.

Proposition 10. *For every a_1, a_2, b_1, b_2, the equation*

$$p_1 f_{a_1,b_1}(x) + p_2 f_{a_2,b_2}(x)$$

has exactly one solution for x in $(0, 1)$.

Proof. Without loss of generality let x_1 and x_2 with $x_1 \leq x_2$ be the unique roots of $f_{a_1,b_1}(x)$ and $f_{a_2,b_2}(x)$ respectively. Assume there is some fixed x^* such that $p_1 f_{a_1,b_1}(x^*) + p_2 f_{a_2,b_2}(x^*) = 0$. Clearly $x_1 \leq x^* \leq x_2$. This is because for $x < x_1$ we have $f_{a_1,b_1}(x), f_{a_2,b_2}(x) > 0$, and similarly for $x > x_2$ we have $f_{a_1,b_1}(x), f_{a_2,b_2}(x) < 0$.

It suffices to show that $q(x) := -\frac{f_{a_1,b_1}(x)}{f_{a_2,b_2}(x)}$ is injective in the range $x \in (x_1, x_2)$. (This would imply different values of x within this interval are roots at different p_1, p_2.) We shall show that q is, in fact, monotone increasing in the interval $x \in (x_1, x_2)$.

We have that

$$\lim_{x \to x_1^+} \frac{f'_{a_1,b_1}(x)}{f_{a_1,b_1}(x)} = \infty > \frac{1}{x_1}$$

and

$$\lim_{x \to x_2^-} \frac{f'_{a_2,b_2}(x)}{f_{a_2,b_2}(x)} = -\infty < \frac{1}{x_2}.$$

By Lemma 9, we can lower bound $\frac{f'_{a_1,b_1}(x)}{f_{a_1,b_1}(x)}$ by $\frac{1}{x}$ and upper bound $\frac{f'_{a_2,b_2}(x)}{f_{a_2,b_2}(x)}$ by $\frac{1}{x}$, therefore giving

$$\frac{f'_{a_1,b_1}(x)}{f_{a_1,b_1}(x)} \geq \frac{1}{x} \geq \frac{f'_{a_2,b_2}(x)}{f_{a_2,b_2}(x)},$$

with equality only when $a = b = 0$. After differentiating q, it is clear that this implies the statement that q is monotone increasing, proving our lemma. □

We now continue to the case where the cardinality is at least 3. Assume now that we have $p_1, p_2, \ldots p_n$ such that $\sum_{i=1}^n p_i = 1$ and $0 \leq p_1, p_2, \ldots, p_n \leq 1$, for $n \geq 3$.

Proposition 11. *The equation*

$$\sum_{i=1}^{n} p_i f_{a_i,b_i}(x) = 0,$$

for $n \geq 3$, has exactly one solution in x in the interval $[0, 1]$.

Proof. For sake of convenience let us define

$$g_{\boldsymbol{p},\boldsymbol{a},\boldsymbol{b}}(x) := \sum_{i=1}^{n} p_i f_{a_i,b_i}(x)$$

where, again, $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ are the points $(a_1, a_2, \ldots, a_n)$, $(b_1, b_2, \ldots, b_n)$, and $(p_1, p_2, \ldots, p_n)$ respectively.

Assume for the sake of contradiction $g_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{p}}(x) = 0$ has at least two roots in x in the interval $[0, 1]$. We know that at least one root has odd multiplicity, because $f_{a_i,b_i}(0) \geq 0 \geq f_{a_i,b_i}(1)$, and it is not difficult to check that for any sufficiently small $\epsilon > 0$, $f(-\epsilon) > 0$ and $f(1+\epsilon) < 0$. Let $x = x_1, x_2$ be two roots such that x_1 is the smallest root with odd multiplicity and such that x_2 is the smallest root which is not x_1.

Consider the locus of points $(\tilde{p_1}, \tilde{p_2}, \ldots, \tilde{p_n})$ which satisfy the condition $0 \leq \tilde{p_1}, \tilde{p_2}, \ldots, \tilde{p_n} \leq 1$ and the following:

$$\sum \tilde{p_i} = 1 \tag{5}$$

$$g_{\tilde{\boldsymbol{p}},\boldsymbol{a},\boldsymbol{b}}(x_1) = 0 \tag{6}$$

$$g_{\tilde{\boldsymbol{p}},\boldsymbol{a},\boldsymbol{b}}(x_2) = 0 \tag{7}$$

where $\tilde{\boldsymbol{p}}$ is defined as the point with coordinates $(\tilde{p_1}, \tilde{p_2}, \ldots, \tilde{p_n})$.

If $n \geq 4$ we can let all but 3 components of $\tilde{\boldsymbol{p}}$ be 0, because there are only 3 linear equations in $\tilde{p_i}$ which need to be satisfied. We now only need to consider the case when $n = 3$.

Consider the set of points $\tilde{\boldsymbol{p}}$ that only satisfy (5) and (6). The set of these points forms a line segment, with the endpoints satisfying the condition that either $\tilde{p_1} = 0$, $\tilde{p_2} = 0$, or $\tilde{p_3} = 0$. Clearly, (p_1, p_2, p_3) is on this line segment.

We first consider the case where $x_1 < x_2$.

Consider $\boldsymbol{p'}$, which we define to be a point along this aforementioned line segment. Clearly, setting $\boldsymbol{p'} = (p_1, p_2, p_3)$ results in $g_{\boldsymbol{p'},\boldsymbol{a},\boldsymbol{b}}(x_2)$ becoming 0, by the definition of x_2. Now, consider what happens if $\boldsymbol{p'}$ is moved to each of the two endpoints of the line segment. It is clear that for at least one of these endpoints, $g_{\boldsymbol{p'},\boldsymbol{a},\boldsymbol{b}}(x_2) \geq 0$ by linearity.

Take $\boldsymbol{p'} = \boldsymbol{p''}$ to be at this endpoint, with $\boldsymbol{p''}$ satisfying $g_{\boldsymbol{p''},\boldsymbol{a},\boldsymbol{b}}(x_2) \geq 0$. We also have $g_{\boldsymbol{p''},\boldsymbol{a},\boldsymbol{b}}(x_1 + \epsilon) < 0$ for arbitrarily small $\epsilon > 0$, and so by the Intermediate Value Theorem $g_{\boldsymbol{p''},\boldsymbol{a},\boldsymbol{b}}(x)$ must therefore have another root in the interval $x \in (x_1, x_2]$. Because we have that either $\tilde{p_1} = 0$, $\tilde{p_2} = 0$, or $\tilde{p_3} = 0$, we can disregard one of the $p_i f_{a_i,b_i}(x)$ terms in the summation, and still have at least 2 roots. However, this contradicts Proposition 10, proving Proposition 11.

Now, if $x_1 > x_2$, we choose $p' = p''$ to satisfy $g_{\boldsymbol{p'},\boldsymbol{a},\boldsymbol{b}}(x_2) \leq 0$. We have that $g_{\boldsymbol{p''},\boldsymbol{a},\boldsymbol{b}}(x_1 - \epsilon) \geq 0$ for arbitrarily small $\epsilon > 0$. Then using the Intermediate Value Theorem implies we have a root of $g_{\boldsymbol{p''},\boldsymbol{a},\boldsymbol{b}}(x)$ which gives a contradiction, using the same reasoning as above. □

Now we are ready to prove Theorem 5.

Proof of Theorem 5. First, Proposition 7 implies that we must have $\frac{\partial G}{\partial c_i} = 0$ for each i and for any minimum point $\boldsymbol{c}$, assuming $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}$ give the maximum possible value of the objective function. Also, Proposition 8, Proposition 10, and Proposition 11 together imply that there exists a unique minimum point for $n \geq 1$. Combining these two results completes the proof. □

5 Optimality for $n = 1$

In this section we explicitly compute the value of OPT_n for $n = 1$ given a conjecture.

We begin by defining $h(x) := H(x, \bar{x})$. We wish to maximize $h(a) + h(b)$ across points satisfying

$$L(a\bar{b} + b\bar{a} + 2c) - H(ab - c, a\bar{b} + c, b\bar{a} + c, \bar{a}\bar{b} - c) = 0,$$

where $a, b \in [0, 1]$ and c is the unique minimum point. Define $X = a\bar{b} + b\bar{a} + 2c$, $E_1 = ab - c$, $E_2 = a\bar{b} + c$, $E_3 = b\bar{a} + c$, and $E_4 = \bar{a}\bar{b} - c$.

Theorem 12. *For a given a, there is a unique value of b satisfying $b \le \bar{a}$ and $L(X) - H(E_1, E_2, E_3, E_4) = 0$, which we denote $B(a)$.*

Proof. Consider $L(X) - H(E_1, E_2, E_3, E_4) = 0$ for a constant. We claim that, if $b < \bar{a}$,

$$\frac{\partial}{\partial b}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) < 0.$$

Let $\frac{\partial E_1}{\partial b} = Q$. Note that when a is held constant,

$$\frac{\partial E_2}{\partial b} = -Q, \frac{\partial E_3}{\partial b} = -Q + 1, \frac{\partial E_4}{\partial b} = Q - 1, \frac{\partial X}{\partial b} = 1 - 2Q.$$

This means that

$$\begin{aligned}
&\frac{\partial}{\partial b}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) \\
&= L'(X) \cdot (1 - 2Q) - H'(E_1) \cdot Q - H'(E_2) \cdot (-Q) - H'(E_3) \cdot (-Q + 1) - H'(E_4) \cdot (Q - 1) \\
&= Q\left(-2\log\tfrac{1-X}{2X} - \log\tfrac{1}{eE_1} + \log\tfrac{1}{eE_2} + \log\tfrac{1}{eE_3} - \log\tfrac{1}{eE_4}\right) + L'(X) - H'(E_3) + H'(E_4) \\
&= \left(\log\left(\left(\frac{2X}{1-X}\right)^2 \cdot \frac{eE_1 \cdot eE_4}{eE_2 \cdot eE_3}\right)\right) + L'(X) - H'(E_3) + H'(E_4).
\end{aligned}$$

By the optimality of c,

$$\left(\frac{1-X}{2X}\right)^2 = \frac{E_1 \cdot E_4}{E_2 \cdot E_3}$$

meaning the expression inside the logarithm is simply equal to 1. Thus, the entire expression simplifies to

$$\log\left(\frac{(1-X)}{2X} \cdot \left(\frac{1}{eE_3}\right)^{-1} \cdot \frac{1}{eE_4}\right) = \log\left(\frac{(1-X)E_3}{2XE_4}\right).$$

Now, we wish to prove

$$b < \bar{a} \implies \log\left(\frac{(1-X)E_3}{2XE_4}\right) \iff (1-X)E_3 < 2XE_4.$$

Note that, when $b < \bar{a}$,

$$2E_4 = 2(\bar{a}\bar{b} - c) = 2 - 2a - 2b + 2ab - 2c > 1 - a - b + 2ab - 2c = 1 - (a\bar{b} + b\bar{a} + 2c) = 1 - X.$$

In addition, $X = a\bar{b} + b\bar{a} + 2c > b\bar{a} + c = E_3$ as, for optimal c, $a\bar{b} + c > 0$. This means, for $b < \bar{a}$, $(1-X)E_3 < 2XE_4$, proving that

$$a < \bar{b} \implies \frac{\partial}{\partial b}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) < 0.$$

We claim that $L(X) - H(E_1, E_2, E_3, E_4) \geq 0$ at $b = 0$. This holds because, in this case, the only value of c that keeps $E_1, E_2, E_3, E_4 \geq 0$ is $c = 0$, meaning $c = 0$ must be the optimal value of c. Then, $X = a, E_1 = 0, E_2 = \bar{a}, E_3 = a, E_4 = 0$. Since $H(0) = 0$ and $L(a) = H(a) + H(\bar{a}) + \bar{a}$,

$$L(X) - H(E_1, E_2, E_3, E_4) = \bar{a} \geq 0,$$

proving this claim.

We also claim that $L(X) - H(E_1, E_2, E_3, E_4) \leq 0$ at $b = \bar{a}$. To prove this, we must first determine the optimal value of c. Such a c must satisfy

$$\left(\frac{1-X}{2X}\right)^2 = \frac{E_1 \cdot E_4}{E_2 \cdot E_3}$$

$$\implies \left(\frac{1-(a^2+\bar{a}^2+2c)}{2(a^2+\bar{a}^2+2c)}\right)^2 = \frac{(a\bar{a}-c)(a\bar{a}-c)}{(a^2+c)(\bar{a}^2+c)}.$$

This holds at $c = a\bar{a}$ which, by uniqueness, is the only c satisfying this equation. When $b = \bar{a}$ and $c = a\bar{a}$, $X = 1, E_1 = 0, E_2 = a, E_3 = \bar{a}, E_4 = 0$ so

$$(L(X) - H(E_1, E_2, E_3, E_4)) = 0 - H(a, \bar{a}).$$

Since $H(a, \bar{a}) \geq 0$, this proves

$$(L(X) - H(E_1, E_2, E_3, E_4)) \leq 0$$

when $b = \bar{a}$. Combining these pieces of information shows that $B(a)$ is well-defined. This is because $L(X) - H(E_1, E_2, E_3, E_4)$ for fixed a is greater than or equal to 0 at $b = 0$, less than or equal to 0 at $b = \bar{a}$ and strictly decreasing, meaning it has exactly one root in $[0, \bar{a}]$ and so $B(a)$ is unique. □

We wish to maximize $h(a) + h(b)$ across all points (a, b) satisfying $L(X) - H(E_1, E_2, E_3, E_4) = 0$. We claim this set of points is defined as points either of the form $(a, B(a))$ or $(1 - B(a), a)$ for $a \in [0, 1]$. This holds because, as proved earlier, all points with $b \leq \bar{a}$ are of the form $(a, B(a))$. In addition, since $L(X) - H(E_1, E_2, E_3, E_4) = 0$ at (a, b) if and only if $L(X) - H(E_1, E_2, E_3, E_4) = 0$ at $(\bar{a}, \bar{b})$ because c and X are the same at these two points. In addition the values of E_i are simply permuted, proving this claim. Thus, $L(X) - H(E_1, E_2, E_3, E_4) = 0$ only at points of the form $(a, B(a))$ or of the form $(1 - B(a), a)$. However, since $h(a) = h(\bar{a})$, $h(a) + h(b) = h(\bar{a}) + h(\bar{b})$ so the same maximum is produced on either of these sets. Thus, we only need to consider points of the form $(a, B(a))$ to determine the maximum value of $h(a) + h(b)$.

Conjecture 13. *We claim that $h(a) + h(B(a))$ is maximized when $a = B(a)$.*

First, note that $B(a) = B^{-1}(a)$ as $L(X) - H(E_1, E_2, E_3, E_4) = 0$ at $(a, B(a))$ if and only if $L(X) - H(E_1, E_2, E_3, E_4) = 0$ at $(B(a), a)$. Also, $B(a)$ is strictly decreasing because $B(0) = 1$, $B(1) = 0$ and B is invertible. Since $B(a)$ is strictly decreasing, so is $B_*(a) = B(a) - a$ meaning there is a unique value a_i such that $a_i = B(a_i)$. Note that,

$$(B^{-1})'(a_i) = \frac{1}{B'(B^{-1}(a_i))} \implies B'(a_i) = \frac{1}{B'(a_i)} \implies B'(a_i) = -1$$

as $B(a)$ is strictly decreasing. Thus, when $a = a_i$,

$$\frac{d}{da}(h(a) + h(B(a))) = h'(a) + h'(B(a))B'(a) = h'(a_i) + h'(a_i) \cdot (-1) = 0.$$

Computer experimentation described in Section 6 suggests there are no other values of a satisfying $h'(a) + h'(B(a))B'(a) = 0$ and that the point where $a = B(a)$ is a global maximum.

We now seek to determine the unique value of $a = a_0 \leq 0.5$ such that, when $a = b = a_0$, $L(X) - H(E_1, E_2, E_3, E_4) = 0$. We first wish to compute

$$\frac{\partial}{\partial a}\left(L(X) - H(E_1, E_2, E_3, E_4)\right)$$

along the curve $a = b$. Note that, when $a = b$, $E_1' = 2a$, $E_2' = E_3' = 1 - 2a$, $E_4' = 2a - 2$, and $X' = 2 - 4a$. Thus,

$$\begin{aligned}
&\frac{\partial}{\partial a}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) \\
&= \log\left(\frac{1-X}{2X}\right)^{2-4a} - \log\left(\frac{1}{eE_1}\right)^{2a} - \log\left(\frac{1}{eE_2}\right)^{1-2a} - \log\left(\frac{1}{eE_3}\right)^{1-2a} - \log\left(\frac{1}{eE_4}\right)^{2a-2} \\
&= 2a\log\left(\left(\frac{2X}{1-X}\right)^2 \cdot \frac{eE_1 \cdot eE_4}{eE_2 \cdot eE_3}\right) + \log\left(\left(\frac{1-X}{2X}\right)^2 \frac{E_2 \cdot E_3}{E_4^2}\right).
\end{aligned}$$

Once again, by the optimality of c, the expression within the logarithm is simply equal to 1. In addition, when $a = b$, $E_2 = E_3 = \frac{X}{2}$ so the expression simplifies to

$$\frac{\partial}{\partial a}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) = \log\left(\left(\frac{1-X}{4E_2}\right)^2 \frac{E_2^2}{E_4^2}\right) = \log\left(\frac{(1-X)^2}{16E_4^2}\right).$$

Note that the optimal value of c when $a = b$ must satisfy the equation

$$\frac{(a^2 - c)(\bar{a}^2 - c)}{(a\bar{a} + c)^2} = \left(\frac{1 - (2a\bar{a} + 2c)}{2(2a\bar{a} + 2c)}\right)^2.$$

Algebraic manipulation shows that, when $a \leq 0.5$, $c = \frac{1}{6}(6a^2 - 6a - 2\sqrt{3}(1 - 2a) + 3)$ satisfies this equation. In addition, when $a \geq \frac{2-\sqrt{3}}{4}$, this value for c fits within the necessary bounds and is therefore the optimal value of c. We now focus in on the case where $\frac{2-\sqrt{3}}{4} \leq a$ to show a_0 must be in this region. By the equation for c,

$$E_4 = \frac{(1-2a)}{\sqrt{3}} - a + \frac{1}{2}, \quad 1 - X = \frac{2}{\sqrt{3}}(1 - 2a).$$

Thus,

$$\frac{\partial}{\partial a}\left(L(X) - H(E_1, E_2, E_3, E_4)\right) = \left(\frac{\left(\frac{2}{\sqrt{3}}(1-2a)\right)^2}{16\left(\left(\frac{1}{\sqrt{3}} - \frac{1}{2}\right)(1-2a)\right)^2}\right).$$

Note that at $a = b = \frac{1}{2}$, $L(X) - H(E_1, E_2, E_3, E_4) = -1$, so $a_0 < \frac{1}{2}$ and this expression simplifies to

$$\log\left(\frac{\frac{4}{3}}{16\left(\frac{1}{\sqrt{3}} - \frac{1}{2}\right)^2}\right) = \log(7 + 4\sqrt{3}).$$

This value, combined with the value of $L(X) - H(E_1, E_2, E_3, E_4)$ at $a = b = \frac{1}{2}$ means that $L(X) - H(E_1, E_2, E_3, E_4)$ along the curve $a = b$, for $\frac{2-\sqrt{3}}{4} \leq a \leq \frac{1}{2}$,

$$= \log(7 + 4\sqrt{3})\left(a - \frac{1}{2}\right) + 1.$$

Solving for where this line intersects 0 shows that

$$a_0 = \frac{\log(2+\sqrt{3}) - 1}{2\log(2+\sqrt{3})} \approx 0.23684.$$

At this point the maximum value of $h(a) + h(b)$ along the curve $L(X) - H(E_1, E_2, E_3, E_4) = 0$ can be computed, giving an approximate value of 1.57948.

6 Numerical experiments

We conducted several numerical experiments which suggest that $n = 1$ is sufficient for finding an optimal rate.

First, we randomly selected each of $\boldsymbol{p} = (p_1, p_2, p_3), \boldsymbol{a} = (a_1, a_2, a_3), \boldsymbol{b} = (b_1, b_2, b_3)$ from the range $[0, 1]$, independently and from an uniform distribution. This was done a million times. For each trial, we normalized $\boldsymbol{p}$ such that $p_1 + p_2 + p_3 = 1$, and checked that the constraint (2) held. Of the trials where (2) held, we found a maximum rate of 1.57777.

Then, we numerically checked the following conjecture:

Conjecture 14. *There exists a constant λ such that the optimization problem OPT_n and the following optimization problem (which we shall denote as OPT_n^{**}) have the same optimal value.*

Maximize the objective function

$$\begin{aligned} F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}) := & \sum_{i=1}^{n} p_i (H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i)) \\ & + \lambda \cdot \left[L\left(\sum_{i=1}^{n} p_i (a_i \bar{b}_i + \bar{a}_i b_i + 2c_i) \right) - \sum_{i=1}^{n} p_i H\left(a_i b_i - c_i, a_i \bar{b}_i + \bar{a}_i b_i + 2c_i, \bar{a}_i \bar{b}_i - c_i \right) \right] \end{aligned} \tag{8}$$

over all points $\boldsymbol{a} = (a_1, ..., a_n), \boldsymbol{b} = (b_1, ..., b_n), \boldsymbol{p} = (p_1, ..., p_n) \in [0, 1]^n$ that satisfy:

$$\sum_{i=1}^{n} p_i = 1$$

for all points $\boldsymbol{c} = (c_1, ..., c_n)$ that make the terms inside $H(., ., ., .)$ nonnegative.

We can rewrite (8) as:

$$\begin{aligned} F(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p}) := & \sum_{i=1}^{n} p_i (H(a_i, \bar{a}_i) + H(b_i, \bar{b}_i) - H\left(a_i b_i - c_i, a_i \bar{b}_i + \bar{a}_i b_i + 2c_i, \bar{a}_i \bar{b}_i - c_i \right)) \\ & + \lambda \cdot L\left(\sum_{i=1}^{n} p_i (a_i \bar{b}_i + \bar{a}_i b_i + 2c_i) \right). \end{aligned} \tag{9}$$

If this conjecture were to be true, we could only consider $n = 2$ in OPT_n^{**} to 2, using the same trick of moving $\boldsymbol{p}$ in a linear fashion.

For this conjecture to be true, it must be true in the neighborhood of $\boldsymbol{p} = (1), \boldsymbol{a} = (0.23684), \boldsymbol{b} = (0.23684)$. In particular, λ must make the total derivative of (9) at this point, which gives that $\lambda = 0.8889$.

We can also try to check that this conjecture is true computationally. For ease of notation, denote

$$A := L\left(\sum_{i=1}^{n} p_i (a_i \bar{b}_i + \bar{a}_i b_i + 2c_i) \right) - \sum_{i=1}^{n} p_i H\left(a_i b_i - c_i, a_i \bar{b}_i + \bar{a}_i b_i + 2c_i, \bar{a}_i \bar{b}_i - c_i \right),$$

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

and M^* to be the maximal value of OBJ in OPT_n, as before. Our conjecture is equivalent to the statement that

$$\text{OBJ} + \lambda \cdot A \leq M^*.$$

is true for all points, or that

$$\lambda \cdot A \leq M^* - \text{OBJ}.$$

Now, if $A \geq 0$, we have

$$\lambda \leq \frac{M^* - \text{OBJ}}{A},$$

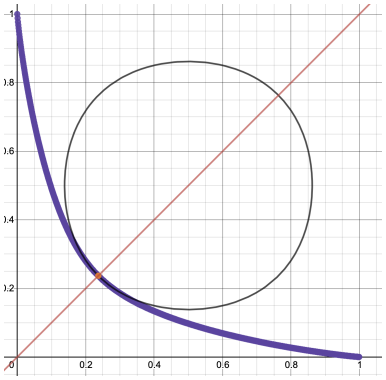
and otherwise we have

$$\lambda \geq \frac{M^* - \text{OBJ}}{A}.$$

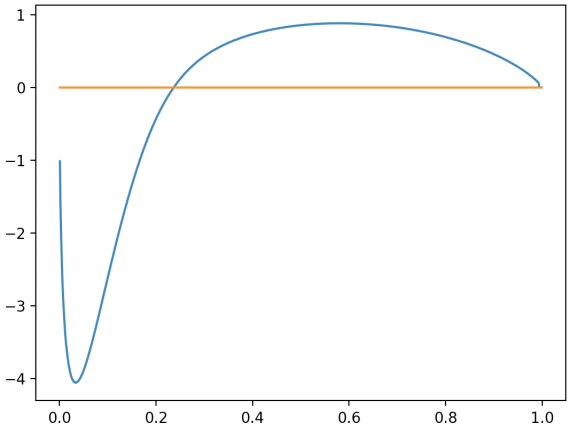
After once again repeatedly randomly selecting $\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}$ uniformly from the range $[0, 1]$ a million times in the same manner as before (though now without (2) needing to be satisfied), we checked computationally that every randomly selected point satisfies these inequalities. Indeed, all points with $A \geq 0$ had $\frac{M^* - \text{OBJ}}{A} \geq 0.903512$, while all points with $A \leq 0$ had $\frac{M^* - \text{OBJ}}{A} \leq 0.858611$, as desired.

The source code in C++ for both experiments have been submitted as attachments.

The image below shows the plot of $B(a)$ in purple, the line $a = b$ in red, the curve of $h(a) + h(b) = 1.57948$ in black, and the point $(0.23684, 0.23684)$. This shows that no other point on the plot of $B(a)$ enters the ring of $h(a) + h(b) = 1.57948$ meaning no other point has a greater value of $h(a) + h(b)$.



The image below plots a on the x-axis and $h(a) + h(B(a))B'(a)$ on the y-axis and shows that this function has exactly one root, which occurs when $a = B(a)$.



Acknowledgements

We would like to thank our mentor, Dr. Zilin Jiang, for suggesting our research topic, for his advice, and for his assistance throughout the project. We met with Dr. Jiang once a week, and this paper is the result of our collective work. We are grateful to the MIT PRIMES-USA program for the opportunity to work on this project, which otherwise would not have been possible, and we thank Dr. Tanya Khovanova and Dr. Alexander Vitanov for their feedback on the manuscript.

References

[Aig86] M. Aigner. Search problems on graphs. *Discrete Appl. Math.*, 14(3):215–230, 1986.

[BL87] 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.

[Due85] 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.

[GMSV92] L. Gargano, V. Montouri, G. Setaro, and U. Vaccaro. An improved algorithm for quantitative group testing. *Discrete Applied Mathematics*, 36(3):299 – 306, 1992.

[JPV19] 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.

[ZBM87] 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.