On a Combinatorial Problem Arising in Machine Teaching

Joakim Sunde [1] **Brigt Håvardstun** [1] **Jan Kratochvíl** [2] **Jan Arne Telle** [1]

Abstract

We study a model of machine teaching where the teacher mapping is constructed from a size function on both concepts and examples. The main question in machine teaching is the minimum number of examples needed for any concept, the so-called teaching dimension. A recent paper (Ferri et al., 2024) conjectured that a worst case for this model, as a function of the size of the concept class, occurs when the consistency matrix contains the binary representations of numbers from zero and up. In this paper we prove their conjecture. The result can be seen as a generalization of a theorem resolving the edge isoperimetry problem for hypercubes (Hart, 1976). Our proof is based on a generalization of a lemma of (Graham, 1970).

1. Introduction

In formal models of *machine learning* (Valiant, 1984) we have a concept class C of possible hypotheses, an unknown target concept $c^* \in C$ and training data given by correctly labelled random examples. The concept class C is given by a binary matrix M whose rows are concepts and whose column set is the domain of examples X, with $M(c, x) = 1$ if c is consistent with x labelled positively, i.e. with $(x, 1)$ rather than with $(x, 0)$. In formal models of *machine teaching* a set of labelled examples w called a *witness* is instead carefully chosen by a teacher T, i.e. $T(c^*) = w$, so the learner can reconstruct c^*. The common goal is to keep the *teaching dimension*, i.e., the cardinality of the witness set, $\max_{c \in C} |T(c)|$, as small as possible. In recent years, the field of machine teaching has seen various applications in fields like pedagogy (Shafto et al., 2014), trustworthy AI (Zhu et al., 2018), reinforcement learning (Zhang et al., 2021), active learning (Wang et al., 2021) and explainable AI (Yang et al., 2021).

Various models of machine teaching have been proposed, e.g. the classical teaching dimension model (Goldman & Kearns, 1995), the optimal teacher model (Balbach, 2008), recursive teaching (Zilles et al., 2011), preference-based teaching (Gao et al., 2017), no-clash teaching (Fallat et al., 2023), and probabilistic teaching (Ferri et al., 2022). In (Telle et al., 2019) a model focusing on teaching size is introduced, and in (Ferri et al., 2024) an algorithm called Greedy constructing the teacher mapping in this model is given.

Greedy assumes two total orderings $\prec_C$ on C and $\prec_X$ on X, with $\prec_X$ extended to $\prec_W$ on subsets of labelled examples $W = 2^{X \times \{0,1\}}$ by shortlex ordering. In the Greedy algorithm the teacher defines its mapping iteratively: go through W in the order of $\prec_W$, and for a given witness $w = \{(x_1, b_1)...(x_q, b_q)\}$, find the earliest (in $\prec_C$ order) $c \in C$ consistent with w (i.e. with $M(c, x_i) = b_i$ for all $1 \leq i \leq q$) such that $T(c)$ is not yet defined, then set $T(c) = w$ and continue with next witness (or drop this w if no such c exists).

To compare the teaching dimension achievable by Greedy to that of other models, the authors of (Ferri et al., 2024) argued as follows when a large witness is used: If Greedy assigns $T(c) = w$ for some $w = \{(x_1, b_1)...(x_q, b_q)\}$, then we may ask why was c not assigned to a smaller witness? Assuming there are $|X| = n$ examples, any subset $Q \subseteq X$ of size $q-1$ when labelled consistent with c has already been tried by Greedy, and hence some other concept must already have been assigned to any such Q, and all these concepts are distinct. This means we must have taught $\binom{n}{q-1} = k$ other concepts already. But then we have already taught at least $k + 1$ concepts and we can again ask why were any of these not taught by a smaller witness of size $q - 2$? It must be that any such witnesses (labelled to be consistent with some concept among the $k + 1$ we already have) must have been used to teach other, again distinct, concepts.

Note that, to verify how many distinct witnesses exist, corresponding to new concepts, that are labelled consistently with one of these $k + 1$ concepts, one must sum up the number of distinct rows when projecting on $q - 2$ columns, for all choices of these columns. Note that the number of distinct rows, i.e., witnesses, and hence the number of con-

[1]Department of Informatics, University of Bergen, Bergen, Norway [2]Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic. Correspondence to: Joakim Sunde <Joakim.Sunde@uib.no>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

cepts, when projecting on $q-2$ columns, for all choices of these columns, depends on the matrix M one does the projection on. The authors of (Ferri et al., 2024) wanted to find the matrix M minimizing the sum of unique rows after doing the projection, thus arriving at the following combinatorial question. What is the binary matrix M on k distinct rows and n columns that would give the smallest sum when projecting on q columns? They conjectured that this was achieved by the matrix $H_{n,k}$ consisting of the k rows corresponding to the binary representations of the numbers between zero and $k-1$, with leading 0s to give them length n. In this paper we prove this conjecture.

Consider the binary consistency graph G_C on the set of concepts versus the set W of subsets of labelled examples, with a concept c adjacent to $w \in W$ if c is consistent with each labelled example in w. We can view the Greedy Matching algorithm as working on G_C. Note that the above-mentioned sum for a matrix M when projecting on q columns (called $m_q(M)$ in the next section) is then the number of W-vertices on q examples that have at least one neighbor among the concepts. Since we prove that $H_{n,k}$ minimizes this value for all q, it means that it minimizes the number of W-vertices having a neighbor in the consistency graph, over all concept classes on k concepts over a domain of size n. As the consistency graph is of importance in machine teaching, this is an indication that our result has a general relevance in that field.

When $q = n-1$ this minimization question is equivalent to asking for the induced subgraph on k vertices of the hypercube of dimension n having the maximum number of edges, for the following reason. The rows of the k by n binary matrix M are viewed as k vertices of the hypercube of dimension n, labelled in the standard way, with two vertices adjacent iff their labels differ in exactly one coordinate. When $q = n-1$ we have $\binom{n}{n-1} = n$ choices for the projection on q columns and each such projection leaves out exactly one column (and a column corresponds to a dimension of the hypercube). Each such projection could give at most k unique rows, so the maximum achievable sum of unique projection rows is k times n. The main observation when $q = n-1$ is the following: three or more rows cannot have the same projection row, but two rows can, and two rows of M give the same projection row (when leaving out a column/dimension) if and only if the corresponding pair of vertices are adjacent (across the dimension we left out), and thus when $q = n-1$, the sum of unique projection rows for M is k times n minus the number of edges induced in the hypercube. Thus, a matrix minimizing the sum of unique projection rows for $q = n-1$ will also maximize the number of induced edges in the hypercube of dimension n.

The question of finding the matrix achieving the maximum mentioned above is called the *edge isoperimetry problem for the hypercube*. This has been shown (Hart, 1976) to be achieved by $H_{n,k}$, and the edge isoperimetry of the hypercube has been studied extensively in (McIlroy, 1974; Delange, 1975; Hart, 1976; Greene & Knuth, 1990; Agnarsson, 2013) to name a few articles. The result we give in this paper is thus a generalization of the edge isoperimetry problem on the hypercube, as we show that $H_{n,k}$ is the solution not only when $q = n-1$, but for all values of $1 \le q \le n$.

The rest of our paper is organized as follows. In Section 2 we give the formal definition of the conjecture. In Section 3 we show that the conjecture would be settled if we could prove a stronger theorem. Then in Section 4 we prove this stronger theorem, using a generalization of an old result from (Graham, 1970).

2. Statement of the main theorem

Let M be a $k \times n$ binary matrix whose all k rows are distinct. Let $\mathcal{M}_{n,k}$ be the set of all such matrices. For any binary matrix A, let dif(A) denote the number of unique rows in the matrix A. For $Q \subseteq \{1, 2, ..., n\}$, let $M(Q)$ be the submatrix of $M \in \mathcal{M}_{n,k}$ formed by taking the columns with indices from Q. Finally for integers a and b where $a \le b$ let $[a, b] = \{a, a+1, ..., b\}$. Our main interest is the number

$$m_q(M) = \sum_{Q \in \binom{[1,n]}{q}} \text{dif}(M(Q))$$

which is the sum of the numbers of unique rows for each submatrix of M created by picking a subset of the columns of size q. For fixed positive integers k, n and q, we are interested in finding a matrix $M \in \mathcal{M}_{k,n}$ with the minimum value of $m_q(M)$. Let $m_q(n, k)$ be this minimum value, i.e.,

$$m_q(n, k) = \min_{M \in \mathcal{M}_{n,k}} m_q(M).$$

We show that the $k \times n$ binary matrix $H_{n,k}$ whose rows are the binary representations of all numbers between zero and $k-1$ achieves this minimum value of $m_q(n, k)$.

Example 1.

$$H_{4,5} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \cong \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

with $m_1(H_{4,5}) = 7$, $m_2(H_{4,5}) = 16$, $m_3(H_{4,5}) = 15$ and $m_4(H_{4,5}) = 5$.

It will be useful for us to use the following recursive definition of $H_{n,k}$. Note that the invariants we are interested in remain unchanged by permutations of rows or columns of the matrices under consideration. In this sense we consider all such matrices equivalent. Let $\mathbf{0}$ be the all 0 row vector and let $\mathbf{0}^T$ be the all 0 column vector, and similarly for $\mathbf{1}$ and $\mathbf{1}^T$. Then

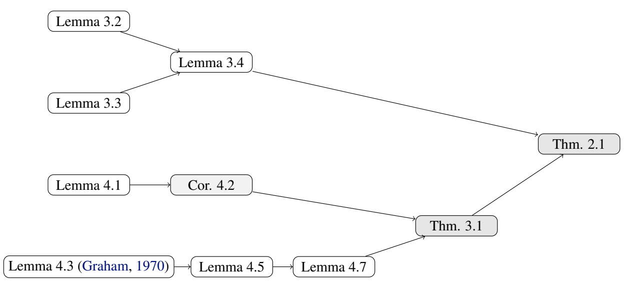
$$H_{n,k} = \begin{cases} \mathbf{0} & k=1 \\ \begin{pmatrix} H_{n-1,\lceil \frac{k}{2}\rceil} & \mathbf{0}^T \\ H_{n-1,\lfloor \frac{k}{2}\rfloor} & \mathbf{1}^T \end{pmatrix} & k>1. \end{cases}$$

Let $h_q(n,k) = m_q(H_{n,k})$. Our goal is thus to prove the following theorem.

Theorem 2.1. *For any positive integers q, n, k where $q \le n$ and $k \le 2^n$,*

$$m_q(n,k) = h_q(n,k).$$

Here is a diagram showing how we will prove Theorem 2.1.



3. A sufficient condition

The goal of this section is to prove that the following theorem (whose proof we leave to the next section) implies Theorem 2.1.

Theorem 3.1. *For any positive integers q, n, k where $q \le n$ and $k \le 2^n$,*

$$\min_{\lceil \frac{k}{2}\rceil \le x \le k-1} h_q(n,x) + h_{q-1}(n-1,k-x)$$
$$= h_q(n, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lfloor \frac{k}{2} \rfloor).$$

which is just stating that the minimum value of the expression on the left occurs when $x = \lceil \frac{k}{2} \rceil$.

Lemma 3.2. *The h numbers satisfy the recurrence relation*

$$h_q(n,1) = \binom{n}{q}$$

and

$$h_q(n,k) = h_q(n, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lfloor \frac{k}{2} \rfloor)$$

for $k > 1$.

Proof. Let Q be a q-element subset of the column-index set $\{1, 2, \dots, n\}$. If $n \notin Q$, then each of the bottom $\lfloor \frac{k}{2} \rfloor$ rows of $H_{n,k}(Q)$ appears as a row among the $\lceil \frac{k}{2} \rceil$ top ones, and hence $m_q(H_{n,k}(\{1, 2, \dots, n-1\})) = m_q(H_{n-1,\lceil \frac{k}{2}\rceil}) = h_q(n-1, \lceil \frac{k}{2} \rceil)$. Since the value of the last column (the n-th column) is 0 for the $\lfloor \frac{k}{2} \rfloor$ rows and 1 for the rest we have that if $n \in Q$, every row from the bottom $\lfloor \frac{k}{2} \rfloor$ rows of $H_{n,k}(Q)$ differs from any row from the $\lceil \frac{k}{2} \rceil$ top ones, and so the sum over those Q that contain n contributes exactly $m_{q-1}(H_{n-1,\lceil \frac{k}{2}\rceil}) + m_{q-1}(H_{n-1,\lfloor \frac{k}{2}\rfloor}) = h_{q-1}(n-1, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lfloor \frac{k}{2} \rfloor)$. Thus

$$h_q(n,k)$$
$$= h_q(n-1, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lfloor \frac{k}{2} \rfloor).$$

This can be slightly simplified as follows. Note that $h_q(n-1, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lceil \frac{k}{2} \rceil)$ is exactly the contribution of the $\lceil \frac{k}{2} \rceil$ top rows of $H_{n,k}$ to $m_q(H_{n,k})$, i.e., $m_q((H_{n-1,\lceil \frac{k}{2}\rceil} \mathbf{0}^T))$ what equals $m_q((\mathbf{0}^T H_{n-1,\lceil \frac{k}{2}\rceil})) = m_q(H_{n,\lceil \frac{k}{2}\rceil}) = h_q(n, \lceil \frac{k}{2} \rceil)$ and the claim follows. ☐

Lemma 3.3. *For any positive integers q, n, k where $q \le n$ and $k \le 2^n$,*

$$m_q(n,k) \ge \min_{\lceil \frac{k}{2}\rceil \le x \le k-1} m_q(n,x) + m_{q-1}(n-1,k-x)$$

Proof. Let $A \in \mathcal{M}_{n,k}$ be a matrix that minimizes m_q over $\mathcal{M}_{n,k}$, i.e., it satisfies $m_q(A) = m_q(n,k)$.

If $k = 1$, then every $Q \in \binom{[1,n]}{q}$ contributes 1 to the sum $\sum_Q \mathrm{dif}(M(Q))$, and hence $m_q(A) = \binom{n}{q}$.

Let $k > 1$. Suppose w.l.o.g. that the last column contains both 0's and 1's and that the number of 1's does not exceed the number of 0's. Let y be the number of 0's in it, and assume that the 0's are in rows $1, \dots, y$ and the 1's in rows $y+1, \dots, k$, with $y \ge k-y$, i.e., $y \ge \lceil \frac{k}{2} \rceil$. Let T be the submatrix of A determined by rows $1, \dots, y$ and columns $1 \dots, n-1$, and let B be the submatrix determined by rows $y+1, \dots, k$ and columns $1, \dots, n-1$, i.e.,

$$A = \begin{pmatrix} T & \mathbf{0}^T \\ B & \mathbf{1}^T \end{pmatrix}.$$

We further denote by $T^* = (T \ \mathbf{0}^T)$ the submatrix of A formed by its top y rows.

We first observe that

$$\sum_{Q \in \binom{[1,n]}{q} : n \notin Q} \mathrm{dif}(A(Q)) \ge \sum_{Q \in \binom{[1,n-1]}{q}} \mathrm{dif}(T(Q)) \qquad (1)$$

since in this case we do not include the n-th column in Q. Because the n-th column is not included, we observe that any unique row projection counted by $\sum_{Q\in\binom{[1,n-1]}{q}} \mathrm{dif}(T(Q))$ will be a subset of the unique row projections counted by $\sum_{Q\in\binom{n}{q}:n\notin Q} \mathrm{dif}(A(Q))$.

We also see that

$$\sum_{Q\in\binom{[1,n]}{q}:n\in Q} \mathrm{dif}(A(Q)) \geq \sum_{Q'\in\binom{[1,n-1]}{q-1}} \mathrm{dif}(T(Q')) + \sum_{Q'\in\binom{[1,n-1]}{q-1}} \mathrm{dif}(B(Q')) \tag{2}$$

since when $n \in Q$, each row leading to a unique projection in A, the entire row, except that last column, was in T or in B. As we know that each projection of rows in B will differ from rows in T in at least the n-th column, we can count up the number of unique projections in T and B separately, using Q' of size $(q-1)$ as we will increase the size by 1 when we add back n.

We see that $m_q(T^*) \geq m_{q-1}(n-1, k-y)$ since for a given T^* for n columns and y rows, the m_q value will be greater or equal to the minimum value over all matrices of size (n, y). We also see that $m_{q-1}(B) \geq m_{q-1}(n-1, k-y)$ using the same idea. Given a matrix B of size $(n-1, k-y)$, the m_q value of this matrix will be larger or equal to the minimum value over all matrices of size $(n-1, k-y)$. Combining these we get the inequality

$$m_q(T^*) + m_{q-1}(B) \geq m_{q-1}(n-1, k-y) + m_{q-1}(n-1, k-y) \tag{3}$$

.

Finally we need the inequality

$$m_q(n, y) + m_{q-1}(n-1, k-y) \geq \min_{\lceil\frac{k}{2}\rceil \leq x \leq k-1} m_q(n, x) + m_{q-1}(n-1, k-x). \tag{4}$$

To show the soundness of this inequality we observe that when $x = \lceil\frac{k}{2}\rceil$ we have $x \leq y$, as $x = \lceil\frac{k}{2}\rceil \leq y$. We also have $y \leq k-1$, as we assume that the last column has both 0's and 1's. When $x = k-1$, we have $y \leq x$, as we do the minimization over all possible values of x in this range, we know that we are evaluating $x = y$ as well. Hence the minimum will be equal to or less than $m_q(n, y) + m_{q-1}(n-1, k-y)$.

Using these four relations we can now finish the proof.

$$m_q(A) = \sum_{Q\in\binom{[1,n]}{q}:n\notin Q} \mathrm{dif}(A(Q)) + \sum_{Q\in\binom{[1,n]}{q}:n\in Q} \mathrm{dif}(A(Q)) \geq$$

(by combining (1) and (2))

$$\geq \sum_{Q\in\binom{[1,n-1]}{q}} \mathrm{dif}(T(Q)) + \sum_{Q'\in\binom{[1,n-1]}{q-1}} \mathrm{dif}(T(Q')) + \sum_{Q'\in\binom{[1,n-1]}{q-1}} \mathrm{dif}(B(Q')) =$$

$$= \sum_{Q\in\binom{[1,n-1]}{q}} \mathrm{dif}(T(Q)) + \sum_{Q\in\binom{[1,n]}{q},n\in Q} \mathrm{dif}(T^*(Q)) + \sum_{Q'\in\binom{[1,n-1]}{q-1}} \mathrm{dif}(B(Q')) =$$

$$= m_q(T^*) + m_{q-1}(B) \geq$$

(by (3))

$$\geq m_q(n, y) + m_{q-1}(n-1, k-y) \geq$$

(and by (4))

$$\geq \min_{\lceil\frac{k}{2}\rceil \leq x \leq k-1} m_q(n, x) + m_{q-1}(n-1, k-x).$$

□

Lemma 3.4. *Theorem 3.1 implies Theorem 2.1*

Proof. Certainly $m_q(n, k) \leq h_q(n, k)$, we prove the other inequality by induction on k. The base case $k = 1$ follows from $m_q(n, 1) = h_q(n, 1) = \binom{n}{q}$.

Suppose $k > 1$. Lemmas 3.2 and 3.3 imply that

$$m_q(n, k) \geq \min_{\lceil\frac{k}{2}\rceil \leq x \leq k-1} m_q(n, x) + m_{q-1}(n-1, k-x) \geq$$

(by the induction hypothesis)

$$\geq \min_{\lceil\frac{k}{2}\rceil \leq x \leq k-1} h_q(n, x) + h_{q-1}(n-1, k-x) =$$

(by Theorem 3.1)

$$= h_q(n, \lceil\frac{k}{2}\rceil) + h_{q-1}(n-1, \lfloor\frac{k}{2}\rfloor) =$$

(by Lemma 3.2)

$$= h_q(n, k).$$

□

4. Proving Theorem 3.1

In this section we will prove Theorem 3.1 by showing that $h_q(n,x)$ "increases" at least as fast as $h_{q-1}(n-1,k-x)$ "decreases" when x starts at $\lceil\frac{k}{2}\rceil$ and increases until $k-1$. To be more precise, we will show that

$$\begin{aligned} &h_q(n,\lceil\frac{k}{2}\rceil+j)-h_q(n,\lceil\frac{k}{2}\rceil) \\ &\geq h_{q-1}(n-1,\lfloor\frac{k}{2}\rfloor)-h_{q-1}(n-1,\lfloor\frac{k}{2}\rfloor-j) \end{aligned} \tag{5}$$

for any $j\geq 1$ such that $\lceil\frac{k}{2}\rceil+j\leq k-1$. Since the above inequality is equivalent to

$$\begin{aligned} &h_q(n,\lceil\frac{k}{2}\rceil+j)+h_{q-1}(n-1,\lfloor\frac{k}{2}\rfloor-j) \\ &\geq h_q(n,\lceil\frac{k}{2}\rceil)+h_{q-1}(n-1,\lfloor\frac{k}{2}\rfloor), \end{aligned}$$

it follows straightforwardly that the minimum value of $h_q(n,x)+h_{q-1}(n-1,k-x)$ over $x\in[\lceil\frac{k}{2}\rceil,k-1]$ is attained by $x=\lceil\frac{k}{2}\rceil$.

We first need to understand the behavior of the $h_q(n,k)$ numbers as k increases or decreases. Let $|x|$ denote the Hamming weight (number of 1's) in the binary representation of integer x. We recall that the binomial coefficient $\binom{n}{k}$ by definition evaluates to 0 when $k<0$ or $k>n$. Similarly, we define the boundary values of $h_q(n,k)$ for $q=0$ and $k=0$ as $h_0(n,k)=1$ (for $k>0$) and $h_q(n,0)=0$.

Lemma 4.1. *For any integers x,q,n such that $0\leq x\leq 2^n-1$ and $0\leq q\leq n$, we have*

$$h_q(n,x+1)=h_q(n,x)+\binom{n-|x|}{q-|x|},$$

and for integers x,q,n such that $1\leq x\leq 2^{n-1}$ and $1\leq q\leq n$, we have

$$h_{q-1}(n-1,x-1)=h_{q-1}(n-1,x)-\binom{n-1-|x-1|}{q-1-|x-1|}.$$

Proof. We prove the first formula, the second one then follows directly by applying the first one for $x-1,q-1$ and $n-1$.

For the boundary values of q and x, we have $h_0(n,1)=1=0+1=h_0(n,0)+\binom{n}{0}$, $h_0(n,x+1)=1=1+0=h_0(n,x)+\binom{n-|x|}{-|x|}$ for $x\geq 1$, and $h_q(n,1)=\binom{n}{q}=0+\binom{n}{q}=h_q(n,0)+\binom{n-|0|}{q-|0|}$.

For the notrivial cases, suppose that $q\geq 1$ and $x\geq 1$. The only difference between $H_{n,x}$ and $H_{n,x+1}$ is that $H_{n,x+1}$ has one extra row, which is the binary representation of x with zeroes padded to the left if needed. Let S be the set of column indices where the last row of $H_{n,x+1}$ has a 1.

We first observe that $\mathrm{dif}(H_{n,x}(Q))=\mathrm{dif}(H_{n,x+1}(Q))$ whenever $S\not\subseteq Q\subseteq\{1,2,\ldots,n\}$. To see this let $i\in S\backslash Q$ and y be the number with binary representation having the same entry as x in the positions belonging to Q and 0's in all other positions. Then $y<x$ and $H_{n,x}$ contains a row which is the binary representation of y. Since this row of $H_{n,x}$ is equal to the last row of $H_{n,x+1}$ when only looking at the columns with indices in Q, $\mathrm{dif}(H_{n,x}(Q))=\mathrm{dif}(H_{n,x+1}(Q))$.

Then we see that $\mathrm{dif}(H_{n,x+1}(Q))=\mathrm{dif}(H_{n,x}(Q))+1$ whenever $S\subseteq Q$. This is because there is no row in $H_{n,x}$ where all the columns with indices in S are equal to 1, since the number of this row would be greater or equal to x.

So we are left with counting how many subsets Q of $\{1,...,n\}$ satisfy $S\subseteq Q$ and $|Q|=q$. This is exactly $\binom{n-|S|}{q-|S|}=\binom{n-|x|}{q-|x|}$.

□

Corollary 4.2. *For any integers q,n,x,j such that $0\leq q\leq n$, $0\leq x$, $1\leq j$ and $x+j\leq 2^n$, we have*

$$h_q(n,x+j)=h_q(n,x)+\sum_{i=x}^{x+j-1}\binom{n-|i|}{q-|i|}.$$

Moreover, whenever $1\leq q\leq n$ and $1\leq j\leq x\leq 2^{n-1}$, we have

$$h_{q-1}(n-1,x-j)=h_{q-1}(n-1,x)-\sum_{i=x-j}^{x-1}\binom{n-1-|i|}{q-1-|i|},$$

and whenever $1\leq q\leq n$ and $1\leq j\leq x-1\leq 2^{n-1}$, we have

$$\begin{aligned} &h_{q-1}(n-1,x-j-1) \\ &=h_{q-1}(n-1,x-1)-\sum_{i=x-j-1}^{x-2}\binom{n-1-|i|}{q-1-|i|}. \end{aligned}$$

Proof. The first two formulae follow from Lemma 4.1 by induction on j, the third formula follows from the second by substituting $x-1$ for x. □

In view of this corollary, the inequality (5) is equivalent to the claim that our goal is to prove that

$$\sum_{i=\lceil\frac{k}{2}\rceil}^{\lceil\frac{k}{2}\rceil+j-1}\binom{n-|i|}{q-|i|}\geq\sum_{i=\lfloor\frac{k}{2}\rfloor-j}^{\lfloor\frac{k}{2}\rfloor-1}\binom{n-1-|i|}{q-1-|i|}$$

holds true for all feasible q,n,k and j.

We first show some useful properties of Hamming weights which extend the following lemma from (Graham, 1970) whose proof was finalized in (Jones & Torrence, 1999).

Lemma 4.3. *((Graham, 1970; Jones & Torrence, 1999)) Let s, t be non-negative integers. Then there exists a bijective mapping $\theta : [0, r] \to [s, s + r]$ such that $|\theta(k)| \geq |k|$ for every $k \in [0, r]$.*

We will need a generalization of this lemma whose proof depends on the following observation:

Observation 4.4. Let $x \geq t$ be non-negative integers. Then $|x - t| \geq |x| - |t|$.

Proof. This follows directly from the standard subtraction algorithm for integers in binary representation. □

Lemma 4.5. *Let s, r, t be non-negative integers such that $r, t \geq 1$ and $s \geq r + t - 1$. Denote by $T = [s, s + r - 1]$ and $B = [s - r - t + 1, s - t]$. Then there exists a bijective mapping $\theta : T \to B$ such that $|\theta(x)| \geq |x| - |t|$ for all $x \in T$.*

Proof. Our proof works by induction on r. When $r = 1$, we have $T = \{s\}$ and $B = \{s - t\}$. The only possible mapping θ then simply maps s to $s - t$ and we see that $|\theta(s)| = |s - t| \geq |s| - |t|$ by Observation 4.4. Thus, the base case $r = 1$ is established for all values of $t \geq 1$.

Let $r > 1$. Create two matrices with r rows each

$$M_T = \begin{pmatrix} \overrightarrow{s+r-1} \\ \vdots \\ \overrightarrow{s+1} \\ \overrightarrow{s} \end{pmatrix} \text{ and } M_B = \begin{pmatrix} \overrightarrow{s-t} \\ \overrightarrow{s-t-1} \\ \vdots \\ \overrightarrow{s-r-t+1} \end{pmatrix}$$

where $\overrightarrow{x}$ is the base 2 representation of x as a binary vector with 0-s padded to the left so that all vectors have the same length. Finally let $M = \begin{pmatrix} M_T \\ M_B \end{pmatrix}$.

Reformulating the lemma in this matrix context we seek a bijective mapping θ of the rows of M_T to the rows of M_B such that $|\theta(x)| \geq |x| - |t|$ holds true for every row x of M_T. (With a slight abuse of notation we write $\theta : M_T \to M_B$.) The induction hypothesis states that this holds true, for this value of t, if the number of rows of each matrix is less than r.

Without loss of generality we may assume that the first (leftmost) column of M contains at least one 0 and at least one 1 (since we could disregard this column otherwise). Then if we look at the first column of M, there will be a point where a 1 appears for the first time, when moving through the rows from the bottom row up. This could happen either in the M_T part or in the M_B part of the matrix. We will deal with these 2 cases separately.

Case 1 (The first leftmost 1 appears in the M_T part of the matrix) We divide both the M_T and M_B matrices further and write M as

$$M = \begin{pmatrix} T_1 \\ T_2 \\ B_2 \\ B_1 \end{pmatrix}$$

where the bottom row of T_1 is the row where the first 1 appears (thus that row is 100...0), with T_2 being the remainder of M_T, and we let B_1 have the same number of rows as T_1. We will map T_1 to B_1 and T_2 to B_2. Since T_2 and B_2 have fewer rows than r (since T_1 and B_1 always have at least one row) and are on the form specified by the lemma since the smallest number in T_2 are the same as in T and the largest number in B_2 is the same as in B and we simply deleted some of the largest/smallest numbers of T and B to create T_2 and B_2 respectively so it will still be an interval. It follows by the induction hypothesis applied to T_2, B_2 and r as the number of rows of T_2, B_2 that, for the same value of t, there exists the required mapping $\theta_1 : T_2 \to B_2$. Note also that this is vacuously true if T_2 and B_2 are empty. Now if we ignore the first column of T_1, then $T_1(\{2, 3, ...\})$ is the binary representation of the numbers $0, 1, ..., |T_1| - 1$. So by Lemma 4.3 there is a mapping $\theta_2 : T_1(\{2, 3, ...\}) \to B_1$ such that $|\theta_2(x)| \geq |x|$ for every x. Adding back the first column of T_1 and using the same mapping between the rows as θ_2, we get a mapping $\theta_3 : T_1 \to B_1$ where $|\theta_3(x)| \geq |x| - 1$ for every x (since the Hamming weight of x increases by 1). Clearly $|x| - 1 \geq |x| - |t|$ when $t \geq 1$, hence combining θ_1 and θ_3 gives us a bijective mapping $\theta : T \to B$ with the required properties.

Case 2 (The first leftmost 1 appears in the M_B part) We divide the matrix in a similar way as in the first case

$$M = \begin{pmatrix} T_1 \\ T_2 \\ B_2 \\ B_1 \end{pmatrix}$$

so that the bottom row of B_2 is the row where the first 1 appears in the leftmost column (so this row is 100...0) and we let T_2 have the same number of rows as B_2. For a binary vector x let $\overline{x}$ be the complement of x, so $\overline{x} = (1, 1, 1, ..., 1) - x$. For a binary matrix A, let $\overline{A}$ be the matrix whose rows are the complements of the rows of A.

By the induction hypothesis, as there are fewer rows and we have the same value of t, there is a mapping $\theta_1 : T_2 \to B_2$ with the required properties. The top row of B_1 is $(011...1)$ so if we look at $\overline{B_1}$, the top row will be $(100...0)$, the next row will be $(100...01)$ and so on, meaning that if we ignore the first column we are counting up from 0 in binary. Lemma 4.3 then gives us a mapping $\theta_2 : \overline{B_1(\{2, 3, ...\})} \to \overline{T_1}$ with $|\theta_2(x)| \geq |x|$ for every x. Adding back the first column but keeping the row mapping we get a mapping $\theta_3 : \overline{B_1} \to \overline{T_1}$

where $|\underline{\theta_3(x)}| \geq |x| - 1$. Now define $\theta_4 : B_1 \to T_1$ by $\theta_4(x) = \overline{\theta_3(\overline{x})}$ and let $||x||$ be the length of the vector x.

$$\begin{aligned} |\theta_4(x)| - |x| &= |\overline{\theta_3(\overline{x})}| - |x| \\ &= ||x|| - |\theta_3(\overline{x})| - (||x|| - |\overline{x}|) \\ &= |\overline{x}| - |\theta_3(\overline{x})| \\ &\leq 1 \end{aligned} \tag{6}$$

To get (6) we use the fact that $|\theta_3(x)| \geq |x| - 1$.

We will now look at the inverse mapping $\theta_4^{-1} : T_1 \to B_1$. For any $y \in T_1$, there is an $x \in B_1$ such that $\theta_4(x) = y$. We just showed that $|\theta_4(x)| - |x| \leq 1$ which is the same as $|y| - |x| \leq 1$ which we can rewrite as $|y| - |\theta_4^{-1}(y)| \leq 1$. Multiplying both sides by -1 we get $|\theta_4^{-1}(y)| \geq |y| - 1$. Combining θ_4^{-1} and θ_1 in the natural way we get the desired mapping $\theta : T \to B$ satisfying $|\theta(x)| \geq |x| - 1 \geq |x| - |t|$ for every x.

Thus the lemma is proven for any $r, t \geq 1$ and any $s \geq r + t - 1$. ☐

We are now set to prove that the sum in the first formula of the Corollary 4.2 is always larger than the sums in the second and third formulae. We will need the following well known observation:

Observation 4.6. For any integers n, k, j such that $0 \leq j \leq k \leq n$,

$$\binom{n}{k} \geq \binom{n-j}{k-j}.$$

Lemma 4.7. *For all positive integers q, n, x, j such that $x + j - 1 \leq 2^n$ and $x - j \geq 0$,*

$$\sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} \geq \sum_{i=x-j}^{x-1} \binom{n-1-|i|}{q-1-|i|},$$

and if $x - j \geq 1$, then

$$\sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} \geq \sum_{i=x-j-1}^{x-2} \binom{n-1-|i|}{q-1-|i|}.$$

Proof. We show that there exists a bijection θ from $[x, x+j-1]$ to $[x-j, x-1]$ (or to $[x-j-1, x-2]$, respectively) such that $|\theta(i)| \geq |i| - 1$ for all i. This will prove the lemma since then for every term $\binom{n-|i|}{q-|i|}$ in the sum of the left hand side there will be a corresponding term in the sum on the right side

$$\binom{n-1-|\theta(i)|}{q-1-|\theta(i)|}$$

and we see that by Observation 4.6

$$\binom{n-1-|\theta(i)|}{q-1-|\theta(i)|} \leq \binom{n-1-(|i|-1)}{q-1-(|i|-1)} = \binom{n-|i|}{q-|i|}.$$

So if such a bijection exists, then for every term in the sum on the left hand side there will be a unique element in the sum on the right hand side which is no greater than the element on the left hand side. So then the sum on the left hand side must be greater than or equal to the sum on the right hand one. Now we just need to show that there exist such bijections θ. We will use Lemma 4.5.

Case 1. Let $s = x, r = j$ and $t = 1$. Then by Lemma 4.5 there is a mapping $\theta : [x, x+j-1] \to [x-1, x-j]\}$ such that $|\theta(i)| \geq |i| - |t| = |i| - 1$ for every i, which is the first bijection we wanted.

Case 2. If we set $t = 2$, then Lemma 4.5 gives us a mapping $\theta : [x, x+j-1] \to [x-2, x-j-1]$ such that $|\theta(i)| \geq |i| - |t| = |i| - 1$ for every i, which is the second bijection we wanted. ☐

We are now ready to prove Theorem 3.1.

Theorem 3.1. *For any positive integers q, n, k where $q \leq n$ and $k \leq 2^n$,*

$$\min_{\lceil \frac{k}{2} \rceil \leq x \leq k-1} h_q(n, x) + h_{q-1}(n-1, k-x)$$
$$= h_q(n, \lceil \frac{k}{2} \rceil) + h_{q-1}(n-1, \lfloor \frac{k}{2} \rfloor).$$

Proof. We consider the two cases depending on whether k is either even or odd.

Case 1 - k is even. Set $x = \frac{k}{2}$ and rewrite the left hand side of Theorem 3.1 as

$$\min_{0 \leq j \leq x-1} h_q(n, x+j) + h_{q-1}(n-1, x-j)$$

Then by the first and second part of Corollary 4.2 whenever $1 \leq j \leq x-1$, we can rewrite the expression which is minimized as

$$h_q(n, x) + \sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} + h_{q-1}(n-1, x)$$
$$- \sum_{i=x-j}^{x-1} \binom{n-1-|i|}{q-1-|i|}$$

By Lemma 4.7 we have $\sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} \geq \sum_{i=k-j}^{x-1} \binom{n-1-|i|}{q-1-|i|}$ for any $1 \leq j \leq x-1$, which means that the smallest value occurs when $j = 0$.

Case 2 - k is odd. Set $x = \lceil \frac{k}{2} \rceil$ and rewrite the expression of the left hand side of Theorem 3.1 as

$$\min_{0 \leq j \leq x-1} h_q(n, x+j) + h_{q-1}(n-1, x-1-j).$$

By the first and third part of Corollary 4.2 we can rewrite the part which is minimized for $1 \leq j \leq x-1$ as

$$h_q(n, x) + \sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} + h_{q-1}(n-1, x-1)$$

$$- \sum_{i=x-j-1}^{x-2} \binom{n-1-|i|}{q-1-|i|}.$$

By Lemma 4.7 we have $\sum_{i=x}^{x+j-1} \binom{n-|i|}{q-|i|} \geq \sum_{i=k-j-1}^{x-2} \binom{n-1-|i|}{q-1-|i|}$ for $1 \leq j \leq x-1$ so the smallest value of the expression will occur for $j = 0$.

□

By Lemma 3.4 this proves Theorem 2.1.

5. Conclusion

We have proven a conjecture of (Ferri et al., 2024), identifying a binary matrix M minimizing $m_q(M)$, the sum of the numbers of distinct rows over all submatrices on q columns. Let us further consider the complexity of computing $m_q(M)$ when the binary matrix M with k rows and n columns is given as input. There is a straightforward algorithm with runtime $O(n^q kq \log k)$. The question arises if computing $m_q(M)$ is FPT (Fixed Parameter Tractable, see Cygan et al. (2015)) when parameterized by q. In other words, is there an algorithm whose runtime is polynomial in the size of M, with any superpolynomial dependency restricted to q only? We leave this as an open problem.

Acknowledgements

Thanks to the reviewers for their careful reading and helpful suggestions.

Impact Statement

Advancing the field of machine teaching.

References

Agnarsson, G. On the number of hypercubic bipartitions of an integer. *Discrete Mathematics*, 313(24):2857–2864, 2013.

Balbach, F. J. Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, 397 (1-3):94–113, 2008.

Cygan, M., Fomin, F. V., Kowalik, L., Lokshtanov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., and Saurabh, S. *Parameterized Algorithms*. Springer, 2015. ISBN 978-3-319-21274-6. doi: 10.1007/978-3-319-21275-3. URL https://doi.org/10.1007/978-3-319-21275-3.

Delange, H. Sur la fonction sommatoire de la fonction" somme des chiffres". *Enseign. Math.*, 21(2):31–47, 1975.

Fallat, S., Kirkpatrick, D., Simon, H. U., Soltani, A., and Zilles, S. On batch teaching without collusion. *Journal of Machine Learning Research*, 24:1–33, 2023.

Ferri, C., Hernández-Orallo, J., and Telle, J. A. Non-cheating teaching revisited: A new probabilistic machine teaching model. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 2973–2979. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/412. URL https://doi.org/10.24963/ijcai.2022/412.

Ferri, C., Garigliotti, D., Håvardstun, B. A. T., Hernández-Orallo, J., and Telle, J. A. When redundancy matters: Machine teaching of representations, 2024. arXiv:2401.12711.

Gao, Z., Ries, C., Simon, H. U., and Zilles, S. Preference-based teaching. *Journal of Machine Learning Research*, 18:31:1–31:32, 2017. URL http://jmlr.org/papers/v18/16-460.html.

Goldman, S. A. and Kearns, M. J. On the complexity of teaching. *Journal of Computer and System Sciences*, 50 (1):20–31, 1995.

Graham, R. L. On primitive graphs and optimal vertex assignments. *Annals of the New York academy of sciences*, 175(1):170–186, 1970.

Greene, D. H. and Knuth, D. E. *Mathematics for the Analysis of Algorithms*, volume 504. Springer, 1990.

Hart, S. A note on the edges of the n-cube. *Discrete Mathematics*, 14(2):157–163, 1976.

Jones, J. C. and Torrence, B. F. The case of the missing case: the completion of a proof by rl graham. *Pi Mu Epsilon Journal*, 10(10):772–778, 1999.

McIlroy, M. D. The number of 1's in binary integers: bounds and extremal properties. *SIAM Journal on Computing*, 3 (4):255–261, 1974.

Shafto, P., Goodman, N. D., and Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55 – 89, 2014. ISSN 0010-0285. doi: https://doi.org/10.1016/j.cogpsych.2013.12.004. URL http://www.sciencedirect.com/science/article/pii/S0010028514000024.

Telle, J. A., Hernández-Orallo, J., and Ferri, C. The teaching size: computable teachers and learners for universal languages. *Machine Learning*, 108(8-9):1653–1675, 2019.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27 (11):1134–1142, 1984. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Wang, C., Singla, A., and Chen, Y. Teaching an active learner with contrastive examples. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17968–17980, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958adb57686c2fdec5796398de5f317a-Abstract.html.

Yang, S. C.-H., Vong, W. K., Sojitra, R. B., Folke, T., and Shafto, P. Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, 11(1):1–17, 2021.

Zhang, X., Bharti, S. K., Ma, Y., Singla, A., and Zhu, X. The sample complexity of teaching by reinforcement on q-learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 10939–10947. AAAI Press, 2021. doi: 10.1609/AAAI.V35I12.17306. URL https://doi.org/10.1609/aaai.v35i12.17306.

Zhu, X., Singla, A., Zilles, S., and Rafferty, A. N. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Zilles, S., Lange, S., Holte, R., and Zinkevich, M. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12(Feb):349–384, 2011.